Streaming source coding with delay

Cheng Chang

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-164
http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-164.html

December 19, 2007

Streaming Source Coding with Delay

by

Cheng Chang

M.S. (University of California, Berkeley) 2004
B.E. (Tsinghua University, Beijing) 2000

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineeing and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Anant Sahai, Chair
Professor Kannan Ramchandran
Professor David Aldous

Fall 2007

The dissertation of Cheng Chang is approved.

Chair	Date
	Date
	Date

University of California, Berkeley

Fall 2007

Streaming Source Coding with Delay

by

Cheng Chang

Abstract

Streaming Source Coding with Delay

by

Cheng Chang

Doctor of Philosophy in Engineering - Electrical Engineeing and Computer Sciences

University of California, Berkeley

Professor Anant Sahai, Chair

Traditionally, information theory is studied in the block coding context — all the source symbols are known in advance by the encoder(s). Under this setup, the minimum number of channel uses per source symbol needed for *reliable* information transmission is understood for many cases. This is known as the capacity region for channel coding and the rate region for source coding. The block coding error exponent (the convergence rate of the error probability as a function of block length) is also studied in the literature.

In this thesis, we consider the problem that source symbols stream into the encoder in real time and the decoder has to make a decision within a finite delay on a symbol by symbol basis. For a finite delay constraint, a fundamental question is how many channel uses per source symbol are needed to achieve a certain symbol error probability. This question, to our knowledge, has never been systematically studied. We answer the source coding side of the question by studying the asymptotic convergence rate of the symbol error probability as a function of delay — the delay constrained error exponent.

The technical contributions of this thesis include the following. We derive an upper bound on the delay constrained error exponent for lossless source coding and show the achievability of this bound by using a fixed to variable length coding scheme. We then extend the same treatment to lossy source coding with delay where a tight bound on the error exponent is derived. Both delay constrained error exponents are connected to their block coding counterpart by a "focusing" operator. An achievability result for lossless multiple terminal source coding is then derived in the delay constrained context. Finally, we borrow the genie-aided feed-forward decoder argument from the channel coding literature

to derive an upper bound on the delay constrained error exponent for source coding with decoder side-information. This upper bound is strictly smaller than the error exponent for source coding with both encoder and decoder side-information. This "price of ignorance" phenomenon only appears in the streaming with delay setup but not the traditional block coding setup.

These delay constrained error exponents for streaming source coding are generally different from their block coding counterparts. This difference has also been recently observed in the channel coding with feedback literature.

Professor Anant Sahai
Dissertation Committee Chair

Contents

List of Figures **v**

List of Tables **viii**

Acknowledgements **ix**

1 Introduction **1**

1.1 Sources, channels, feedback and side-information 3

1.2 Information, delay and random arrivals 6

1.3 Error exponents, focusing operator and bandwidth 9

1.4 Overview of the thesis 12

1.4.1 Lossless source coding with delay 12

1.4.2 Lossy source coding with delay 13

1.4.3 Distributed lossless source coding with delay 14

1.4.4 Lossless Source Coding with Decoder Side-Information with delay . 15

2 Lossless Source Coding **16**

2.1 Problem Setup and Main Results 16

2.1.1 Source Coding with Delay Constraints 16

2.1.2 Main result of Chapter 2: Lossless Source Coding Error Exponent with Delay 22

2.2 Numerical Results 22

2.2.1 Comparison of error exponents 23

2.2.2 Non-asymptotic results: prefix free coding of length 2 and queueing delay 24

2.3 First attempt: some suboptimal coding schemes 28

2.3.1 Block Source Coding's Delay Performance 28

2.3.2 Error-free Coding with Queueing Delay . . . 29
2.3.3 Sequential Random Binning . . . 31
2.4 Proof of the Main Results . . . 40
2.4.1 Achievability . . . 40
2.4.2 Converse . . . 44
2.5 Discussions . . . 46
2.5.1 Properties of the delay constrained error exponent $E_s(R)$ . . . 46
2.5.2 Conclusions and future work . . . 52

3 Lossy Source Coding **54**
3.1 Lossy source coding . . . 54
3.1.1 Lossy source coding with delay . . . 55
3.1.2 Delay constrained lossy source coding error exponent . . . 56
3.2 A brief detour to peak distortion . . . 56
3.2.1 Peak distortion . . . 57
3.2.2 Rate distortion and error exponent for the peak distortion measure . 57
3.3 Numerical Results . . . 59
3.4 Proof of Theorem 2 . . . 60
3.4.1 Converse . . . 60
3.4.2 Achievability . . . 60
3.5 Discussions: why peak distortion measure? . . . 63

4 Distributed Lossless Source Coding **65**
4.1 Distributed source coding with delay . . . 65
4.1.1 Lossless Distributed Source Coding with Delay Constraints . . . 66
4.1.2 Main result of Chapter 4: Achievable error exponents . . . 69
4.2 Numerical Results . . . 72
4.2.1 Example 1: symmetric source with uniform marginals . . . 73
4.2.2 Example 2: non-symmetric source . . . 74
4.3 Proofs . . . 76
4.3.1 ML decoding . . . 76
4.3.2 Universal Decoding . . . 85
4.4 Discussions . . . 88

5 Lossless Source Coding with Decoder Side-Information **90**

5.1 Delay constrained source coding with decoder side-information 90

5.1.1 Delay Constrained Source Coding with Side-Information 92

5.1.2 Main results of Chapter 5: lower and upper bound on the error exponents 93

5.2 Two special cases 95

5.2.1 Independent side-information 95

5.2.2 Delay constrained encryption of compressed data 96

5.3 Delay constrained Source Coding with Encoder Side-Information 99

5.3.1 Delay constrained error exponent 100

5.3.2 Price of ignorance 102

5.4 Numerical results 103

5.4.1 Special case 1: independent side-information 103

5.4.2 Special case 2: compression of encrypted data 103

5.4.3 General cases 106

5.5 Proofs 108

5.5.1 Proof of Theorem 6: random binning 108

5.5.2 Proof of Theorem 7: Feed-forward decoding 110

5.6 Discussions 117

6 Future Work **118**

6.1 Past, present and future 119

6.1.1 Dominant error event 119

6.1.2 How to deal with both past and future? 120

6.2 Source model and common randomness 122

6.3 Small but interesting problems 123

Bibliography **124**

A Review of fixed-length block source coding **131**

A.1 Lossless Source Coding and Error Exponent 131

A.2 Lossy source coding 133

A.2.1 Rate distortion function and error exponent for block coding under average distortion 133

A.3 Block distributed source coding and error exponents 135

A.4 Review of block source coding with side-information 137

B Tilted Entropy Coding and its delay constrained performance 140

B.1 Tilted entropy coding 140

B.2 Error Events 142

B.3 Achievability of delay constrained error exponent $E_s(R)$ 143

C Proof of the concavity of the rate distortion function under the peak distortion measure 145

D Derivation of the upper bound on the delay constrained lossy source coding error exponent 146

E Bounding source atypicality under a distortion measure 149

F Bounding individual error events for distributed source coding 152

F.1 ML decoding: Proof of Lemma 9 152

F.2 Universal decoding: Proof of Lemma 11 155

G Equivalence of ML and universal error exponents and tilted distributions 158

G.1 case 1: $\gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}) < R^{(\gamma)} < \gamma H(\bar{p}^1_{x|y}) + (1-\gamma)H(p^1_{xy})$. . . . 160

G.2 case 2: $R^{(\gamma)} \geq \gamma H(\bar{p}^1_{x|y}) + (1-\gamma)H(p^1_{xy})$. 165

G.3 Technical Lemmas on tilted distributions 167

H Bounding individual error events for source coding with side-information 175

H.1 ML decoding: Proof of Lemma 13 175

H.2 Universal decoding: Proof of Lemma 14 177

I Proof of Corollary 1 179

I.1 Proof of the lower bound 179

I.2 Proof of the upper bound 180

J Proof of Theorem 8 181

J.1 Achievability of $E_{ei}(R)$ 181

J.2 Converse 185

List of Figures

1.1 Information theory in 50 years 2
1.2 Shannon's original problem 3
1.3 Separation, feedback and side-informfation 4
1.4 Noiseless channel 5
1.5 Delay constrained source coding for streaming data 7
1.6 Random arrivals of source symbols 7
1.7 Block length vs decoding error 10
1.8 Delay vs decoding error 11
1.9 Focusing bound vs sphere packing bound 11

2.1 Timeline of source coding with delay 17
2.2 Timeline of source coding with universal delay 18
2.3 Source coding error exponents 23
2.4 Ratio of error exponent with delay to block coding error exponent 24
2.5 Streaming prefix-free coding system 25
2.6 Transition graph of random walk B_k for source distribution 26
2.7 Error probability vs delay (non-asymptotic results) 27
2.8 Block coding's delay performance 28
2.9 Error free source coding for a fixed-rate system 29
2.10 Union bound of decoding error 35
2.11 A universal delay constrained source coding system 41
2.12 Plot of $G_R(\rho)$ 47
2.13 Derivative of $E_s(R)$ 51

3.1 Timeline of lossy source coding with delay 55
3.2 Peak distortion measure and valid reconstructions under different D 58

3.3 $R-D$ curve under peak distortion constraint is a staircase function 60
3.4 Lossy source coding error exponent 61
3.5 A delay optimal lossy source coding system. 61

4.1 Slepian-Wolf source coding 66
4.2 Timeline of delay constrained source coding with side-information 66
4.3 Rate region for the example 1 source 73
4.4 Error exponents plot: $E_{SW,x}(R_x, R_y)$ plotted for $R_y = 0.71$ and $R_y = 0.97$. 75
4.5 Rate region for the example 2 source 76
4.6 Error exponents plot for source x for fixed R_y as R_x varies 77
4.7 Error exponents plot for source y for fixed R_y as R_x varies 78
4.8 Two dimensional plot of the error probabilities $p_n(l,k)$, corresponding to error events (l,k), contributing to $\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}]$ in the situation where $\inf_{\gamma\in[0,1]} E_y(R_x, R_y, \rho, \gamma) \geq \inf_{\gamma\in[0,1]} E_x(R_x, R_y, \rho, \gamma)$. 82
4.9 Two dimensional plot of the error probabilities $p_n(l,k)$, corresponding to error events (l,k), contributing to $\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}]$ in the situation where $\inf_{\gamma\in[0,1]} E_y(R_x, R_y, \gamma) < \inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma)$. 83
4.10 2D interpretation of the *score* of a sequence pair 87

5.1 Lossless source coding with decoder side-information 91
5.2 Lossless source coding with side-information: DMC 91
5.3 Timeline of delay constrained source coding with side-information 92
5.4 Encryption of compressed data with delay 97
5.5 Compression of encrypted data with delay 98
5.6 Information-theoretic model of compression of encrypted data 98
5.7 Lossless source coding with both encoder and decoder side-information . . . 99
5.8 Timeline of Delay constrained source coding with encoder/decoder side-information 101
5.9 Delay constrained Error exponents for source coding with independent decoder side-information 104
5.10 A binary stream cipher for source $\{\epsilon, 1-\epsilon\}$ 104
5.11 Delay constrained Error exponents for uniform source with symmetric decoder side-information (symmetric channel) 105
5.12 Uniform source and side-information connected by a symmetric erasure channel 105
5.13 Delay constrained Error exponents for uniform source with symmetric decoder side-information (erasure channel) 106

5.14 Delay constrained Error exponents for general source and decoder side-information . . . 107
5.15 Feed-forward decoder . . . 111

6.1 Past, present and future . . . 119
6.2 Dominant error event . . . 120

A.1 Block lossless source coding . . . 131
A.2 Block lossy source coding . . . 133
A.3 Achievable region for Slepian-Wolf source coding . . . 138

J.1 Another FIFO variable length coding system . . . 182

List of Tables

1.1 Delay constrained information theory problems 8

5.1 A non-uniform source with uniform marignals 106

6.1 Type of dominant error events for different coding problems 120

Acknowledgements

I would like to thank my thesis advisor, Professor Anant Sahai, for being a great source of ideas and inspiration through my graduate study. His constant encouragement and support made this thesis possible. I am truly honored to be Anant's first doctoral student.

Professors and post-docs in the Wireless Foundations Center have been extremely helpful and inspirational through my graduate study. Professor Kannan Ramchandran's lectures in EE225 turned my research interest from computer vision to fundamental problems in signal processing and communication. Professor Venkat Anantharam introduced me to the beautiful information theory and Professor Michael Gastpar served in my qualification exam committee. For more than a year Dr. Stark Draper and I worked on the problem that built the foundations of this thesis. I appreciate the help from all of them.

I would like to thank all my colleagues in the Wireless Foundations Center for making my life in Cory enjoyable. Particularly, my academic kid brothers Rahul Tandra, Mubaraq Mishra, Hari Palaiyanur and Pulkit Grover shouldered my workload in group meetings. Dan Schonberg captained our softball team for six years. My cubicle neighbors, Vinod Prabhakaran and Jonathan Tsao, were always ready to talk to me on a wide range of topics. I am also grateful to our wonderful EECS staff members Amy Ng and Ruth Gjerde for making my life a lot easier.

My years at Berkeley have been a wonderful experience. I would like to thank all my friends for all those nights out, especially Keda Wang, Po-lung Chia, Will Brett, Beryl Huang, Jing Xiang and Jiangang Yao for being the sibling figures that I never had.

My parents have been extremely supportive through all these years, I hope I made you proud. Last but not least, I would like to thank the people of Harbin, China and Berkeley, USA, for everything.

Chapter 1

Introduction

In 1948 Claude Shannon published The Mathematical Theory of Communication [72], which "single-handedly started the field of information theory" [82]. 60 years later, what Shannon started in that paper has evolved into what is shown in Figure 1.1. Some of the most important aspects of information theory are illustrated in this picture such as sources, noisy channels, universality, feedback, security, side-information, interaction between multiple agents etc. For trained eyes, it is easy to point out some classical information theory problems such as lossless/lossy source coding [72, 74], channel coding [72], Slepian-Wolf source coding [79], multiple-access channels [40], broadcast channels [37, 24], arbitrarily varying channels [53], relay channels [25] and control over noisy channels [69] as sub problems of Figure 1.1. Indeed we can come up with some *novel* problems by just changing the connections in Figure 1.1, for example a problem named "On the security of joint multiple access channel/correlated source coding with partial channel information and noisy feedback" is well in the scope of that picture. And this may have already been published in the literature.

In this thesis, we study information theory from another dimension— delay. Traditionally, the issue of delay is ignored, as in most information theory problems a *message* which contains all the information is known at the encoder prior to the whole communication task. Or the issue of delay is studied in a strict *real time* setup such as [58]. We instead study the intermediate regime where a finite delay for *every* information symbol is part of the system design criteria. Related works are found in the networking literature on delay and protocol [38] and recently on implications of finite delay on channel coding [67]. In

this thesis, we focus on the implications of end-to-end delay on a variety of source coding problems.

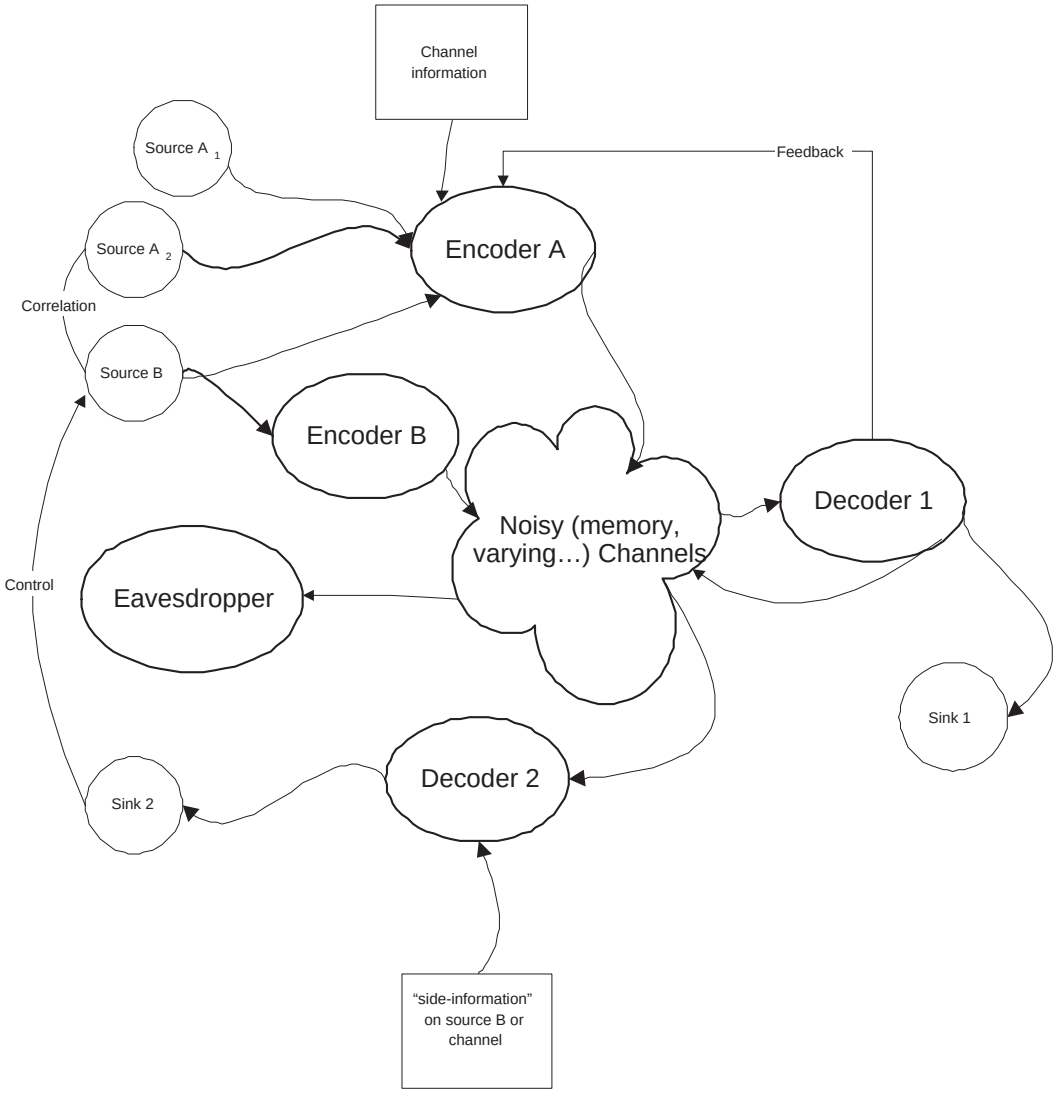

Figure 1.1. Sources, channels, feedback, side-information and eavesdroppers

1.1 Sources, channels, feedback and side-information

Instead of looking at information theory as a big complicated system illustrated in Figure 1.1, we focus on a discrete time system that is consists five parts as shown in Figure 1.2. We study the source model and the channel model in their simplest forms. A source is a series of independent identically distributed random variables on a finite alphabet. A noisy channel is characterized by a probability transition matrix with finite input and finite output alphabets. The encoder is a map from a realization of the source to channel inputs and the decoder is a map from the channel outputs to the reconstruction of the source.

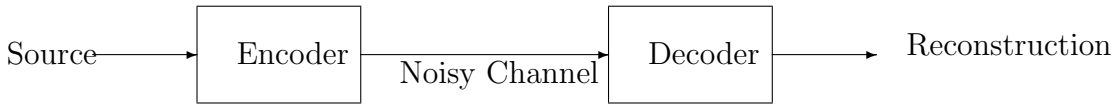

Figure 1.2. Shannon's original problem in [72]

This is the original system setup in Shannon's 1948 paper [72]. One of the most important results in that paper is the *separation* theorem. The separation theorem states that one can reconstruct the source with a high fidelity [1] if and only if the capacity of the channel is at least as large as the entropy of the source. The capacity of the channel is the number of independent *bits* that can be reliably communicated across the noisy channel per channel use. The entropy of the source is the number of *bits* needed on average to describe the source. Both the entropy of the source and the capacity of the channel are defined in terms of *bits*. And hence the separation theorem guarantees the validity of the digital interface shown in Figure 1.3. In this system setup, where the channel coding and source coding are two separate parts, the reconstruction has high fidelity as long as the entropy of the source is less than the capacity of the channel. This theorem, however, does not state that separate source-channel coding is *optimal* under every criteria. For example, the large deviation performance of the joint source channel coding system is strictly better than separate coding even in the block coding case [27]. However, to simplify the problem, we study source coding and channel coding separately. In this thesis, we focus on the source coding problems.

Figure 1.3 is the system we are concerned with. It illustrates the separation of source

[1] For simplicity, we define high fidelity as *lossless* in this section. The *lossy* version, reconstruction under a distortion measure, is discussed in Chapter 3.

coding and channel coding and two other important elements in information theory: channel feedback and side-information of the source. It is well known that feedback does not increase the channel *capacity* [26] for memoryless channels. However, as shown in [61, 47, 85, 70, 67, 68] and numerous other papers in the literature, channel feedback can be used to reduce complexity and increase reliability.

Another important aspect we are going to explore is source side-information. Intuitively speaking, knowing some information about the source *for free* can only help the decoder to reconstruct the source. In the system model shown in Figure 1.3, the source is modeled as iid random variables and the noisy channel is a discrete time memoryless channel, the side-information is another iid random sequence which is correlated with the source in a memoryless fashion. The problem is reduced to a standard source coding problem if both the encoder and the decoder have the side-information. In [79], it is shown that with decoder only side-information the source coding system can operate at the *conditional* entropy rate instead of the generally higher entropy rate of the source.

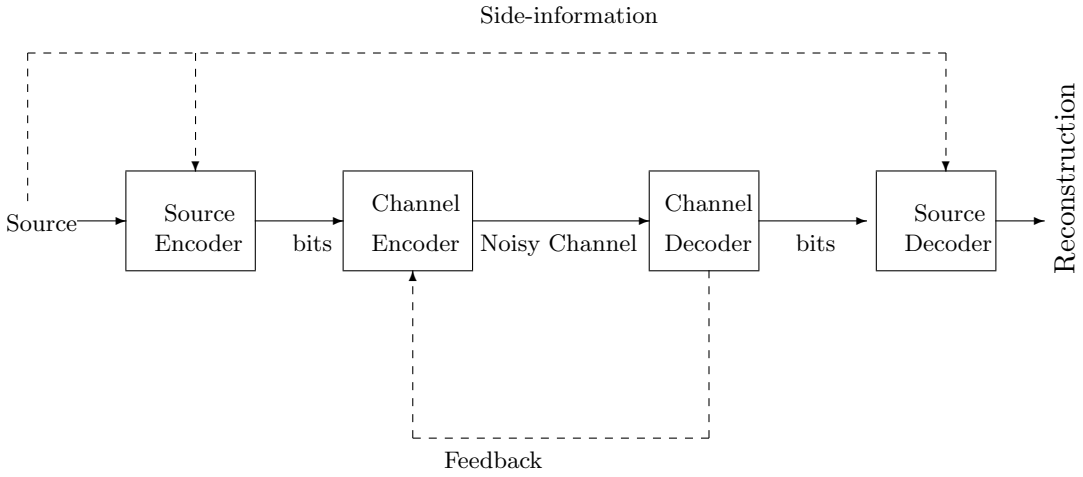

Figure 1.3. Separation, channel feedback and source side-information

By replacing the noisy channel in Figure 1.3 with a noiseless channel through which R bits can be instantaneously communicated, we have a *source* coding system as shown in Figure 1.4. In the main body of this thesis, Chapters 2-4, we study the model shown in Figure 1.4.

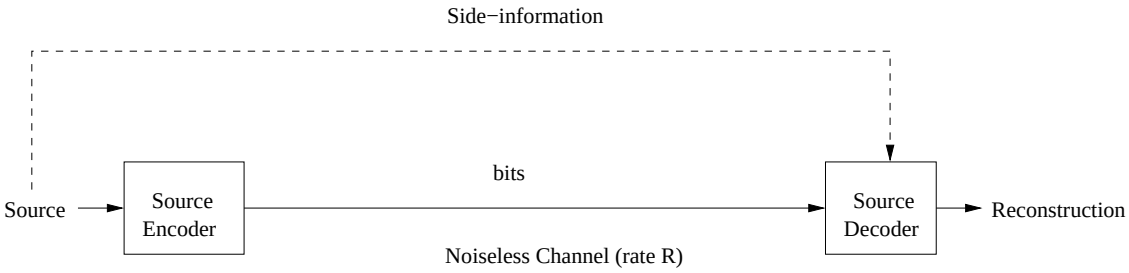

Figure 1.4. Source coding problem

1.2 Information, delay and random arrivals

A very important aspect that is missing in the model shown in Figure 1.3 is the timing of the source symbols and the channel uses. In most information theory problems, the source sequence is assumed to be ready for the source encoder before the beginning of compression. And hence the source encoder has whole knowledge of the entire sequence. Similarly, the channel encoder has the message represented by a binary string before the beginning of communication and thus the channel outputs depend on every bit in that message. This is a reasonable assumption for many applications. For example this is a valid model for the transmission of Kong Zi (Confucius)'s books to another planet. However, there are a wide range of problems that cannot be modeled that way. Recent results in the interaction of information theory and control [69] show that we need to study the whole information system in an *anytime* formulation. It is also shown in that paper that the traditional notion of channel capacity is not enough. On the source coding side, modern communication applications such as video conferencing [54, 48] require short delay between the generation of the source and the consumption by the end user. Instead of transmitting Kong Zi's books, we are facing the problem of transmitting Wang Shuo [84]'s writings to our impatient audience on another planet, while Wang Shuo only types one character per second in his usual unpredictable "I'm your daddy" [78] manner.

In this thesis, we study delay constrained coding for streaming data, where a finite delay is required between the realization time of a source symbol (a random variable in a random sequence) and the consumption time of the source symbol at the sink. Without loss of generality, the source generates *one* source symbol per second and within a finite delay Δ seconds, the decoder consumes that source symbol as shown in Figure 1.5. We assume the source is iid. Implicitly, this model states that the reconstruction of the source symbol at the decoder is only dependent on what the decoder receives until Δ seconds after the realization of that source symbol. This is in line with the notion of *causality* defined in [58]. The formal setup is in Section 2.1.1. In this thesis, we freely interchange "coding with delay" with "delay constrained coding".

Another timing issue is the randomness in the arrival times of the information (source symbols) to the encoder. As shown in Figure 1.6, Wang Shuo types a character in a random fashion. The readers may not care about the random inter-arrivals between his typing. But with a finite delay, a positive rate (protocol information) has to be used to encode the timing information [38]. If the readers care about the arrival time, the encoding of the

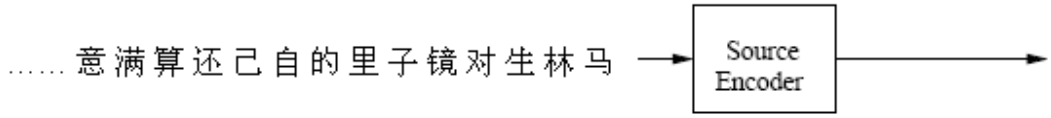

Figure 1.5. Wang Shuo types one character per second while impatient readers want to read his writings within a finite delay. Rightmost character is the first in "I'm your daddy" [78]. Read from right to left.

timing information might pose new challenges to the coding system if the inter-arrival time is unbounded. We only have some partial result for geometric distributed arrival times. We are currently working on Poisson random arrivals but neither of these results is included in this thesis. Another interesting case is when the channel is a timing channel. Then Wang Shuo can send messages by properly choosing his typing time [3].

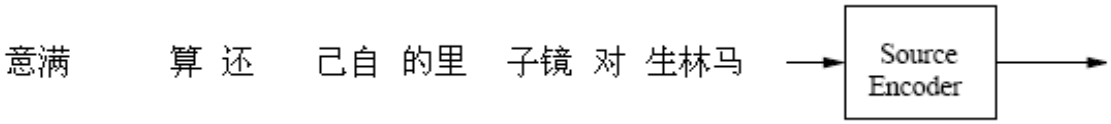

Figure 1.6. Wang Shuo decides to type in a pattern with random inter-character time. Rightmost character is the first in "I'm your daddy" [78]. Read from right to left.

We focus on the case where source symbols come to the encoder at a constant rate (no random arrivals). The problems of interest are summarized in Table 1.1. Each problem is "coded" by a binary string of length six. A fundamental question is how these six elements affect the coding problem. This is not fully understood. A brief discussion is given at the end of the thesis. There are quite a few ($\leq$ 62) interesting problems in the table. For example, the ultimate problem in the table is 111111: "Delay constrained joint source channel Wyner-Ziv coding with feedback and random arrivals". To our knowledge, nobody has *published* anything about this before. There are many other open questions dealing with the timing issues of the communication system as shown in the table. We leave them to future researchers to explore.

	Random arrival	Non-uniform source	Noisy channel	Feedback	Side-information	Distortion
Chapter 2	0	1	0	0	0	0
Chapter 3	0	1	0	0	0	1
Chapter 5	0	1	0	0	1	0
Ongoing work	1	1	0	0	0	0
[15]	0	1	1	0	0	0
[67]	0	0	1	1	0	0
Ultimate	1	1	1	1	1	1

Table 1.1. Delay constrained information theory problems

1.3 Error exponents, focusing operator and bandwidth

From the perspective of *the large deviation principle* [31], an error exponent is the logarithm of the probability of the exceptionally rare behavior of source or channel. This error exponent result is especially useful when the source entropy is much lower than the channel capacity. The channel capacity and the source entropy rate characterize the behavior of the source and channel in the regime of the *central limit theorem*. And hence error exponent results always play second fiddle to channel capacity and source entropy rate. Recently, we studied the *third* order problem— redundancy rate in the block coding setup [21, 19]. This is particularly interesting when the channel capacity is close to the source entropy.

In Shannon's original paper on information theory [72], he studied the source coding and channel coding problem from the perspective of the minimum average number of channel uses to communicate 1 source symbol *reliably*. The setting is asymptotic in nature— long block length, where the number of source symbols is big. Reliable communication means an arbitrary small decoding error. However, it is not clear how fast the error converges to zero with longer block length. Later, researchers studied the convergence rate problem and it was shown that the error converges to zero *exponentially* fast with block length as long as there is some redundancy in the system, i.e. channel capacity is strictly higher than the entropy rate of the source. This exponent is defined as the error exponent, as shown in Figure 1.7. For channel coding, a lower bound and an upper bound on this error exponent are derived in some of the early works [75, 76] and [41]. A very nice upper bound derivation appears in Gallager's technical report [35]. The lossless source coding error exponent is completely identified in [29], the lossy source coding error exponent is studied in [55]. The joint source channel coding error exponent was first studied in [27].

In the delay constrained setup of streaming source coding and channel coding problems, we study the convergence rate of the symbol error probability. As shown in Figure 1.8, the error probabilities go to zero exponentially fast with *delay*. We ask the fundamental question: is *delay* for delay constrained streaming coding the same as the *block length* for classical fixed-length block coding?

In [67], Sahai first answered the question for streaming channel coding. It is shown that without feedback, the delay constrained error exponent does not beat the block coding upper bound. More importantly, it is shown that, with feedback, the delay constrained error exponent is higher than its block coding counterpart as illustrated in Figure 1.9. This new exponent is termed the "focusing" bound.

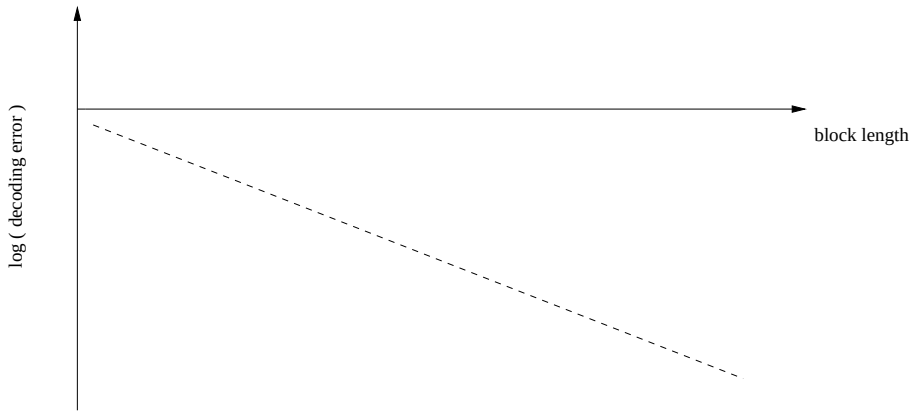

Figure 1.7. Decoding error converges to zero exponentially fast with block length given system redundancy. The slope of the curve is the block coding error exponent.

In this thesis, we answer the question for source coding. For both lossless source coding and lossy source coding with delay, we show that the delay constrained source coding error exponents are higher than their block coding counterparts in Chapter 2 and Chapter 3 respectively. Similar to channel coding, the delay constrained error exponent and the block coding error exponent are connected by a "focusing" operator.

$$E_{delay}(R) = \inf_{\alpha>0} \frac{1}{\alpha} E_{block}((1+\alpha)R) \tag{1.1}$$

where $E_{delay}(R)$ and $E_{block}(R)$ are the rate R error exponents of delay constrained and fixed-length block coding respectively. The conceptual information-theoretic explanation of this operator is that the coding system can borrow some *resources* (channel uses) from the future to deal with the delay constrained *dominant error events*. This is not the case for block coding as for fixed length block coding, the amount of *resources* is given before the realization of the randomness from the source or the channel.

This thesis is an information-theoretic piece of work. But we also *care* about applications (non-asymptotics). Now suppose Wang Shuo's impatient readers want to know what Wang Shuo just typed $\Delta = 20$ seconds ago, but they can tolerate an average $P_e = 0.1\%$ decoding error. A natural system design question is: what is the sufficient and necessary bandwidth R. We give a one line derivation here and will not further study the design issue in this thesis. The error probability decays to zero exponentially with delay:

$$P_e \approx 2^{-\Delta E_{delay}(R)} \tag{1.2}$$

Because $E_{delay}(R)$ is monotonically increasing with R, the sufficient and necessary condition

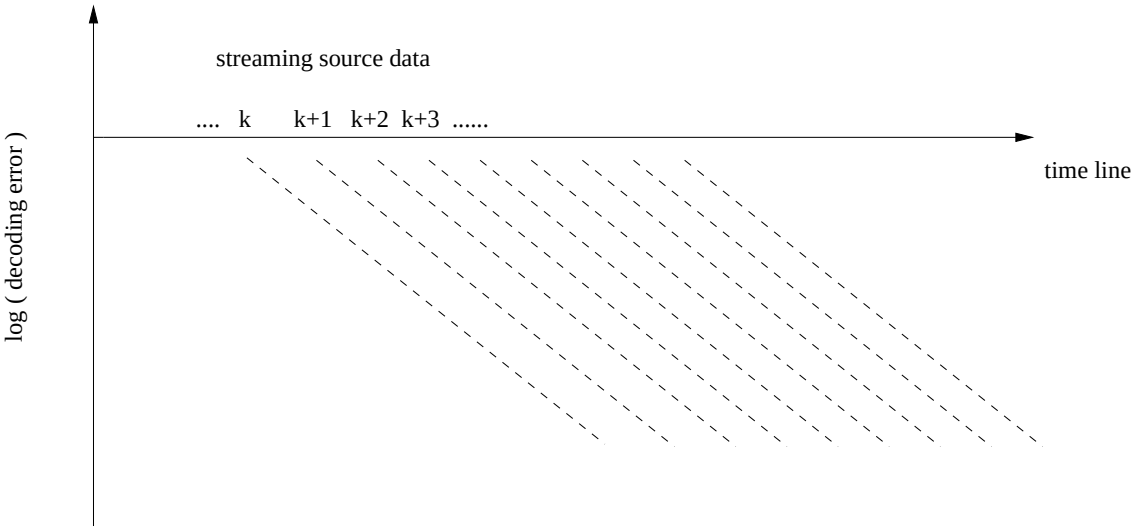

Figure 1.8. Decoding error converges to zero exponentially fast with block length given system redundancy. The slope of the curves are the delay constrained error exponent.

to pleasing our impatient readers is thus:

$$R \approx E_{delay}^{-1}(\frac{1}{\Delta}\log(P_e)) \tag{1.3}$$

An explicit lower bound on R can be trivially derived from (1.3), our recent work on the block coding redundancy rate problems [21, 19] and a manipulation of the focusing operator in (1.1). It is left for future researchers.

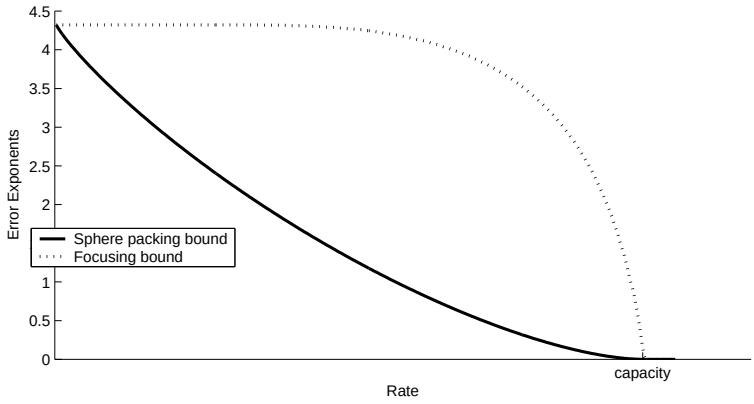

Figure 1.9. Focusing bound vs sphere packing bound for a binary erasure channel with erasure rate 0.05

1.4 Overview of the thesis

As discussed in previous sections, there is a huge number of information-theoretic problems that we are interested in. We only focus on source coding problems with an end to end decoding delay in this thesis. Technical results of the thesis are developed in Chapters 2-5. The structure of every chapter roughly follows the same flow: problem setup → main theorems → examples → proofs → discussion and ideas for future work. Section 2.1 serves as the foundation of the thesis. Formal definitions of streaming data and delay constrained encoder and decoder are introduced. Every chapter is mostly presented in a self-contained fashion.

The organization of the thesis is as follows. We first study the simplest and the most fundamental problem of all, delay constrained lossless source coding in Chapter 2. In Chapter 3, we add a distortion measure to the coding system to explore new aspects of delay constrained problems and give a more general proof. Then in Chapter 4, the achievability (lower bound on the delay constrained error exponent) of the distributed source coding problem is studied, no general upper bound is known yet. However, in Chapter 5, we give an upper bound on the delay constrained error exponent for source coding with decoder side-information, which is a special case of distributed source coding.

1.4.1 Lossless source coding with delay

In Chapter 2, we study the lossless source coding with delay problem. This is the simplest problem of all in terms of problem setup. Some of the most important concepts of this thesis are introduced in this chapter. There is one source that generates one source symbol per second and the encoder can send R bits per second to the decoder. The decoder wants to recover *every* source symbol within a finite delay from when the symbol enters the encoder. We define the delay constrained error exponent $E_s(R)$ as the exponential rate at which the decoding error decays to zero with delay. The delay constrained error exponent is the main object we study in this thesis.

An upper bound on this error exponent is derived by a "focusing" bound argument. The key step is to translate the symbol error with delay to the fixed length block coding error. From there, the classical block coding error exponent [41] result can be borrowed. This bound is shown to be achievable by implementing an optimal universal fixed to variable length encoder together with a FIFO buffer. A similar scheme based on tilted distribution

coding was proposed in [49] when the encoder knows the distribution of the source. A new proof for the tilted distribution based scheme is provided based on the large deviation principle. We analyze the delay constrained error exponent $E_s(R)$ and show that there are several differences between this error exponent and the classical block source coding error exponent in Section 2.5. For example, the derivative of the delay constrained error exponent is positive at the entropy rate instead of 0 for block coding error exponents. Also the delay constrained error exponent approaches infinity at the logarithm of the alphabet, which is not the case for block coding. However, they are connected through the focusing operator defined in (1.1). A similar "focusing" operator was recently observed in streaming channel coding with feedback [67].

For streaming channel coding with delay, the present of feedback is essential for the achievability of the "focusing" bound. But for streaming lossless source coding with delay, no feedback is needed. In order to understand for what kind of information-theoretic problems the focusing operator applies to, we study a more general problem in Chapter 3. For the delay constrained point-to-point source coding problems in Chapters 2 and 3, the focusing bound applies because the encoder has full information of the source. However, for the distributed source coding problems in Chapters 4 and 5 the focusing operator no longer applies because the encoders do not have the full information of source randomness.

1.4.2 Lossy source coding with delay

In Chapter 2, the reconstruction has to be exactly the same as the source or else a decoding error occurs. In Chapter 3, we loosen the requirement for the *exactness* of the reconstruction and study the delay constrained source coding problem under a distortion measure. The source model is the same as that for lossless source coding. We define the delay constrained lossy source coding error exponent as the exponential convergence rate of the *lossy* error probability with delay, where a lossy error occurs if the distortion between a source symbol and its reconstruction is higher than the system requirement.

Technically speaking, lossless source coding can be treated as a special case of lossy source coding by properly defining *error* as a distortion violation. Hence, the results in Chapter 3 can be treated as a natural generalization of Chapter 2. We prove that the delay constrained error exponent and the block coding error exponent for peak distortion measure are connected through the same focusing operator defined in (1.1). The reason is that the rate distortion function under peak distortion is *concave* $\cap$ over the distribution of the

source. This is a property that average distortion measures do not have. The derivations in this chapter are more general than in Chapter 2 since we only use the concavity of the rate distortion function over the distribution. Hence the techniques can be used in other delay constrained coding problems, for example, channel coding with feedback where the variable length code is also concave $\cap$ in the channel behavior.

1.4.3 Distributed lossless source coding with delay

In Chapter 4, we study delay constrained distributed source coding of correlated sources. The block coding counterpart was first studied by Slepian and Wolf in [79], where they determined the *rate region* for two encoders that intend to communicate two correlated sources to one decoder without cooperation.

In Chapter 2, we introduced a sequential random binning scheme that achieves the random coding error exponent. This scheme is not necessary in the point-to-point case because the random coding error exponent is much smaller than the optimal "focusing" bound. In distributed source coding, however, sequential random binning is useful. By using a sequential random binning scheme, we show that a positive delay constrained error exponent can be achieved for *both* sources as long as the rate pair is in the interior of the rate region determined in [79]. The delay constrained error exponents are different from the block coding ones in [39, 29] because the two problems have different *dominant* error events. Similar to fixed length block coding, we show that both maximum-likelihood decoding and universal decoding achieve the same error exponent. This is through an analysis in Appendix G where tilted distributions are used to bridge the two error exponents. Several important Lemmas that are used in other chapters are also proved in Appendix G. The essence of these proofs is to use Lagrange duality to confine the candidate set of distributions of a minimization problem to a one dimensional exponential family.

Unlike what we observed in Chapter 2 and Chapter 3, the delay constrained error exponent is smaller than the fixed length block coding counterpart. Is it because the achievability scheme is suboptimal, or does this new problem have different properties than the point to point source coding problems studied in the previous two chapters? This question leads us to the study of the upper bound for a special case of the delay constrained distributed source coding problem in the next chapter.

Chronologically, this is the first project on delay constrained source coding that I was involved with. Then a post-doc at Cal, Stark Draper, my advisor Anant Sahai and I worked

on this problem from December 2004 to October 2006. Early results were summarized in [32] and a final version has been submitted [12]. Now a professor in Wisconsin, Stark contributed a lot to this project, especially in the universal decoding part. I appreciate his help in the project and introducing me to Imre Csiszár and János Körner's great book [29].

1.4.4 Lossless Source Coding with Decoder Side-Information with delay

What is missing in Chapter 4 is a non-trivial upper bound on the delay constrained error exponents. In Chapter 5, we study one special distributed source coding problem, source coding with decoder side-information, and derive *an* upper bound on the error exponent. This problem is a special case of the problem in Chapter 4 since the decoder side-information can be treated as an encoder with rate higher than the logarithm of the size of the alphabet. It is a well known fact that there is a duality between channel coding and source coding with decoder side-information [2] in the block coding setup. So it is not surprising that we can borrow a delay constrained channel coding technique called the feed-forward decoder that was first developed by Pinsker [62] and recently clarified by Sahai [67] to solve our problem in Chapter 5. However, the lower bound and the upper bound are in general not the same. This leaves a great space for future improvements.

We then derive the error exponent for source coding with both decoder and encoder side-information. This delay constrained error exponent is related to the block coding error exponent by the focusing operator in (1.1). This error exponent is generally strictly higher than the upper bound of the delay constrained error exponent with only decoder side-information. This is similar to the channel coding case, where the delay constrained error exponent is higher with feedback than without feedback. This phenomenon is called "price of ignorance" [18] and is not observed in the block coding context. This shows that in the delay constrained setup, traditional compression first then encryption scheme achieves higher reliability than the novel encryption first then compression scheme developed in [50], although there is no difference in reliability between the two schemes in the block coding setup. The 'price of ignorance" adds to the series of observations that delay is not the same as block length.

Chapter 2

Lossless Source Coding

In this chapter, we begin by reviewing classical results on the error exponents of lossless block source coding. In order to understand the fundamental differences between classical block coding and delay constrained coding[1], we introduce the setup of delay constrained lossless source coding problem and the notion of error exponent with delay constraint. This setup serves as the foundation of the thesis. We then present the main result of this chapter: a tight achievable delay-constrained error exponent for lossless source coding with delay constraints. Some alternative suboptimal coding schemes are also analyzed, especially the sequential random binning scheme which is used as a useful tool in future chapters on distributed lossless source coding.

2.1 Problem Setup and Main Results

Fixed-length lossless block source coding is reviewed in Section A.1 in the appendix. We present our result in the delay constrained setup for streaming lossless source coding.

2.1.1 Source Coding with Delay Constraints

We introduce the delay constrained source coding problem in this section, system model and setup and architecture issues are discussed. These are the basics of the whole thesis.

[1]In this thesis, we freely interchange "coding with delay" with "delay constrained coding", similarly "error exponent with delay" with "delay constrained error exponent".

System Model, Fixed Delay and Delay Universality

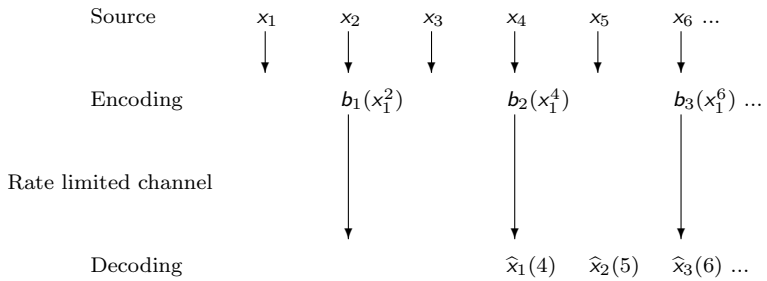

Figure 2.1. Time line of delay constrained source coding: rate $R = \frac{1}{2}$, delay $\Delta = 3$

As shown in Figure 2.1, rather than being known in advance, the source symbols enter the encoder in a streaming fashion. We assume that the discrete memoryless source generates one source symbol x_i per second from a finite alphabet $\mathcal{X}$. Where x_i's are i.i.d from a distribution p_x. Without loss of generality, assume $p_x(x) > 0$, $\forall x \in \mathcal{X}$.

The j^{th} source symbol x_j is not known at the encoder until time j, this is the fundamental difference in the system model from the block source coding setup in Section A.1. The coding system also commits to a rate R. Rate R operation means that the encoder sends 1 bit to the decoder every $\frac{1}{R}$ seconds. It is shown in Proposition 1 and Proposition 2 that we only need to study the problem when the rate R falls in the interval of $[H(p_x), \log|\mathcal{X}|]$.

We define the sequential encoder-decoder pair. This is the coding system we study for the delay constrained setup.

Definition 1 *A fixed-delay Δ sequential encoder-decoder pair $\mathcal{E}, \mathcal{D}$ is a sequence of maps: $\{\mathcal{E}_j\}, j = 1, 2, ...$ and $\{\mathcal{D}_j\}, j = 1, 2, ...$ The outputs of $\mathcal{E}_j$ are the outputs of the encoder $\mathcal{E}$ from time $j-1$ to j.*

$$\mathcal{E}_j : \mathcal{X}^j \longrightarrow \{0,1\}^{\lfloor jR \rfloor - \lfloor (j-1)R \rfloor}$$
$$\mathcal{E}_j(x_1^j) = b_{\lfloor (j-1)R \rfloor + 1}^{\lfloor jR \rfloor}$$

The output of the fixed-delay Δ decoder $\mathcal{D}_j$ is the decoding decision of x_j based on the received

binary bits up to time $j + \Delta$.

$$\mathcal{D}_j : \{0,1\}^{\lfloor (j+\Delta)R \rfloor} \longrightarrow \mathcal{X}$$
$$\mathcal{D}_j(b_1^{\lfloor (j+\Delta)R \rfloor}) = \widehat{x}_j(j+\Delta)$$

Hence $\widehat{x}_j(j+\Delta)$ *is the estimation of* x_j *at time* $j+\Delta$ *and thus there is an end-to-end delay of* Δ *seconds between when* x_j *enters the encoder and when the decoder outputs the estimate of* x_j. *A rate* $R = \frac{1}{2}$, *fixed-delay* $\Delta = 3$, *sequential source coding system is illustrated in Figure 2.1.*

In this thesis, we focus our study on the fixed delay coding problem with a fixed delay Δ. Another interesting problem is the delay-universal coding problem defined as follows.

The outputs of delay-universal decoder $\mathcal{D}_j$ are the decoding decisions of all the arrived source symbols at the encoder by time j based on the received binary bits up to time j.

$$\mathcal{D}_j : \{0,1\}^{\lfloor jR \rfloor} \longrightarrow \mathcal{X}^j$$
$$\mathcal{D}_j(b_1^{\lfloor jR \rfloor}) = \widehat{x}_1^j(j)$$

Where $\widehat{x}_1^j(j)$ is the estimation, at time j, of x_1^j and thus the end-to-end delay of symbol x_i at time j is $j - i$ seconds for $i \leq j$. In a delay-universal scheme, the decoder emits revised estimates for all source symbols so far. This coding system is illustrated in Figure 2.2.

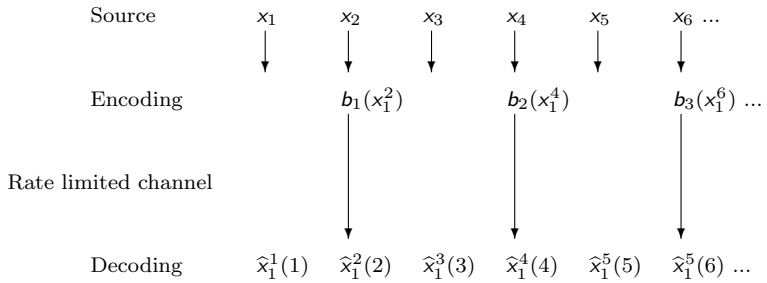

Figure 2.2. Time line of delay universal source coding: rate $R = \frac{1}{2}$

Without giving the proof, we state that all the error exponent results in this thesis for fixed-delay problems also apply to delay universal problems.

Symbol Error Probability, Delay and Error Exponent

For a streaming source coding system, a source symbol that enters the system earlier could get a higher decoding accuracy than a symbol that enters the system later. This is because the earlier source symbol could use more *resource*— noiseless channel uses. This is in contrast to block coding systems where *all* source symbols share the same noiseless channel uses. Thus for delay constrained source coding, it is important to study the symbol decoding error probability instead of the block coding error probability.

Definition 2 *A delay constrained error exponent $E_s(R)$ is said to be achievable if and only if for all $\epsilon > 0$, there exists $K < \infty$, $\forall \Delta > 0$, $\exists$ fixed-delay Δ encoder decoder pairs $\mathcal{E}$, $\mathcal{D}$, s.t. $\forall i$:*

$$\Pr[x_i \neq \widehat{x}_i(i+\Delta)] \leq K 2^{-\Delta(E_s(R)-\epsilon)}$$

Note: the order of the conditions of this definition is extremely important. Simplistically speaking, a delay constrained error exponent is said to be achievable (meaning the symbol error decays exponentially with delay with the claimed exponent for all symbols), if it is achievable for all fixed-delays.

We have the following two propositions stating that the error exponent is only interesting if $R \in [H(p_x), \log|\mathcal{X}|]$. First we show that the error exponent is only interesting if the rate $R \leq \log|\mathcal{X}|$.

Proposition 1 *$E_s(R)$ is not well defined[2] if $R > \log|\mathcal{X}|$*

Proof: Suppose $R > \log|\mathcal{X}|$, for integer N large enough, we have

$$2^{\lfloor NR-2 \rfloor} > 2^{N \log|\mathcal{X}|} = |\mathcal{X}|^N.$$

Now we construct a very simple block coding system as follows.

The encoder first queues up N source symbols from time $(k-1)N+1$ to kN, $k = 1, 2, ...$, those N symbols are $x_{(k-1)N+1}, ...x_{kN}$. Now since $2^{\lfloor NR-2 \rfloor} > |\mathcal{X}|^N$, use an injective map $\mathcal{E}$ from $\mathcal{X}^N$ to $\{1, 2, 3, ..., 2^{\lfloor NR-2 \rfloor}\}$. Notice that the encoder can send at least $\lfloor NR-2 \rfloor$ bits in the time interval $(kN, (k+1)N)$. So the encoder can send $\mathcal{E}(x_{(k-1)N+1}, ...x_{kN})$ to the decoder within time interval $(kN, (k+1)N)$. Thus the decoding error for source symbols

[2]Less mathematically strictly speaking, the error exponent $E_s(R)$ is infinite if $R > \log|\mathcal{X}|$.

$x_{(k-1)N+1}, ...x_{kN}$ is zero at time $(k+1)N$. This is true for all $k = 1, 2...$ So the error probability for any source symbol x_n at time $n+\Delta$ is zero for any $\Delta \geq 2N$. Thus the error exponent $E_s(R)$ for $R > \log|\mathcal{X}|$ is not defined because $\log 0$ is not defined. Or conceptually we say the error exponent is *infinite* for $R > \log|\mathcal{X}|$. □

The above lemma is for the delay constrained source coding problem in this chapter. But it should be obvious that similar results hold for the other source coding problems discussed in future chapters. That is, if the rate of the system is above the logarithm of the alphabet size, an *almost* instantaneous error free coding is possible. This implies that the error exponent is not well defined in that case.

Secondly, the delay constrained error exponent is zero, i.e. the error probability does not *universally* converge to zero if the rate is below the entropy rate of the source. This result should not be surprising given that it is also true for block coding. However, for the completeness of the thesis, we give a proof. The proof here is to translate a delay constrained problem to a block coding problem, then by using the classical block coding error probability result we lower bound the symbol error probability, and thus give a lower upper bound on the delay constrained error exponent.

Proposition 2 *$E_s(R) = 0$ if $R < H(p_{\mathsf{x}})$.*

Proof: We prove the lemma by contradiction. Suppose that for some source p_{x} the delay constrained source coding error exponent $E_s(R) > 0$ for some $R < H(p_{\mathsf{x}})$. Then from Definition 2 we know that for any $\epsilon > 0$, there exists $K < \infty$, such that for all Δ, there exists a delay constrained source coding system $\mathcal{E}$, $\mathcal{D}$, and

$$\Pr[\mathsf{x}_n \neq \widehat{\mathsf{x}}_n(n+\Delta)] \leq K 2^{-\Delta(E_s(R)-\epsilon)} \text{ for all } n. \tag{2.1}$$

Notice that (2.1) is true for all n, Δ. We pick n and Δ that are *big enough* as needed.

Now we can design a block coding scheme of block length n, the encoder $\mathcal{E}$ is derived from the delay constrained source coding system $\mathcal{D}_i$, where the output of the block encoder is the same as the accumulate of the delay constrained encoders.

$$\mathcal{E}(x_1^n) = \big(\mathcal{E}_1(x_1), \mathcal{E}_2(x_1^2), ..., \mathcal{E}_{n+\Delta}(x_1^{n+\Delta})\big) = b_1^{\lfloor (n+\Delta)R \rfloor} \tag{2.2}$$

Notice that some of the output of the encoders $\mathcal{E}_i(x_1^i)$ can be empty.

The decoder part is similar. The block code decoder uses the output of all the delay constrained decoders up to time $n+\Delta$

$$\mathcal{D}(b_1^{(n+\Delta)R}) = \big(\widehat{x}_1(1+\Delta), \widehat{x}_2(2+\Delta), ..., \widehat{x}_n(n+\Delta)\big) \triangleq \widehat{x}_1^n \tag{2.3}$$

Now we look at the block error of the block coding system built from the delay constrained source coding system. First, we fix the ratio of n and Δ at $\Delta = \alpha n$, we will choose a sufficient large n and a small enough α to construct the contradiction.

$$\begin{aligned} \Pr[x_1^n \neq \widehat{x}_1^n] &\leq \sum_{i=1}^{n} \Pr[x_i \neq \widehat{x}_i] \\ &\leq \sum_{i=1}^{n} K 2^{-\Delta(E-\epsilon)} \\ &\leq nK 2^{-\alpha n(E-\epsilon)} \end{aligned} \tag{2.4}$$

However, from the classical block source coding theorem in [72], we know that

$$\Pr[x_1^n \neq \widehat{x}_1^n] \geq 0.5 \tag{2.5}$$

if the effective rate of the block coding system $\frac{n+\Delta}{n}R$ is smaller than the entropy rate of the source $H(p_x)$. This only requires n be much larger than Δ, such that $\frac{n+\Delta}{n}R < H(p_x)$.

Now combining (2.4) and (2.5), we have

$$nK 2^{-\alpha n(E-\epsilon)} \geq 0.5 \tag{2.6}$$

Notice that $E > \epsilon$, $K < \infty$ is a constant, and $\alpha > 0$ is also a constant, hence there exists n big enough such that $nK 2^{-\alpha n(E-\epsilon)} < 0.5$. This gives the contradiction we need. The lemma is proved. □

These two lemmas tell us that the delay constrained error exponent $E_s(R)$ is only interesting for $R \in [H(p_x), \log|\mathcal{X}|]$.

In the proof of Proposition 2, we constructed a block coding system from a delay constrained source coding system with some delay performance and then build the contradiction. In this thesis, we also use this technique to show other theorems. We borrow the classical block coding results to serve as the building blocks of the delay constrained information theory.

The above two lemmas are for the delay constrained lossless source coding problem in this chapter. But it should be obvious that similar result holds for other source coding problems discussed in future chapters. That is, delay constrained error exponents are *zero*, or the error probabilities do not converge to zero universally if the rate is lower than the relevant entropy. And if the rate is above the logarithm of the alphabet size of the source, any exponent is achievable.

2.1.2 Main result of Chapter 2: Lossless Source Coding Error Exponent with Delay

Following the definition of the delay-constrained error exponent for lossless source coding in Definition 2, we have the following theorem which describes the convergence rate of the symbol-wise error probability of lossless source coding with delay problem.

Theorem 1 *Delay constrained lossless source coding error exponent: For source $x \sim p_x$, the delay constrained source coding error defined in Definition 2 is*

$$E_s(R) = \inf_{\alpha>0} \frac{1}{\alpha} E_{s,b}((\alpha+1)R) \tag{2.7}$$

Where $E_{s,b}(R)$ is the block source coding error exponent [29] defined in (A.4). This error exponent $E_s(R)$ is both achievable and an upper bound, hence our result is complete.

Recall the definition of delay constrained error exponent, for all $\epsilon > 0$, there exists a finite constant K, s.t. for all i, Δ,

$$\Pr[x_i \neq \widehat{x}_i(i+\Delta)] \leq K 2^{-\Delta(E_s(R)-\epsilon)}$$

The result has two parts. First, it states that there exists a coding scheme, such that the error exponent $E_s(R)$ can be achieved in a universal setup. This is summarized in Proposition 5 in Section 2.4.1. Second, there is no coding scheme can achieve better delay constrained error exponent than $E_s(R)$. This is summarized in Proposition 6 in Section 2.4.2.

Before showing the proof of Theorem 1, we give some numerical results and discuss some other coding schemes in the next two sections.

2.2 Numerical Results

In this section we evaluate the delay constrained performance for different coding schemes via an example. For a simple source x with alphabet size 3, $\mathcal{X} = \{A, B, C\}$ and the following distribution

$$p_x(A) = a \quad p_x(B) = \frac{1-a}{2} \quad p_x(C) = \frac{1-a}{2}$$

Where $a \in [0, 1]$, in this section we set $a = 0.65$.

2.2.1 Comparison of error exponents

The error exponents for both block and delay constrained source coding predict the asymptotic performance of different source coding systems when the delay is long. We plot the delay constrained error exponent $E_s(R)$, the block source coding error exponent $E_{s,b}(R)$ and the random coding error exponent $E_r(R)$ in Figure 2.3. As shown in Theorem 1 and Theorem 9, these error exponents tell the asymptotic performance of the two delay constrained source coding systems. We give a simple streaming prefix-free code at rate $\frac{3}{2}$for this source. It will be shown that, although this prefix-free coding is suboptimal as the error exponent is much smaller than $E_s(\frac{3}{2})$, its error exponent is much larger than $E_r(\frac{3}{2})$ which is equal to $E_{s,b}(\frac{3}{2})$ in this case. We give the details of the streaming prefix-free coding system and analyze its decoding error probability in Section 2.2.2.

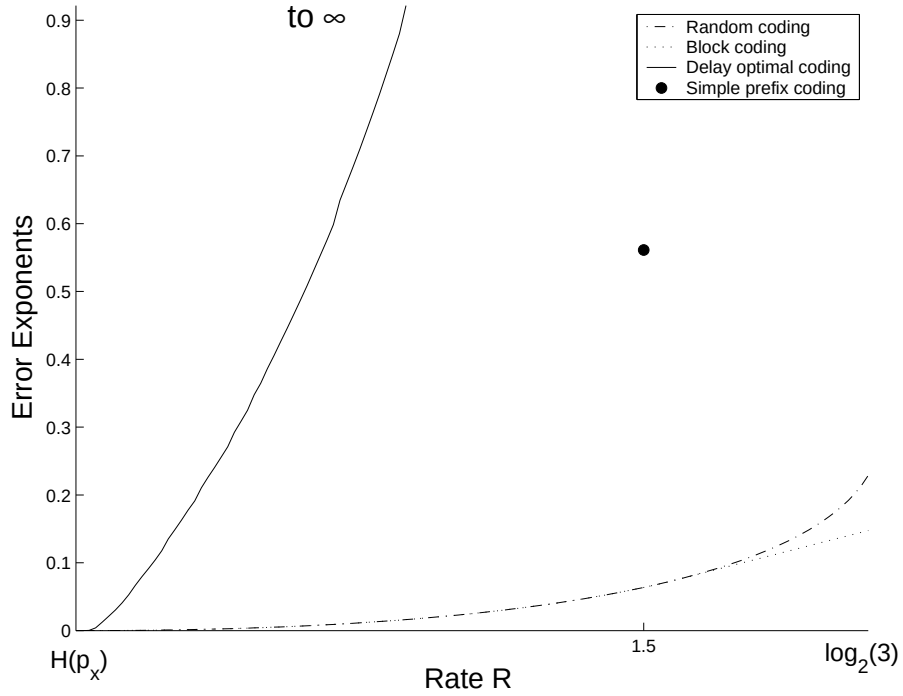

Figure 2.3. Different source coding error exponents: delay constraint error exponent $E_s(R)$, block source coding error exponent $E_{s,b}(R)$, random coding error exponent $E_r(R)$

In Figure 2.4, we plot the ratio of the optimal delay constrained error exponent over the block coding error exponent. The ratio tells to achieve the same error probability, how many times longer the delay has to be for the block coding system that is studied in Section 2.3.1. The smallest ratio is around 52 at rate around 1.45.

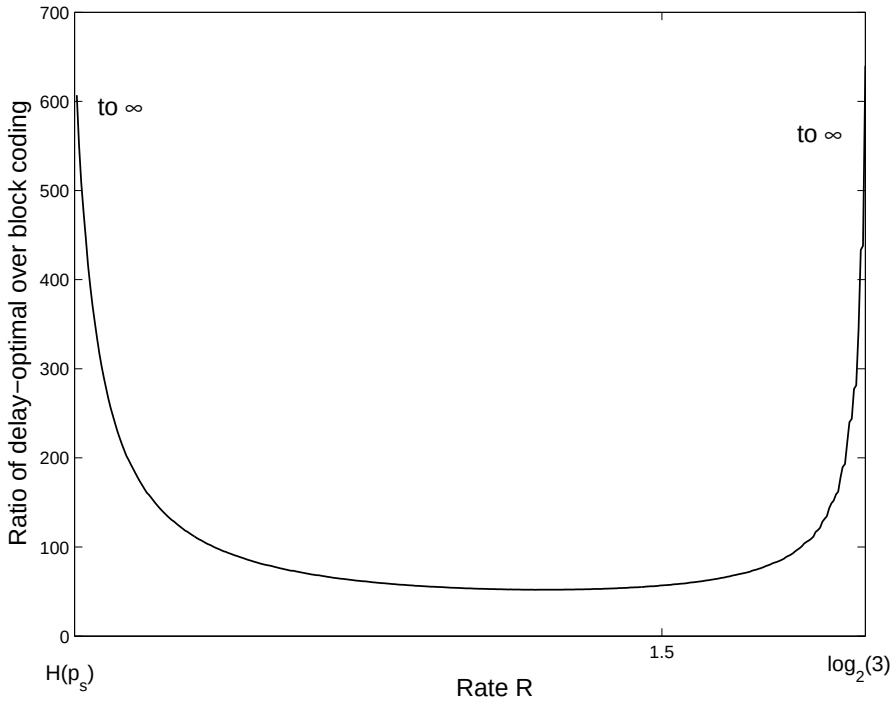

Figure 2.4. Ratio of delay optimal error exponent $E_s(R)$ over block source coding error exponent $E_{s,b}(R)$,

2.2.2 Non-asymptotic results: prefix free coding of length 2 and queueing delay

In this section we show a very simple coding scheme which use a prefix-free code [26] instead of the asymptotical optimal universal code studied in Section 2.4.1. The length two prefix-free code for source x we are going to use is:

$$AA \to 0$$
$$AB \to 1000 \quad AC \to 1001 \quad BA \to 1010 \quad BB \to 1011$$
$$BC \to 1100 \quad CA \to 1101 \quad CB \to 1110 \quad CC \to 1111$$

This prefix-free code is obviously sub-optimal because the code length is only 2, also the code length is not adapted to the distribution. However, it will be shown that even this obviously non optimal code outperforms the block coding schemes in the delay constrained setups when a is not too small. We analyze the performance of the streaming prefix-free coding system at $R = \frac{3}{2}$. That is, the source generates 1 symbol per second, while 3 bits can be sent through the rate constrained channel every 2 seconds. The prefix-free stream coding

system like a FIFO queueing system with infinite buffer which is similar to the problem studied in [49]. The prefix-free stream coding system is illustrated in Figure 2.5. Following the prefix code defined earlier, the prefix-free encoder group two source symbols x_{2k-1}, x_{2k} together at time $2k$, $k = 1, 2, ...$ into the prefix-free code. the length of the codeword is either 1 or 4. The buffer is drained out by 3 bits per 2 seconds. Write the number of bits in the buffer as B_k at time $2k$. Every two seconds, the number of bits B_k in the buffer either goes down by 2 if $x_{2k-1}, x_{2k} = AA$ or goes up by 1 if $x_{2k-1}x_{2k} \neq AA$. 3 bits are drained out every 2 seconds from the FIFO queue, notice that if the queue is empty, the encoder can send random bits through the channel without causing confusion at the decoder because the source generates 1 source symbol per second.

Source	AA	AB	BA	CC	AA	CA	AA	AA	AA	CB	AA	CC	
Prefix code	0	1000	1010	1111	0	1101	0	0	0	1110	0	1111	
Buffer	/	/	0	10	111	0	01	/	/	/	0	/	1
Rate R bit−stream	***	0**	100	010	101	111	011	010	0**	0**	111	00*	111
Decision		AA		AB	BA	CC	AA	CA AA	AA	AA		CB AA	

Figure 2.5. Streaming prefix-free coding system (/ indicates empty queue, * indicates meaningless random bits)

Clearly $B_k, k = 1, 2, ...$ form a Markov chain with following transition matrix: $B_k = B_{k-1} + 1$ with probability $1 - a^2$, $B_k = B_{k-1} - 2$ with probability a^2, notice that $B_k \geq 0$ thus the boundary conditions. We have the state transition graph in Figure 2.6. For this Markov chain, we can easily derive the stationary distribution (if it exists) [33].

$$\mu_k = L\big(\frac{-1 + \sqrt{1 + \frac{4(1-q)}{q}}}{2}\big)^k \tag{2.8}$$

Where $q = a^2$ and L is the normalizer, notice that μ_k is a geometric series and the stationary distribution exists as long as the geometric series goes to 0 as index goes to infinity, i.e. $4\frac{1-q}{q} < 8$ or equivalently $q > \frac{1}{3}$. In this example, we have $a = 0.65$, thus $q = a^2 = 0.4225 > \frac{1}{3}$, thus the stationary distribution μ_k exists.

For the above simple prefix-free coding system, we can easily derive the decoding error for symbols x_{2k-1}, x_{2k} at time $2k + \Delta - 1$, thus the effective delay for x_{2k-1} is Δ. The decoding error can only happen if at time $2k + \Delta - 1$, at least one bits of the prefix-free

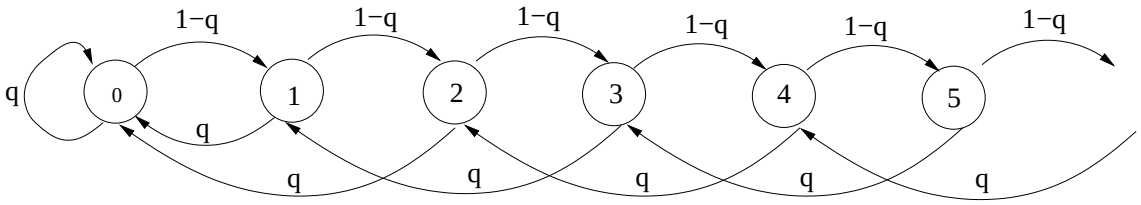

Figure 2.6. Transition graph of random walk B_k for source distribution $\{a, \frac{1-a}{2}, \frac{1-a}{2}\}$, $q = a^2$ is the probability that B_k goes down by 2.

code describing x_{2k-1}, x_{2k} are still in the queue. This implies that the number of bits in the buffer at time $2k$, B_k, is larger than

$$\lfloor \frac{3}{2}(\Delta - 1) \rfloor - l(x_{2k-1}, x_{2k})$$

where $l(x_{2k-1}, x_{2k})$ is the length of the prefix-free code for x_{2k-1}, x_{2k}, thus it is 1 with probability $q = a^2$ and it is 4 with probability $1 - q = 1 - a^2$. Notice that the length of the prefix-free code for x_{2k-1}, x_{2k} is independent with B_k, we have the following upper bound on the error probability of decoding with delay Δ when the system is at stationary state:

$$\begin{aligned}
\Pr[\hat{x}_{2k-1} \neq x_{2k-1}] &\leq \Pr[l(x_{2k-1}, x_{2k}) = 1] \Pr[B_k > \lfloor \frac{3}{2}(\Delta - 1) \rfloor - 1] \\
&\quad \Pr[l(x_{2k-1}, x_{2k}) = 4] \Pr[B_k > \lfloor \frac{3}{2}(\Delta - 1) \rfloor - 4] \\
&= q \sum_{j=\lfloor \frac{3}{2}(\Delta-1) \rfloor}^{\infty} \mu_j + (1-q) \sum_{j=\lfloor \frac{3}{2}(\Delta-1) \rfloor - 3}^{\infty} \mu_j \\
&= G\left(\frac{-1 + \sqrt{1 + \frac{4(1-q)}{q}}}{2}\right)^{\lfloor \frac{3}{2}(\Delta-1) \rfloor - 3}
\end{aligned}$$

The last line is by substituting in (2.8) for stationary distribution μ_j. Where G is a constant, we omit the detail expression of G here. Following the last line, the error exponent for this prefix-free coding system is obviously

$$\frac{3}{2} \log \left(\frac{-1 + \sqrt{1 + \frac{4(1-q)}{q}}}{2}\right)$$

With the above evaluation, we now compare three different coding schemes in the non-asymptotic setups. In Figure 2.7, the error probability vs delay curves are plotted for three different coding schemes. First we plot the causal random coding error probability. Shown in Figure 2.3, at $R = \frac{3}{2}$, random coding error exponent $E_r(R)$ is the same as the block

coding error exponent $E_{s,b}(R)$. The block coding curve is for a so called *simplex* coding scheme, where the encoder first queues up $\frac{\Delta}{2}$ symbols, encode them into a length $\frac{\Delta}{2}R$ binary sequence and use the next $\frac{\Delta}{2}$ seconds to transmit the message. This coding scheme gives an error exponent $\frac{E_{s,b}(R)}{2}$. As can be seen in Figure 2.3, the slope of the *simplex* block coding is roughly half of the block source coding's slope.

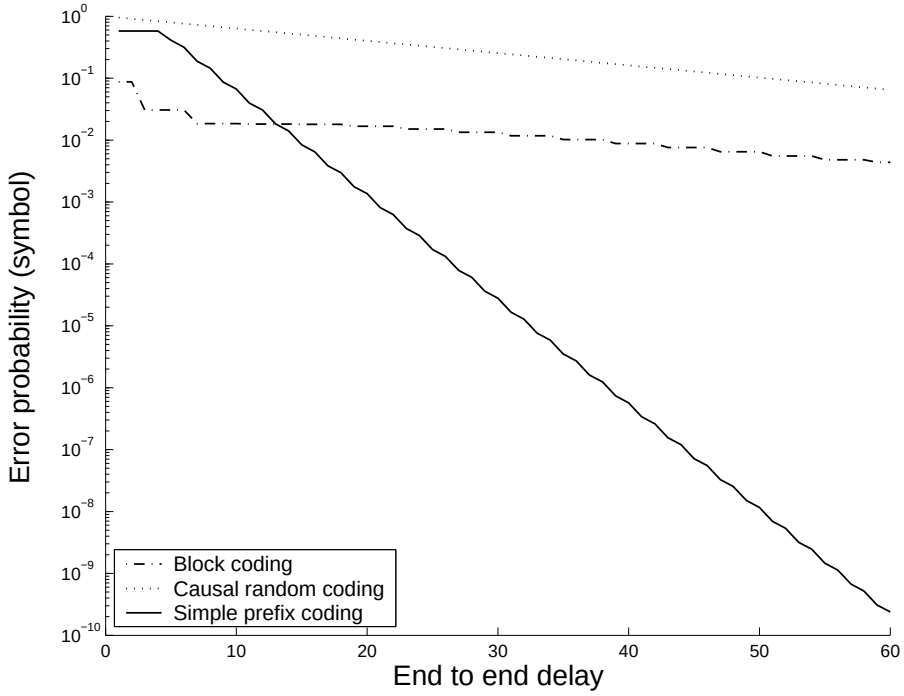

Figure 2.7. Error probability vs delay (non-asymptotic results)

The slope of these curves in Figure 2.7 indicates how fast the error probability goes to zero with delay, i.e. the error exponents. Although smaller than the delay optimal error exponent $E_s(R)$, the simple prefix-free coding has a much higher error exponent than both both random coding and the *simplex* block coding error exponent. A trivial calculation tells us that in order to get 10^{-6} symbol error probability, the delay requirement for delay optimal coding is $\sim$ 40, for causal random coding is around $\sim$ 303, for *simplex* block coding is at around$\sim$ 374. Here we run a linear regression on the data: $y_\Delta = \log_{10} P_e(\Delta)$, $x_\Delta = \Delta$ as shown in Figure 2.3 from $\Delta = 80$ to $\Delta = 100$. Then we extrapolate the Δ, s.t. $\log_{10} P_e(\Delta) = -6$. Thus we see a major delay performance improvement for delay optimal source coding.

2.3 First attempt: some suboptimal coding schemes

On our way to find the optimal delay constrained performance, we first studied the classical block coding's delay constrained performance in Section 2.3.1, the random tree code later discussed in Section 2.3.3 and several other known coding systems. These coding systems in the next three subsections are shown to be suboptimal in the sense that they all achieve strictly smaller delay constrained error exponents than that in Theorem 1. This section is by no means an exhaustive study of the delay constrained performance of all possible coding systems. Our intention is to look at the obvious candidates and help readers to understand the key issues with delay constrained coding.

2.3.1 Block Source Coding's Delay Performance

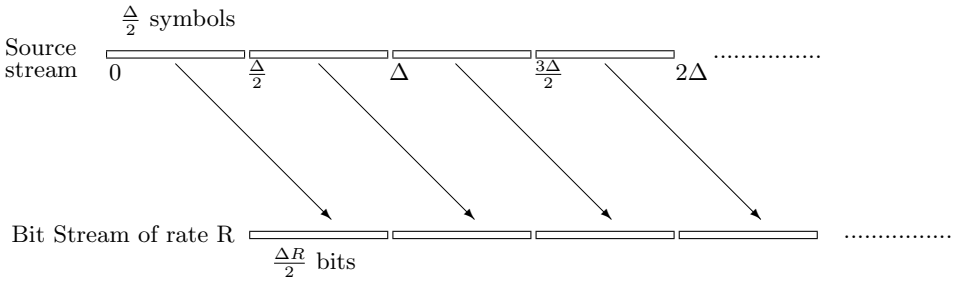

Figure 2.8. Block coding's delay performance

We analyze the delay constrained performance for traditional block source coding systems. As shown in Figure 2.8, the encoder first encode $\frac{\Delta}{2}$ source symbols, during a period of $\frac{\Delta}{2}$ seconds, into a block of binary bits. During the next $\frac{\Delta}{2}$ seconds, the encoder uses all the bandwidth to transmit the binary bits through the rate R noiseless channel. After all the binary bits are received by the decoder at time Δ, the effective coding system is a length $\frac{\Delta}{2}$, rate R block source coding system. Thus the error probability of each symbol is lower bounded by $2^{-\frac{\Delta}{2}E_{s,b}(R)}$ as shown in Theorem 9. The delay constrained error exponent for source symbol x_1 is

$$\begin{aligned} E &= -\frac{1}{\Delta}\log\{2^{-\frac{\Delta}{2}E_{s,b}(R)}\} \\ &= \frac{E_{s,b}(R)}{2} \end{aligned} \tag{2.9}$$

This is half of the block coding error exponent. Another problem of this coding scheme

is that the encoder is delay-specific as Δ is a parameter in the infrastructure. While the optimal coding scheme as shown in Section 2.4.1 is not.

2.3.2 Error-free Coding with Queueing Delay

For both the block coding in the previous section and the sequential random coding in Section 2.3.3, the coding system commits to some none-zero error probability for symbol x_i at time $i + \Delta$, for any i, Δ pair as long as the rate R is smaller than $\log |\mathcal{X}|$. The next two coding schemes are different in the sense that the error probability for any source symbol x_i is going to be exactly zero after some finite variant delay $\Delta_i(x_1^\infty)$. These encoding schemes have two parts, first part is a zero-error fixed or variable to variable length source encoder, the second part is a FIFO (first in first out) buffer. The queue in the buffer is drained out by a constant rate R bits per second. If the queue is empty, the encoder buffer simply send arbitrary bits through the noiseless channel. The decoder, knowing the rate of the source, simply discards the arbitrary gibberish bits added by the encoder buffer and now the the task is only to decode the source symbols by the received bits that describe the source stream. This coding scheme makes no decoding error on x_i as long as the bits that describe x_i are all received by the decoder. Thus the symbol error only occurs if there are too many bits are used to describe the source block that x_i is in and some previous source symbols. This class of coding schemes are illustrated in Figure 2.9. The error free code can be either variable to fixed length coding or variable to variable length coding.

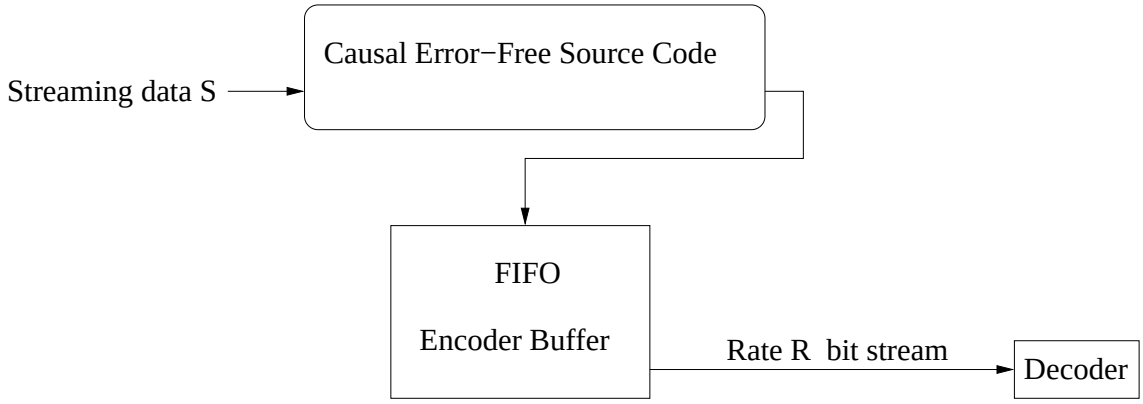

Figure 2.9. Error free source coding for a fixed-rate system

Prefix-free Entropy Coding with Queueing Delay

Entropy coding [26] is a simple source coding scheme, where the code length of a source symbol is roughly inversely proportional to the logarithm of the probability of that source symbol. For example, Huffman code and arithmetic code are two entropy code that are thoroughly studied in [26]. An important property of the entropy coding is that the average code length per source symbol is the entropy rate $H(p_{\mathsf{x}})$ of the source x if the code block is long enough. The code length is independent of the rate R. We argue that the simple prefix-free entropy coding with queueing delay scheme is not *always* optimal by the following simple example. Consider a binary source $\mathcal{S} = \{a, b, c, d\}$ with distribution $p_{\mathsf{x}}(a) = 0.91$, $p_{\mathsf{x}}(b) = p_{\mathsf{x}}(c) = p_{\mathsf{x}}(d) = 0.03$ and the rate of the system is $R = 2$. Now obviously the best coding scheme is *not to code* by just sending the uncompressed binary expressions of the source across the rate $R = 2$ system. The system achieves zero error for every source symbol x_i *instantaneously*. Since rate R match well with the alphabet size $\log |\mathcal{X}|$, philosophically speaking, this statement is the same to Michael Gastpar's PhD thesis [43] in which the Gaussian source is matched well with the Gaussian channel. The delay constrained error exponent is thus infinity or strictly speaking not well defined. One length-2 optimal prefix-free code of this source is $\mathcal{E}(a) = 0$, $\mathcal{E}(b) = 10$, $\mathcal{E}(c) = 110$, $\mathcal{E}(d) = 111$. And the coding system using this prefix-free code or entropy codes of any length clearly does not achieve zero error probability for any delay Δ for source symbol x_i if i is large enough. The error occurs if the source symbols before i are atypical. For example, a long string of c's. This case is thoroughly analyzed in Section 2.2.2.

We show in Section 2.4.1 that a simple universal coding with queueing delay scheme is indeed optimal. This scheme is inspired by Jelinek's non-universal coding scheme in [49]. We call Jelinek's scheme *tilted entropy coding* because the new codes are entropy codes for a *tilted distribution* generated from the original distribution. We give a simplified proof of Jelinek's result in Appendix B using modern large deviation techniques.

LZ78 Coding with Queueing Delay

In this subsection, we argue that the standard LZ78 coding is not suitable for delay constrained setup due to the nature of its dictionary construction.

In their seminal paper [88], Lempel and Ziv proposed a dictionary based sequential universal lossless data compression algorithm. This coding scheme is causal in nature.

However, the standard LZ78 coding scheme described in [88] and later popularized by Welch in [83] does not achieve any positive delay constrained error exponent. The reason why seemingly causal LZW coding scheme does not achieve any positive delay constrained error exponent is rooted in the incremental nature of the dictionary. If the dictionary is infinite as original designed by Lempel and Ziv, then for a memoryless source and a fixed delay Δ, every word with length 2Δ will be in the dictionary with high probability at sufficiently large time t. Thus the encoding delay at the LZW encoder is at least 2Δ for any source symbol x_i, $i > t$. According to the definition of delay constrained error exponent in Definition 2, the achievable error exponent is 0 for such a Lempel-Ziv coding system.

On the other hand, with a finite dictionary, the dictionary mismatches with the source statistics if the early source symbols are *atypical*. This mis-match occurs with a positive probability and thus the encoding delay for later source symbols is big and thus no positive delay constrained error exponent can be achieved. *Note:* we only discuss the popular LZ78 algorithm here. There are numerous other forms of universal Lempel-Ziv coding, the first one in [87]. The delay constrained performance for such coding schemes are left for future studies.

2.3.3 Sequential Random Binning

In [12] and [32], we proposed a sequential random binning scheme for delay constrained source coding systems. It is the source-coding counterpart to tree and convolutional codes used for channel coding [34]. This sequential random binning scheme follows our definition of delay constrained source coding system in Definition 1 with additional common randomness accessible to both the encoder(s) and decoder. The encoder is universal in nature that it is the same for any source distribution p_x meanwhile the decoder can be ML decoder or universal decoder which will be discussed in great details later. We define the sequential random binning scheme as follows.

Definition 3 *A randomized sequential encoder-decoder pair (a random binning scheme) $\mathcal{E}, \mathcal{D}$ is a sequence of maps: $\{\mathcal{E}_j\}, j = 1, 2, ...$ and $\{\mathcal{D}_j\}, j = 1, 2, ...$ The outputs of $\mathcal{E}_j$ are the outputs of the encoder $\mathcal{E}$ from time $j-1$ to j.*

$$\mathcal{E}_j : \mathcal{X}^j \longrightarrow \{0,1\}^{\lfloor jR \rfloor - \lfloor (j-1)R \rfloor}$$

$$\mathcal{E}_j(x_1^j) = b_{\lfloor (j-1)R \rfloor + 1}^{\lfloor jR \rfloor}$$

The output of the fixed-delay Δ decoder $\mathcal{D}_j$ is the decoding decision of x_j based on the received binary bits up to time $j+\Delta$.

$$\mathcal{D}_j : \{0,1\}^{\lfloor (j+\Delta)R \rfloor} \longrightarrow \mathcal{X}$$
$$\mathcal{D}_j(b_1^{\lfloor (j+\Delta)R \rfloor}) = \widehat{x}_j(j+\Delta)$$

Where $\widehat{x}_j(j+\Delta)$ is the estimation of x_j at time $j+\Delta$ and thus has end-to-end delay of Δ seconds. Common randomness, shared between encoder(s) and decoder(s), is assumed. Availability to common randomness is a common assumption in information theory community. Common randomness can be generated from correlated data [63]. This is not the main topic of this thesis in which we assume enough common randomness. Interested readers may read [56, 1]. This allows us to randomize the mappings independent of the source sequence. In this thesis, we only need pair-wise independence, formally, for all i, n:

$$\Pr[\mathcal{E}(x_1^i x_{i+1}^n) = \mathcal{E}(x_1^i \tilde{x}_{i+1}^n)] = 2^{-(\lfloor nR \rfloor - \lfloor iR \rfloor)} \leq 2 \times 2^{-(n-i)R} \tag{2.10}$$

for all $x_{i+1} \neq \tilde{x}_{i+1}$

(2.10) deserves more staring at. The number of output bits of the encoder for source symbols from time i to time n is $\lfloor nR \rfloor - \lfloor iR \rfloor$, so (2.10) means that the chance that the binary representations of two length-n source strings which diverge at time i is $2^{-(\lfloor nR \rfloor - \lfloor iR \rfloor)}$. We define *bins* as follows. A bin is a set of source strings that share the same binary representations:

$$\mathcal{B}_x(x_1^n) = \{\tilde{x}_1^n \in \mathcal{S}^n : \mathcal{E}(\tilde{x}_1^n) = \mathcal{E}(x_1^n)\} \tag{2.11}$$

With the notion of bins, (2.10) is equivalent to the following equality for $x_{i+1} \neq \tilde{x}_{i+1}$:

$$\Pr[\mathcal{E}(x_1^i \tilde{x}_{i+1}^n) \in \mathcal{B}_x(x_1^i x_{i+1}^n)] = 2^{-(\lfloor nR \rfloor - \lfloor iR \rfloor)} \leq 2 \times 2^{-(n-i)R} \tag{2.12}$$

In this thesis, the sequential random encoder always works by assigning random *parity bits* in a causal fashion to the observed source sequence. That is the bits generated at each time in Definition 3 are iid Bernoulli-(0.5) random variables. Since parity bits are assigned causally, if two source sequences x_1^n and $\tilde{x}_1^n$share the same length-l prefix, i.e. $x_1^l = \tilde{x}_1^l$ then their first $\lfloor lR \rfloor$ parity bits must match. Subsequent parity bits are drawn independently.

At the decoder side, the decoder receives the binary parity check $b_1^{\lfloor nR \rfloor}$ at time n. Hence the decoder knows the bin number which the source sequence x_1^n is in. Now the decoder has to pick one source sequence out of the bin $\mathcal{B}_x(x_1^n)$ as the estimate of x_1^n. If the decoder knows the distribution of the source p_{x}, it can simply pick the sequence with the maximum likelihood. Otherwise, which is called universal case, the decoder can use a minimum empirical entropy decoding rule. We first summarize the delay constrained error exponent results for both ML decoding and universal decoding.

First the maximum-likelihood decoding where the decoder knows the distribution of the source p_{x}.

Proposition 3 *Given a rate $R > H(p_{\mathsf{x}})$, there exists a randomized streaming encoder and maximum likelihood decoder pair (per Definition 3) and a finite constant $K > 0$, such that $\Pr[\widehat{\mathsf{x}}_i(i+\Delta) \neq \mathsf{x}_i] \leq K 2^{-\Delta E^{ML}(R)}$ for all $i, \Delta \geq 0$, or equivalently for all $n \geq \Delta \geq 0$*

$$\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}] \leq K 2^{-\Delta E^{ML}(R)} \tag{2.13}$$

$$\text{where } E^{ML}(R) = \sup_{\rho \in [0,1]} \{\rho R - (1+\rho) \log \left(\sum_x p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}} \right)\} \tag{2.14}$$

Secondly, if the decoder does not know the distribution, it can use the minimum empirical entropy rule to pick the sequence with the smallest empirical entropy inside the received bin. More interestingly, the decoder may prefer not to use its *knowledge* of the distribution of the source. The benefit of doing so is due to the uncertainty of the distribution of the source, for example a source distribution of $\{0.9, 0.051, 0.049\}$ may be mistaken as a distribution of $\{0.9, 0.049, 0.051\}$ and the decoding may as well be incorrect due to this nature. We summarize the universal coding result in the following proposition.

Proposition 4 *Given a rate $R > H(p_{\mathsf{x}})$, there exists a randomized streaming encoder and universal decoder pair (per Definition 3) such that for all $\epsilon > 0$ there exists finite $K > 0$ such that $\Pr[\widehat{\mathsf{x}}_i(i+\Delta) \neq \mathsf{x}_i] \leq K 2^{-\Delta(E^{UN}(R)-\epsilon)}$ for all $n, \Delta \geq 0$ where*

$$E^{UN}(R) = \inf_q D(q \| p_{\mathsf{x}}) + |R - H(q)|^+, \tag{2.15}$$

where q is an arbitrary probability distribution on $\mathcal{X}$ and where $|z|^+ = \max\{0, z\}$.

Remark: The error exponents of Propositions 3 and 4 both equal random block-coding exponents shown in (A.5).

Proof of Propositions 3: ML decoding

To show Propositions 3 and 4, we first develop the common core of the proof in the context of ML decoding. First, we describe the ML decoding rule.

ML decoding rule:

Denote by $\widehat{x}_1^n(n)$ the estimate of the source sequence x_1^n at time n.

$$\widehat{x}_1^n(n) = \underset{\tilde{x}_1^n \in \mathcal{B}_x(x_1^n)}{\arg\max}\, p_x(\tilde{x}_1^n) = \underset{\tilde{x}_1^n \in \mathcal{B}_x(x_1^n)}{\arg\max} \prod_{i=1}^{n} p_x(\tilde{x}_i) \tag{2.16}$$

The ML decoding rule in (2.16) is very simple. At time n, the decoder simply picks the sequence $\widehat{x}_1^n(n)$ with the highest likelihood which is in the same bin as the true sequence x_1^n). Now the estimate of source symbol $n-\Delta$ is simply the $(n-delay)^{th}$ symbol of $\widehat{x}_1^n(n)$, denoted by $\widehat{x}_{n-\Delta}(n)$.

Details of the proof:

The proof strategy is as follows. To lead to a decoding error, there must be some false source sequence $\tilde{x}_1^n$ that satisfies three conditions: (i) it must be in the same bin (share the same parities) as x_1^n, i.e., $\tilde{x}_1^n \in \mathcal{B}_s(x_1^n)$, (ii) it must be more likely than the true sequence, i.e., $p_x(\tilde{x}_1^n) > p_x(x_1^n)$, and (iii) $\tilde{x}_l \neq x_l$ for some $l \leq n-\Delta$.

The error probability can be union bounded as follows which is also illustrated in Figure 2.10.

$$\begin{aligned}
\Pr[\widehat{x}_{n-\Delta}(n) \neq x_{n-\Delta}] &\leq \Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}] \\
&= \sum_{x_1^n} \Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta} | x_1^n = x_1^n] p_x(x_1^n) && (2.17) \\
&= \sum_{x_1^n} \sum_{l=1}^{n-\Delta} \Pr\left[\exists\, \tilde{x}_1^n \in \mathcal{B}_s(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } p_x(\tilde{x}_1^n) \geq p_x(x_1^n)\right] p_x(x_1^n) && (2.18) \\
&= \sum_{l=1}^{n-\Delta} \left\{ \sum_{x_1^n} \Pr\left[\exists\, \tilde{x}_1^n \in \mathcal{B}_s(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } p_x(\tilde{x}_1^n) \geq p_x(x_1^n)\right] p_x(x_1^n) \right\} \\
&= \sum_{l=1}^{n-\Delta} p_n(l). && (2.19)
\end{aligned}$$

After conditioning on the realized source sequence in (2.17), the remaining randomness is only in the binning. In (2.18) we decompose the error event into a number of mutually

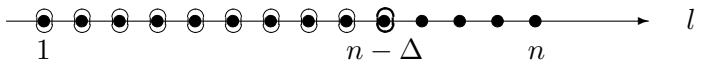

Figure 2.10. Decoding error probability at $n-\Delta$ can be union bounded by the sum of probabilities of first decoding error at l, $1 \le l \le n-\Delta$. The dominant error event $p_n(n-\Delta)$ is the one in the highlighted oval(shortest delay).

exclusive events by partitioning all source sequences $\tilde{x}_1^n$ into sets $\mathcal{F}_n(l, x_1^n)$ defined by the time l of the first sample in which they differ from the realized source x_1^n,

$$\mathcal{F}_n(l, x_1^n) = \{\tilde{x}_1^n \in \mathcal{X}^n | \tilde{x}_1^{l-1} = x_1^{l-1}, \tilde{x}_l \neq x_l\}, \tag{2.20}$$

and define $\mathcal{F}_n(n+1, x_1^n) = \{x_1^n\}$. Finally, in (2.19) we define

$$p_n(l) = \sum_{x_1^n} \Pr\left[\exists\, \tilde{x}_1^n \in \mathcal{B}_s(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } p_x(\tilde{x}_1^n) \ge p_x(x_1^n)\right] p_x(x_1^n). \tag{2.21}$$

We now upper bound $p_n(l)$ using a Chernoff bound argument similar to [39].

Lemma 1 $p_n(l) \le 2 \times 2^{-(n-l+1)E^{ML}(R)}$.

Proof:

$$\begin{aligned}
p_n(l) =& \sum_{x_1^n} \Pr\left[\exists\, \tilde{x}_1^n \in \mathcal{B}_s(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } p_x(\tilde{x}_1^n) \ge p_x(x_1^n)\right] p_x(x_1^n) \\
\le& \sum_{x_1^n} \min\Big[1, \sum_{\substack{\tilde{x}_1^n \in \mathcal{F}_n(l, x_1^n) \text{s.t.} \\ p_x(x_1^n) \le p_x(\tilde{x}_1^n)}} \Pr[\tilde{x}_1^n \in \mathcal{B}_s(x_1^n)]\Big] p_x(x_1^n) && (2.22) \\
\le& \sum_{x_1^{l-1}, x_l^n} \min\Big[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ p_x(x_l^n) < p_x(\tilde{x}_l^n)}} 2 \times 2^{-(n-l+1)R}\Big] p_x(x_1^{l-1}) p_x(x_l^n) && (2.23) \\
\le& 2 \times \sum_{x_l^n} \min\Big[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ p_x(x_l^n) < p_x(\tilde{x}_l^n)}} 2^{-(n-l+1)R}\Big] p_x(x_l^n) \\
=& 2 \times \sum_{x_l^n} \min\Big[1, \sum_{\tilde{x}_l^n} \mathit{1}[p_x(\tilde{x}_l^n) > p_x(x_l^n)] 2^{-(n-l+1)R}\Big] p_x(x_l^n) && (2.24) \\
\le& 2 \times \sum_{x_l^n} \min\left[1, \sum_{\tilde{x}_l^n} \min\left[1, \frac{p_x(\tilde{x}_l^n)}{p_x(x_l^n)}\right] 2^{-(n-l+1)R}\right] p_x(x_l^n) \\
\le& 2 \times \sum_{x_l^n} \left[\sum_{\tilde{x}_l^n} \left[\frac{p_x(\tilde{x}_l^n)}{p_x(x_l^n)}\right]^{\frac{1}{1+\rho}} 2^{-(n-l+1)R}\right]^{\rho} p_x(x_l^n) && (2.25)
\end{aligned}$$

$$
\begin{aligned}
&= 2 \times \sum_{x_l^n} p_{\mathsf{x}}(x_l^n)^{\frac{1}{1+\rho}} \left[\sum_{\tilde{x}_l^n} [p_{\mathsf{x}}(\tilde{x}_l^n)]^{\frac{1}{1+\rho}} \right]^{\rho} 2^{-(n-l+1)\rho R} \\
&= 2 \times \left[\sum_{x} p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}} \right]^{(n-l+1)} \left[\sum_{x} p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}} \right]^{(n-l+1)\rho} 2^{-(n-l+1)\rho R} \qquad (2.26) \\
&= 2 \times \left[\sum_{x} p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}} \right]^{(n-l+1)(1+\rho)} 2^{-(n-l+1)\rho R} \\
&= 2 \times 2^{-(n-l+1)\left[\rho R - (1+\rho)\log\left(\sum_x p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}}\right)\right]} \qquad (2.27)
\end{aligned}
$$

In (2.22) we apply the union bound. In (2.23) we use the fact that after the first symbol in which two sequences differ, the remaining parity bits are independent, and use the fact that only the likelihood of the differing suffixes matter in (2.12). That is, if $x_1^{l-1} = \tilde{x}_1^{l-1}$, then $p_{\mathsf{x}}(x_1^n) < p_{\mathsf{x}}(\tilde{x}_1^n)$ if and only if $p_{\mathsf{x}}(x_l^n) < p_{\mathsf{x}}(\tilde{x}_l^n)$. In (2.24) $1(\cdot)$ is the indicator function, taking the value one if the argument is true, and zero if it is false. We get (2.25) by limiting ρ to the range $0 \leq \rho \leq 1$ since the arguments of the minimization are both positive and upper-bounded by one. We use the iid property of the source, exchanging sums and products to get (2.26). The bound in (2.27) is true for all ρ in the range $0 \leq \rho \leq 1$. Maximizing (2.27) over ρ gives $p_n(l) \leq 2 \times 2^{-(n-l+1)E^{ML}(R)}$ where $E^{ML}(R)$ is defined in Proposition 3, in particular (2.14). □

Using Lemma 1 in (2.19) gives

$$
\begin{aligned}
\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}] \leq & 2 \times \sum_{l=1}^{n-\Delta} 2^{-(n-l+1)E^{ML}(R)} \qquad (2.28) \\
&= \sum_{l=1}^{n-\Delta} 2 \times 2^{-(n-l+1-\Delta)E^{ML}(R)} 2^{-\Delta E^{ML}(R)} \\
&\leq K_0 2^{-\Delta E^{ML}(R)} \qquad (2.29)
\end{aligned}
$$

In (2.29) we pull out the exponent in Δ. The remaining summation is a sum over decaying exponentials, can thus can be bounded by some constant K_0. This proves Proposition 3.

Proof of Proposition 4: Universal decoding

We use the union bound on the symbol-wise error probability introduced in (2.19), but with minimum empirical entropy, rather than maximum-likelihood, decoding.

Universal decoding rule:

$$\widehat{x}_l(n) = w[l]_l \quad \text{where} \quad w[l]_1^n = \underset{\bar{x}^n \in \mathcal{B}_s(x_1^n) \text{ s.t. } \bar{x}_1^{l-1} = \widehat{x}_1^{l-1}(n)}{\arg\min} H(\bar{x}_l^n). \tag{2.30}$$

We term this a sequential minimum empirical-entropy decoder. The reason for using this decoder instead of the standard minimum block-entropy decoder is that the block-entropy decoder has a polynomial term in n (resulting from summing over the type classes) that multiplies the exponential decay in Δ. For n large, this polynomial can dominate. Using the sequential minimum empirical-entropy decoder results in a polynomial term in Δ.

Details of the proof: With this decoder, errors can only occur if there is some sequence $\tilde{x}_1^n$ such that (i) $\tilde{x}_1^n \in \mathcal{B}_s(x_1^n)$, (ii) $\tilde{x}_1^{l-1} = x^{l-1}$, and $\tilde{x}_l \neq x_l$, for some $l \leq n - \Delta$, and (iii) the empirical entropy of $\tilde{x}_l^n$ is such that $H(\tilde{x}_l^n) < H(x_l^n)$. Building on the common core of the achievability (2.17)–(2.19) with the substitution of universal decoding in the place of maximum likelihood results in the following definition of $p_n(l)$ (cf. (2.31) with (2.21),

$$p_n(l) = \sum_{x_1^n} \Pr\left[\exists\, \tilde{x}_1^n \in \mathcal{B}_s(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } H(\tilde{x}_l^n) \leq H(x_l^n)\right] p_{\mathsf{x}}(x_1^n) \tag{2.31}$$

The following lemma gives a bound on $p_n(l)$.

Lemma 2 *For sequential minimum empirical entropy decoding,*

$$p_n(l) \leq 2 \times (n - l + 2)^{2|\mathcal{X}|} 2^{-(n-l+1)E^{UN}(R)}.$$

Proof: We define P^{n-l} to be the type of length-$(n-l+1)$ sequence x_l^n, and $\mathcal{T}_{P^{n-l}}$ to be the corresponding type class so that $x_l^n \in \mathcal{T}_{P^{n-l}}$. Analogous definitions hold for $\tilde{P}^{n-l}$ and

$\tilde{x}_l^n$. We rewrite the constraint $H(\tilde{x}_l^n) < H(\tilde{x}_l^n)$ as $H(\tilde{P}^{n-l}) < H(P^{n-l})$. Thus,

$$\begin{aligned}
p_n(l) =& \sum_{x_1^n} \Pr\left[\exists\ \tilde{x}_1^n \in \mathcal{B}_s(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } H(\tilde{x}_l^n) \le H(x_l^n)\right] p_x(x_1^n) \\
\le& \sum_{x_1^n} \min\Big[1, \sum_{\substack{\tilde{x}_1^n \in \mathcal{F}_n(l,\, x^n) \text{ s.t.} \\ H(\tilde{x}_l^n) \le H(x_l^n)}} \Pr[\tilde{x}_1^n \in \mathcal{B}_s(x_1^n)]\Big] p_x(x^n) \\
\le& \sum_{x_1^{l-1}, x_l^n} \min\Big[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ H(\tilde{x}_l^n) \le H(x_l^n)}} 2 \times 2^{-(n-l+1)R}\Big] p_x(x_1^{l-1}) p_x(x_l^n) \qquad (2.32) \\
\le& 2 \times \sum_{x_l^n} \min\Big[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ H(\tilde{x}_l^n) \le H(x_l^n)}} 2^{-(n-l+1)R}\Big] p_x(x_l^n) \qquad (2.33) \\
=& 2 \times \sum_{P^{n-l}} \sum_{x_l^n \in \mathcal{T}_{P^{n-l}}} \min\Big[1, \sum_{\substack{\tilde{P}^{n-l} \text{ s.t.} \\ H(\tilde{P}^{n-l}) \le H(P^{n-l})}} \sum_{\tilde{x}_l^n \in \mathcal{T}_{\tilde{P}^{n-l}}} 2^{-(n-l+1)R}\Big] p_x(x_l^n) \qquad (2.34) \\
\le& 2 \times \sum_{P^{n-l}} \sum_{x_{l+1}^n \in \mathcal{T}_{P^{n-l}}} \min\Big[1, (n-l+2)^{|\mathcal{X}|} 2^{-(n-l)[R - H(P^{n-l})]}\Big] p_x(x_l^n) \qquad (2.35) \\
\le& 2 \times (n-l+2)^{|\mathcal{X}|} \sum_{P^{n-l}} \sum_{x_l^n \in \mathcal{T}_{P^{n-l}}} 2^{-(n-l+1)[|R - H(P^{n-l})|^+]} \\
& \qquad 2^{-(n-l+1)[D(P^{n-l} \| P_s) + H(P^{n-l})]} \qquad (2.36) \\
\le& 2 \times (n-l+2)^{|\mathcal{X}|} \sum_{P^{n-l}} 2^{-(n-l+1) \inf_q [D(q \| P_s) + |R - H(q)|^+]} \qquad (2.37) \\
\le& 2 \times (n-l+2)^{2|\mathcal{X}|} 2^{-(n-l+1) E^{UN}(R)} \qquad (2.38)
\end{aligned}$$

To show (2.32), we use the bound in (2.12). In going from (2.34) to (2.35) first note that the argument of the inner-most summation (over $\tilde{x}_l^n$) does not depend on x_1^n. We then use the following relations: (i) $\sum_{\tilde{x}_l^n \in \mathcal{T}_{\tilde{P}^{n-l}}} = |\mathcal{T}_{\tilde{P}^{n-l}}| \le 2^{(n-l+1)H(\tilde{P}^{n-l})}$, which is a standard bound on the size of the type class [29], (ii) $H(\tilde{P}^{n-l}) \le H(P^{n-l})$ by the sequential minimum empirical entropy decoding rule, and (iii) the polynomial bound on the number of types [28], $|\{\tilde{P}^{n-l}\}| \le (n-l+2)^{|\mathcal{X}|}$. In (2.36) we recall the function definition $|\cdot|^+ \triangleq \max\{0, \cdot\}$. We pull the polynomial term out of the minimization and use $p_x(x_l^n) = 2^{-(n-l+1)[D(P^{n-l} \| P_s) + H(P^{n-l})]}$ for all $x_l^n \in \mathcal{T}_{P^{n-l}}$. It is also in (2.36) that we see why we use a sequential minimum empirical entropy decoding rule instead of a block minimum entropy decoding rule. If we had not marginalized out over x_1^{l-1} in (2.33) then we would have a polynomial term out front in terms of n rather than $n-l$, which for large n could dominate the exponential decay in $n-l$. As the expression in (2.37) no longer depends on x_l^n, we simplify by using $|\mathcal{T}_{P^{n-l}}| \le 2^{(n-l+1)H(P^{n-l})}$. In (2.38) we use the definition of the universal error exponent

$E^{UN}(R)$ from (2.15) of Proposition 4, and the polynomial bound on the number of types. Other steps should be obvious. □

Lemma 2 and $\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}] \leq \sum_{l=1}^{n-\Delta} p_n(l)$ imply that:

$$\begin{aligned}\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}] &\leq \sum_{l=1}^{n-\Delta} (n-l+2)^{2|\mathcal{X}|} 2^{-(n-l+1)E^{UN}(R)} \\ &\leq \sum_{l=1}^{n-\Delta} K_1 2^{-(n-l+1)[E^{UN}(R)-\epsilon]} \qquad (2.39) \\ &\leq K 2^{-\Delta[E^{UN}(R)-\epsilon]} \qquad (2.40)\end{aligned}$$

In (2.39) we incorporate the polynomial into the exponent. Namely, for all $a > 0$, $b > 0$, there exists a C such that $z^a \leq C 2^{b(z-1)}$ for all $z \geq 1$. We then make explicit the delay-dependent term. Pulling out the exponent in Δ, the remaining summation is a sum over decaying exponentials, and can be bounded by a constant. Together with K_1, this gives the constant K in (2.40). This proves Proposition 4. Note that the ϵ in (2.40) does not enter the optimization because $\epsilon > 0$ can be picked equal to any constant. The choice of ϵ effects the constant K in Proposition 4.

Discussions on sequential random binning

We propose a sequential random binning scheme, as summarized in Propositions 3 and 4, which achieves the random block coding error exponent defined in (A.5). Although this error exponent is strictly suboptimal as shown in Proposition 8 in Section 2.5, it provides a very useful tool in deriving the achievability results for delay constrained coding (both source coding and channel coding) as will be seen in future chapters where delay constrained source coding with decoder side-information and distributed source coding problems are discussed. Philosophically speaking, the essence of sequential random binning is to leave the uncertainties about the source at the encoder to the future, and let the binning reduce the uncertainties of the source. This is in contrast to the optimal scheme shown in Section 2.4.1, where the encoder queues up the fixed-to-variable code and thus for a particular source symbol and delay, the uncertainties lie in the past.

2.4 Proof of the Main Results

In this section we derive the delay constrained source coding error exponent in Theorem 1. We implement a simple universal coding scheme to achieve this error exponent defined in Theorem 1. As for the converse, we use a *focusing bound* type of argument which is parallel to the analysis on the delay constrained error exponent for channel coding with feedback in [67].

Following the notion of delay constrained source coding error exponent $E_s(R)$ in Definition 2, we have the following result in Theorem 1.

$$E_s(R) = \inf_{\alpha>0} \frac{1}{\alpha} E_{s,b}((\alpha+1)R)$$

In Section 2.4.1, we first show the achievability of $E_s(R)$ by a simple fixed to variable length universal code and a FIFO queue coding scheme. Then in Section 2.4.2, we show that $E_s(R)$ is indeed an upper bound on the delay constrained error exponent. This error exponent was first derived in [49] in a slightly different setup by using a non-universal (encoder needs to know the distribution of the source, p_X, and the rate R) coding scheme, we give a simple proof of Jelinek's result by using large deviation theory in Appendix B.

2.4.1 Achievability

In this section, we introduce a universal coding scheme which achieves the delay constrained error exponent shown in Theorem 1. The coding scheme only depends on the size of the alphabet of the source, not the distribution of the source. We first describe our universal coding scheme. The basic idea is to encode a long block source symbols into a variable-length binary strings. The binary strings first describe the *type* of the source block, then index different source blocks with the same type by a one to one map. This is in line with the Minimum Description Length principle [66, 4]. Source blocks with the same *type* have the same length.

Optimal Universal Coding

A block-length N is chosen that is much smaller than the target end-to-end delays, while still being large enough. This finite block length N will be absorbed into the finite constant K. For a discrete memoryless source and large block-lengths N, the variable-length code

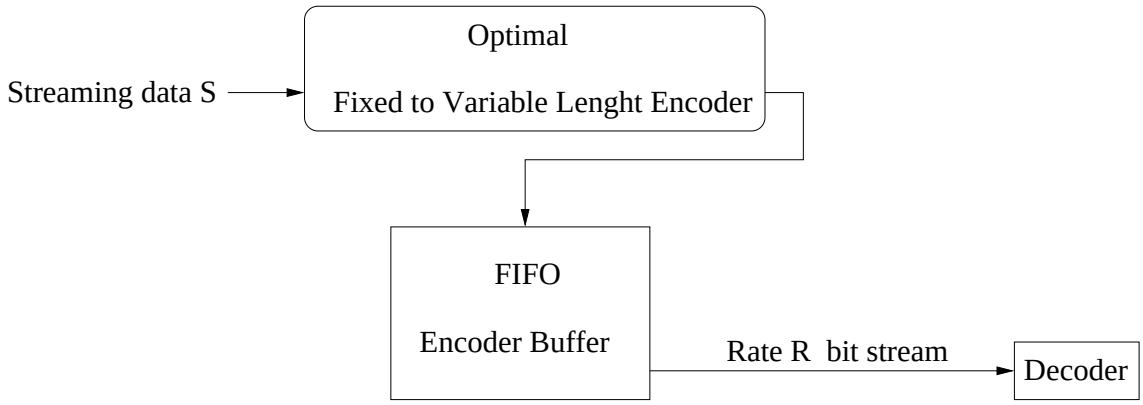

Figure 2.11. A universal delay constrained source coding system

consists of two stages: first describing the type of the block [3] $\vec{x}_i$ using $O(|\mathcal{X}|\log N)$ bits and then describing which particular realization has occurred by using a variable $NH(\vec{x}_i)$ bits. The overhead $O(|\mathcal{X}|\log N)$ is asymptotically negligible and the code is also universal in nature. It is easy to verify that the average code length: $\lim_{N\to\infty}\frac{E_{p_x}(l(\vec{x}))}{N} = H(p_x)$ This code is obviously a prefix-free code. Write $l(\vec{x}_i)$ as the length of the codeword for $\vec{x}_i$, then:

$$NH(\vec{x}_i) \leq l(\vec{x}_i) \leq |\mathcal{X}|\log(N+1) + NH(\vec{x}_i) \tag{2.41}$$

The binary sequence describing the source is fed to the FIFO buffer illustrated in Figure 2.11. Notice that if the buffer is empty, the output of the buffer can be gibberish binary bits. The decoder simply discards these meaningless bits because it is aware that the buffer is empty.

Proposition 5 *For the iid source $\sim p_x$ using the universal delay constrained code described above, for all$\epsilon > 0$, there exists $K < \infty$, s.t. for all t, Δ:*

$$\Pr[\vec{x}_t \neq \widehat{\vec{x}}_t((t+\Delta)N)] \leq K2^{-\Delta N(E_s(R)-\epsilon)}$$

Where $\widehat{\vec{x}}_t((t+\Delta)N)$ is the estimate of $\vec{x}_t$ at time $(t+\Delta)N$. Before the proof, we have the following lemma to bound the probabilities of atypical source behavior.

Lemma 3 *(Source atypicality) for all $\epsilon > 0$, block length N large enough, there exists $K < \infty$, s.t. for all n, if $r < \log|\mathcal{X}|$:*

$$K2^{-nN(E_{s,b}(r)+\epsilon)} \leq \Pr[\sum_{i=1}^{n} l(\vec{x}_i) > nNr] \leq K2^{-nN(E_{s,b}(r)-\epsilon)} \tag{2.42}$$

[3] $\vec{x}_i$ is the i^{th} block of length N.

Proof: Only need to show the case for $r > H(p_x)$. By the Cramér's theorem[31], for all $\epsilon_1 > 0$, there exists K_1, such that:

$$\Pr[\sum_{i=1}^{n} l(\vec{x}_i) > nNr] = \Pr[\frac{1}{n}\sum_{i=1}^{n} l(\vec{x}_i) > Nr] \leq K_1 2^{-n(\inf_{z>Nr} I(z)-\epsilon_1)} \tag{2.43}$$

And

$$\Pr[\sum_{i=1}^{n} l(\vec{x}_i) > nNr] = \Pr[\frac{1}{n}\sum_{i=1}^{n} l(\vec{x}_i) > Nr] \geq K_1 2^{-n(\inf_{z>Nr} I(z)+\epsilon_1)} \tag{2.44}$$

where the rate function $I(z)$ is [31]:

$$I(z) \quad = \quad \sup_{\rho\in\mathcal{R}}\{\rho z - \log(\sum_{\vec{x}\in\mathcal{X}^N} p_x(\vec{x})2^{\rho l(\vec{x})})\} \tag{2.45}$$

Here we use this equivalent $K-\epsilon$ notation instead of limit superior and limit inferior in the Cramér's theorem [31].

Write $I(z,\rho) = \rho z - \log(\sum_{\vec{x}\in\mathcal{X}^N} p_x(\vec{x})2^{\rho l(\vec{x})})$, $I(z,0) = 0$. $z > Nr > NH(p_x)$, for large N:

$$\frac{\partial I(z,\rho)}{\partial \rho}|_{\rho=0} = z - \sum_{\vec{x}\in\mathcal{X}^N} p_x(\vec{x}) l(\vec{x}) \geq 0$$

By the Hölder's inequality, for all ρ_1, ρ_2, and for all $\theta \in (0,1)$:

$$\begin{aligned}(\sum_i p_i 2^{\rho_1 l_i})^{\theta}(\sum_i p_i 2^{\rho_2 l_i})^{(1-\theta)} \quad &\geq \quad \sum_i (p_i^{\theta} 2^{\theta\rho_1 l_i})(p_i^{(1-\theta)} 2^{(1-\theta)\rho_2 l_i}) \\ &= \quad \sum_i p_i 2^{(\theta\rho_1+(1-\theta)\rho_2) l_i}\end{aligned}$$

This shows that $\log(\sum_{\vec{x}\in\mathcal{X}^N} p_x(\vec{x})2^{\rho l(\vec{x})})$ is a convex $\cup$ function on ρ, thus $I(z,\rho)$ is a concave $\cap$ function on ρ for fixed z. Then $\forall z > 0$, $\forall \rho < 0$, $I(z,\rho) < 0$, which means that the ρ to maximize $I(z,\rho)$ is positive. This implies that $I(z)$ is monotonically increasing with z and obviously $I(z)$ is continuous. Thus $\inf_{z>Nr} I(z) = I(Nr)$

For $\rho \geq 0$, using the upper bound on $l(\vec{x})$ in (2.41):

$$\begin{aligned}\log(\sum_{\vec{x}\in\mathcal{X}^N} p_x(\vec{x})2^{\rho l(\vec{x})}) \quad &\leq \quad \log(\sum_{T_P^N \in T^N} 2^{-ND(p\|p_x)} 2^{\rho(|\mathcal{S}|\log(N+1)+NH(p))}) \\ &\leq \quad \log((N+1)^{|\mathcal{X}|} 2^{-N\min_p\{D(p\|p_x)-\rho H(p)\}+\rho|\mathcal{S}|\log(N+1)}) \\ &= \quad N\big(-\min_p\{D(p\|p_x)-\rho H(p)\}+\epsilon_N\big)\end{aligned}$$

where $\epsilon_N = \frac{(1+\rho)|\mathcal{X}|\log(N+1)}{N}$ goes to 0 as N goes to infinity.

For $\rho \geq 0$, using the lower bound on $l(\vec{x})$ in (2.41):

$$\begin{aligned}
\log(\sum_{\vec{x}\in\mathcal{X}^N} p_x(\vec{x})2^{\rho l(\vec{x})}) &\geq \log(\sum_{p\in\mathcal{T}^N} 2^{-ND(p\|p_x)+|\mathcal{S}|\log(N+1)}2^{\rho NH(p)}) \\
&\geq \log(2^{-N\min_{p\in\mathcal{T}^N}\{D(p\|p_x)-\rho H(p)\}+|\mathcal{S}|\log(N+1)}) \\
&= N\big(-\min_p\{D(p\|p_x)-\rho H(p)\}-\epsilon'_N\big)
\end{aligned}$$

where $\epsilon'_N = \frac{|\mathcal{X}|\log(N+1)}{N} - \min_p\{D(p\|p_x)-\rho H(p)\} + \min_{p\in T^N}\{D(p\|p_x)-\rho H(p)\}$ goes to 0 as N goes to infinity, $\mathcal{T}^N$ is the set of all types of $\mathcal{X}^N$.

Substitute the above inequalities to $I(Nr)$ defined in (2.45):

$$I(Nr) \geq N\big(\sup_{\rho>0}\{\min_p \rho(r-H(p))+D(p\|p_x)\}-\epsilon_N\big) \tag{2.46}$$

And

$$I(Nr) \leq N\big(\sup_{\rho>0}\{\min_p \rho(r-H(p))+D(p\|p_x)\}+\epsilon'_N\big) \tag{2.47}$$

First fix ρ, by a simple Lagrange multiplier argument, with fixed $H(p)$, we know that the distribution p to minimize $D(p\|p_x)$ is a tilted distribution of p_x^α. It can be verified that $\frac{\partial H(p_x^\alpha)}{\partial\alpha} \geq 0$ and $\frac{\partial D(p_x^\alpha\|p_x)}{\partial\alpha} = \alpha\frac{\partial H(p_x^\alpha)}{\partial\alpha}$. Thus the distribution to minimize $D(p\|p_x)-\rho H(p)$ is p_x^ρ. Using some algebra, we have

$$D(p_x^\rho\|p_x)-\rho H(p_x^\rho) = -(1+\rho)\log\sum_{s\in\mathcal{S}} p_x(x)^{\frac{1}{1+\rho}}$$

Substitute this into (2.46) and (2.47) respectively:

$$\begin{aligned}
I(Nr) &\geq N\big(\sup_{\rho>0}\rho r-(1+\rho)\log\sum_{s\in\mathcal{X}} p_x(x)^{\frac{1}{1+\rho}}-\epsilon_N\big) \\
&= N\big(E_{s,b}(r)-\epsilon_N\big)
\end{aligned} \tag{2.48}$$

The last equality can again be proved by a simple Lagrange multiplier argument.

Similarly:

$$I(Nr) \leq N\big(E_{s,b}(r)+\epsilon'_N\big) \tag{2.49}$$

Substitute (2.48) and (2.49) into(2.43) and (2.44) respectively, by letting ϵ_1 small enough and N big enough thus ϵ_N and ϵ'_N small enough, we get the the desired bound in (2.42). □

Now we are ready to prove Proposition 5.

Proof: We give an upper bound on the decoding error on $\vec{x}_t$ at time $(t+\Delta)N$. At time $(t+\Delta)N$, the decoder *cannot* decode $\vec{x}_t$ with 0 error probability iff the binary strings describing $\vec{x}_t$ are *not* all out of the buffer. Since the encoding buffer is FIFO, this means that the number of outgoing bits from some time t_1 to $(t+\Delta)N$ is less than the number of the bits in the buffer at time t_1 plus the number of incoming bits from time t_1 to time tN. Suppose the buffer is last empty at time $tN-nN$ where $0 \leq n \leq t$, given this condition, the decoding error occurs only if $\sum_{i=0}^{n-1} l(\vec{x}_{t-i}) > (n+\Delta)NR$. Write $l_{\max}$ as the longest code length, $l_{\max} \leq |\mathcal{X}| \log(N+1) + N|\mathcal{X}|$. Then $\Pr[\sum_{i=0}^{n-1} l(\vec{x}_{t-i}) > (n+\Delta)NR] > 0$ only if $n > \frac{(n+\Delta)NR}{l_{\max}} > \frac{\Delta NR}{l_{\max}} \stackrel{\Delta}{=} \beta\Delta$

$$
\begin{aligned}
\Pr[\vec{x}_t \neq \vec{x}_t((t+\Delta)N)] \quad &\leq \quad \sum_{n=\beta\Delta}^{t} \Pr[\sum_{i=0}^{n-1} l(\vec{x}_{t-i}) > (n+\Delta)NR] \qquad (2.50)\\
&\leq_{(a)} \quad \sum_{n=\beta\Delta}^{t} K_1 2^{-nN(E_{s,b}(\frac{(n+\Delta)NR}{nN})-\epsilon_1)} \\
&\leq_{(b)} \quad \sum_{n=\gamma\Delta}^{\infty} K_2 2^{-nN(E_{s,b}(R)-\epsilon_2)} + \sum_{n=\beta\Delta}^{\gamma\Delta} K_2 2^{-\Delta N(\min_{\alpha>0}\{\frac{E_{s,b}((1+\alpha)R)}{\alpha}\}-\epsilon_2)} \\
&\leq_{(c)} \quad K_3 2^{-\gamma\Delta N(E_{s,b}(R)-\epsilon_2)} + |\gamma\Delta - \beta\Delta| K_3 2^{-\Delta N(E_s(R)-\epsilon_2)} \\
&\leq_{(d)} \quad K 2^{-\Delta N(E_s(R)-\epsilon)}
\end{aligned}
$$

where, $K_i's$ and $\epsilon_i's$ are properly chosen positive real numbers. (a) is true because of Lemma 3. Define $\gamma = \frac{E_s(R)}{E_{s,b}(R)}$, in the first part of (b), we only need the fact that $E_{s,b}(R)$ is non decreasing with R. In the second part of (b), we write $\alpha = \frac{\Delta}{n}$ and take the α to minimize the error exponents. The first term of (c)comes from the sum of a geometric series. The second teFrm of (c) is by the definition of $E_s(R)$. (d) is by the definitions of γ. ■

2.4.2 Converse

To bound the best possible error exponent with fixed delay, we consider a block coding encoder/decoder pair constructed by the delay constrained encoder/decoder pair and translate the block-coding bounds of [29] to the fixed delay context. The argument is analogous to the "focusing bound" derivation in [67] for channel coding with feedback.

Proposition 6 *For fixed-rate encodings of discrete memoryless sources, it is not possible to achieve an error exponent with fixed-delay higher than*

$$\inf_{\alpha>0} \frac{1}{\alpha} E_{s,b}((\alpha+1)R) \qquad (2.51)$$

from the definition of delay constrained source coding error exponent in Definition 2, the statement of this proposition is equivalent to the following mathematical statement:

For any $E > \inf_{\alpha>0} \frac{1}{\alpha} E_{s,b}((\alpha+1)R)$*, there exists an positive real value* ϵ*, such that for any* $K < \infty$*, there exists* $i > 0$*,* $\Delta > 0$ *and*

$$\Pr[x_i \neq \widehat{x}_i(i+\Delta)] > K2^{-\Delta(E_s(R)-\epsilon)}$$

Proof: We show the proposition by contradiction. Suppose that the delay-constrained error exponent can be higher than $\inf_{\alpha>0} \frac{1}{\alpha} E_{s,b}((\alpha+1)R)$. Then according to Definition 2, there exists a delay-constrained source coding system, such that for some $E > \inf_{\alpha>0} \frac{1}{\alpha} E_{s,b}((\alpha+1)R)$, for any $\epsilon > 0$, there exists $K < \infty$, such that for all $i > 0$, $\Delta > 0$

$$\Pr[x_i \neq \widehat{x}_i(i+\Delta)] \leq K2^{-\Delta(E-\epsilon)}$$

so we choose some $\epsilon > 0$, such that

$$E - \epsilon > \inf_{\alpha>0} \frac{1}{\alpha} E_{s,b}((\alpha+1)R) \tag{2.52}$$

Then consider a block coding scheme $(\mathcal{E}, \mathcal{D})$ which is built on the delay constrained source coding system. The encoder of the block coding system is the *same* as the delay-constrained source encoder, and the block decoder $\mathcal{D}$ works as follows:

$$\widehat{x_1^i} = (\widehat{x}_1(1+\Delta), \widehat{x}_2(2+\Delta), ..., \widehat{x}_i(i+\Delta))$$

Now the block decoding error of this coding system can be upper bounded as follows, for any $i > 0$ and $\Delta > 0$:

$$\begin{aligned} \Pr[\widehat{x_1^i} \neq x_1^i] &\leq \sum_{t=1}^{i} \Pr[\widehat{x}_i(i+\Delta) \neq x_i] \\ &\leq \sum_{t=1}^{i} K2^{-\Delta(E-\epsilon)} \\ &= iK2^{-\Delta(E-\epsilon)} \end{aligned} \tag{2.53}$$

The block coding scheme $(\mathcal{E}, \mathcal{D})$ is a block source coding system for i source symbols by using $\lfloor R(i+\Delta) \rfloor$ bits hence has a rate $\frac{\lfloor R(i+\Delta) \rfloor}{i} \leq \frac{i+\Delta}{i} R$. From the classical block coding result in Theorem 9, we know that the source coding error exponent $E_{s,b}(R)$ is monotonically

increasing and continuous in R, so $E_{s,b}(\frac{\lfloor R(i+\Delta)\rfloor}{i}) \leq E_{s,b}(\frac{R(i+\Delta)}{i})$. Again from Theorem 9, we know that the block coding error probability can be bounded in the following way:

$$\Pr[\widehat{x_1^i} \neq x_1^i] \quad > \quad 2^{-i(E_{s,b}(\frac{R(i+\Delta)}{i})+\epsilon_i)} \tag{2.54}$$

where $\lim_{i\to\infty} \epsilon_i = 0$.

Combining (2.53) and (2.54), we have:

$$2^{-i(E_{s,b}(\frac{R(i+\Delta)}{i})+\epsilon_i)} < iK2^{-\Delta(E-\epsilon)}$$

Now let $\alpha = \frac{\Delta}{i}$, $\alpha > 0$, then the above inequality becomes:

$$2^{-i(E_{s,b}(R(1+\alpha))+\epsilon_i)} < 2^{-i\alpha(E-\epsilon-\theta_i)} \text{ and hence:}$$

$$E - \epsilon < \frac{1}{\alpha}(E_{s,b}(R(1+\alpha)+\epsilon_i) + \theta_i \tag{2.55}$$

where $\lim_{i\to\infty} \theta_i = 0$. Now the above inequality is true for all i, $\alpha > 0$, meanwhile $\lim_{i\to\infty} \theta_i = 0$, $\lim_{i\to\infty} \epsilon_i = 0$, taking all these into account, we have:

$$E - \epsilon \leq \inf_{\alpha>0} \frac{1}{\alpha} E_{s,b}(R(1+\alpha)) \tag{2.56}$$

Now (2.56) contradicts with the assumption in (2.52), thus the proposition is proved. ■

Combining Proposition 5 and Proposition 6, we establish the desired results summarized in Theorem 1. For source coding with delay constraints problem, we are able to provide an accurate estimate on the speed of the decaying of the symbol wise error probability. This error exponent is in general larger than the block coding error exponent as shown in Figure 2.3.

2.5 Discussions

We first investigate some of the properties of the delay constraint source coding error exponent $E_s(R)$ defined in (2.7). Then we will conclude this chapter with some ideas for future work.

2.5.1 Properties of the delay constrained error exponent $E_s(R)$

The source coding with delay error exponent can be parameterized by a single real number ρ which is parallel to the delay constrained channel coding with feedback error

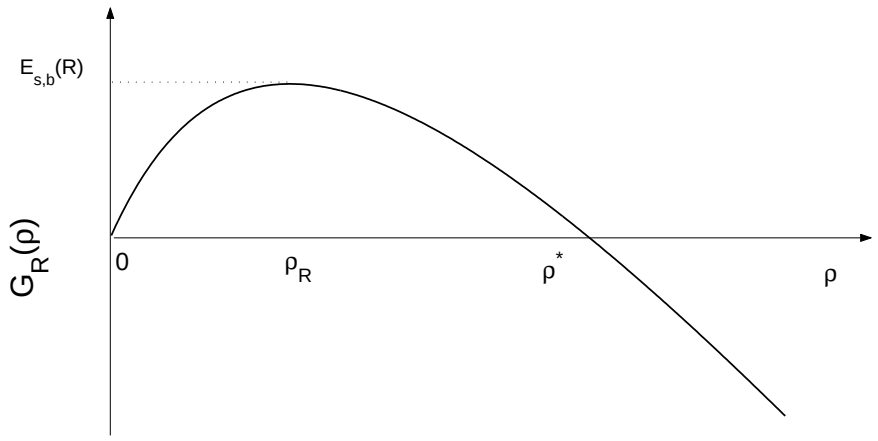

Figure 2.12. $G_R(\rho)$

exponents for symmetric channels derived in [67]. The parametrization of $E_s(R)$ makes the evaluations of $E_s(R)$ easier and potentially enables us to study other properties of $E_s(R)$. We summarize the result in the following corollary first shown in [20].

Proposition 7 *Parametrization of $E_s(R)$:*

$$E_s(R) = E_0(\rho^*) = (1+\rho^*)\log[\sum_s p_x(x)^{\frac{1}{1+\rho^*}}] \tag{2.57}$$

where ρ^* satisfies the following condition: $R = \frac{E_0(\rho^*)}{\rho^*} = \frac{(1+\rho^*)\log[\sum_s p_x(x)^{\frac{1}{1+\rho^*}}]}{\rho^*}$, i.e. $\rho^* R = E_0(\rho^*)$

Proof: For the simplicity of the notations, we define $G_R(\rho) = \rho R - E_0(\rho)$, thus $G_R(\rho^*) = 0$ by the definition of ρ^*. And $\max_\rho G_R(\rho) = E_{s,b}(R)$ by the definition of the block coding error exponent $E_{s,b}(R)$ in Theorem 9. As shown in Lemma 31 and Lemma 22 in Appendix G:

$$\frac{dG_R(\rho)}{d\rho} = R - H(p_x^\rho), \quad \frac{d^2G_R(\rho)}{d\rho^2} = -\frac{dH(p_x^\rho)}{d\rho} \leq 0$$

Furthermore, $\lim_{\rho\to\infty}\frac{1}{\rho}G_R(\infty) = R - \log|\mathcal{X}| < 0$, thus $G_R(\infty) < 0$. Obviously $G_R(0) = 0$. So we know that $G_R(\rho)$ is a concave $\cap$ function with a unique root for $\rho > 0$, as illustrated

in Figure 2.12. Let $\rho_R \geq 0$ be such that maximize $G_R(\rho)$, by the concavity of $G_R(\rho)$ this is equivalent to $R - H(p_x^{\rho_R}) = 0$. By concavity, we know that $\rho_R < \rho^*$.
Write:

$$\begin{aligned} F_R(\alpha, \rho) &= \frac{1}{\alpha}(\rho(\alpha+1)R - E_0(\rho)) \\ &= \rho R + \frac{G_R(\rho)}{\alpha} \end{aligned}$$

Then $E_s(R) = \inf_{\alpha>0} \sup_{\rho \geq 0} F_R(\alpha, \rho)$. And for any $\alpha \in (0, \frac{\log|\mathcal{X}|}{R} - 1)$,

$$\begin{aligned} &\frac{\partial F_R(\alpha,\rho)}{\partial \rho} = \frac{(\alpha+1)R - H(p_x^\rho)}{\alpha} \\ &\frac{\partial^2 F_R(\alpha,\rho)}{\partial \rho^2} = \frac{1}{\alpha}\frac{dH(p_x^\rho)}{d\rho} \leq 0 \\ &F_R(\alpha, 0) = 0 \\ &F_R(\alpha, \infty) < 0 \end{aligned} \tag{2.58}$$

So for any $\alpha \in (0, \frac{\log|\mathcal{X}|}{R} - 1)$, let $\rho(\alpha)$ be so $F_R(\alpha, \rho(\alpha))$ is maximized. $\rho(\alpha)$ is thus the unique solution to:

$$(\alpha+1)R - H(p_x^{\rho(\alpha)}) = 0$$

Define $\alpha^* = \frac{H(p_x^{\rho^*})}{R} - 1$, i.e. $(\alpha^*+1)R - H(p_x^{\rho^*}) = 0$ which implies that,

$$\begin{aligned} \alpha^* &= \frac{H(p_x^{\rho^*})}{R} - 1 < \frac{H(p_x^\infty)}{R} - 1 = \frac{\log|\mathcal{X}|}{R} - 1 \\ \alpha^* &= \frac{H(p_x^{\rho^*})}{R} - 1 > \frac{H(p_x^{\rho_R})}{R} - 1 = 0 \end{aligned}$$

Now we establish that $\alpha^* \in (0, \frac{\log|\mathcal{X}|}{R} - 1)$, by the definition of α^*, ρ^* maximizes $F_R(\alpha^*, \rho)$ over all ρ. From the above analysis:

$$\begin{aligned} E_s(R) &= \inf_{\alpha>0} \frac{1}{\alpha} E_b((\alpha+1)R) \\ &= \inf_{\alpha>0} \sup_{\rho\geq 0} \frac{1}{\alpha}(\rho(\alpha+1)R - E_0(\rho)) \\ &= \inf_{\alpha>0} \sup_{\rho\geq 0} F_R(\alpha, \rho) \\ &\leq \sup_{\rho\geq 0} F_R(\alpha^*, \rho) \\ &= F_R(\alpha^*, \rho^*) \\ &= \rho^* R \end{aligned} \tag{2.59}$$

The last step is from the definition of $F_R(\alpha, \rho)$ and the definition of ρ^*.

On the other hand:

$$\begin{aligned} E_s(R) &= \inf_{\alpha>0} \sup_{\rho\geq 0} F_R(\alpha,\rho) \\ &\geq \sup_{\rho\geq 0} \inf_{\alpha>0} F_R(\alpha,\rho) \\ &\geq \sup_{\rho\geq 0} F_R(\alpha^*,\rho) \\ &= F_R(\alpha^*,\rho^*) \\ &= \rho^* R \end{aligned} \tag{2.60}$$

Combining (2.59) and (2.60), we derive the desired parametrization of $E_s(R)$. Here we actually also prove that (α^*, ρ^*) is the saddle point for function $F_R(\alpha, \rho)$. □

Our next proposition shows that for any rate within the meaningful rate region $(H(p_x), \log|\mathcal{X}|)$, the delay constraint error exponent is strictly larger than the block coding error exponent.

Proposition 8 *Comparison of $E_s(R)$ and $E_{s,b}(R)$:*

$$\forall R \in (H(p_x), \log|\mathcal{X}|), \qquad E_s(R) > E_{s,b}(R)$$

Proof: From the Proposition 7, we have: $E_s(R) = \rho^* R$. Also from Proposition 7, we know that $\rho_R < \rho^*$, where the block coding error exponent $E_{s,b}(R) = \rho_R R - E_0(\rho_R)$. By noticing that $E_0(\rho) \geq 0$, for all $\rho \geq 0$, we have:

$$\begin{aligned} E_s(R) &= \rho^* R \\ &> \rho_R R \\ &\geq \rho_R R - E_0(\rho_R) \\ &\geq E_{s,b}(R) \end{aligned}$$

□

The source coding error exponent $E_{s,b}(R)$ has a zero right derivative at the entropy rate of the source $R = H(p_x)$, this is a simple corollary of Lemma 23 in Appendix G. The delay constrained source coding error exponent is different from its block coding counterpart as shown in Figure 2.3. The next proposition shows that the right derivative at the entropy rate of the source is positive.

Proposition 9 *Right derivative of $E_s(R)$ at $R = H(p_x)$:*

$$\lim_{R\to H(p_x)^+} \frac{dE_s(R)}{dR} = \frac{2H(p_x)}{\sum_x p_x(x)[\log(p_x(x))]^2 - H(p_x)^2}$$

Proof: Proposition 7 shows that the delay constrained error exponent $E_s(R)$ can be parameterized by a real number ρ as $E_s(R) = E_0(\rho)$ and $R(\rho) = \frac{E_0(\rho)}{\rho}$. In particular for $R(\rho) = H(p_x)$, $\rho = 0$. By Lemma 23 in Appendix G, we have

$$\begin{aligned}
\frac{d\frac{E_0(\rho)}{\rho}}{d\rho} &= \frac{\rho\frac{dE_0(\rho)}{d\rho} - E_0(\rho)}{\rho^2} \\
&= \frac{1}{\rho^2}(\rho H(p_x^\rho) - E_0(\rho)) \\
&= \frac{1}{\rho^2} D(p_x^\rho \| p_x) \qquad (2.61) \\
\frac{dE_0(\rho)}{d\rho} &= H(p_x^\rho) \qquad (2.62) \\
\frac{dD(p_x^\rho \| p_x)}{d\rho} &= \rho\frac{dH(p_x^\rho)}{d\rho} \qquad (2.63)
\end{aligned}$$

Now we can follow Gallager's argument on page 143-144 in [41]. By noticing that both $E_0(R)$ and $\frac{E_0(\rho)}{\rho}$ are positive for $\rho \in (0, +\infty)$. Combining (2.61) and (2.62):

$$\begin{aligned}
\frac{dE_s(R)}{dR} &= \frac{\frac{dE_0(\rho)}{d\rho}}{\frac{dR(\rho)}{d\rho}} \\
&= \frac{\rho^2 H(p_x^\rho)}{D(p_x^\rho \| p_x)} \qquad (2.64)
\end{aligned}$$

From (2.64), we know the right derivative of $E_s(R)$ at $R = H(p_x)$:

$$\begin{aligned}
\lim_{R \to H(p_x)^+} \frac{dE_s(R)}{dR} &= \lim_{\rho \to 0^+} \frac{\rho^2 H(p_x^\rho)}{D(p_x^\rho \| p_x)} \\
&= \lim_{\rho \to 0^+} \frac{2\rho H(p_x^\rho) + \rho^2 \frac{dH(p_x^\rho)}{d\rho}}{\rho \frac{dH(p_x^\rho)}{d\rho}} \qquad (2.65) \\
&= \lim_{\rho \to 0^+} \frac{2H(p_x^\rho)}{\frac{dH(p_x^\rho)}{d\rho}} \\
&= \frac{2H(p_x)}{\sum_x p_x(x)[\log(p_x(x))]^2 - H(p_x)^2} \qquad (2.66)
\end{aligned}$$

(2.65) is true because of the L'Hospital rule [86]. (2.66) is true by simple algebra.

By the Cauchy-Schwatz inequality $\sum_x p_x(x)[\log(p_x(x))]^2 - H(p_x)^2 > 0$ unless p_x is uniform. Thus the right derivative of $E_s(R)$ at $R = H(p_x)$ is strictly positive in general. □

In Figure 2.13, we plot the positive right derivative of $E_s(R)$ at $R = H(p_x)$ for Bernoulli (α) source where $\alpha \in (0, 0.25)$.

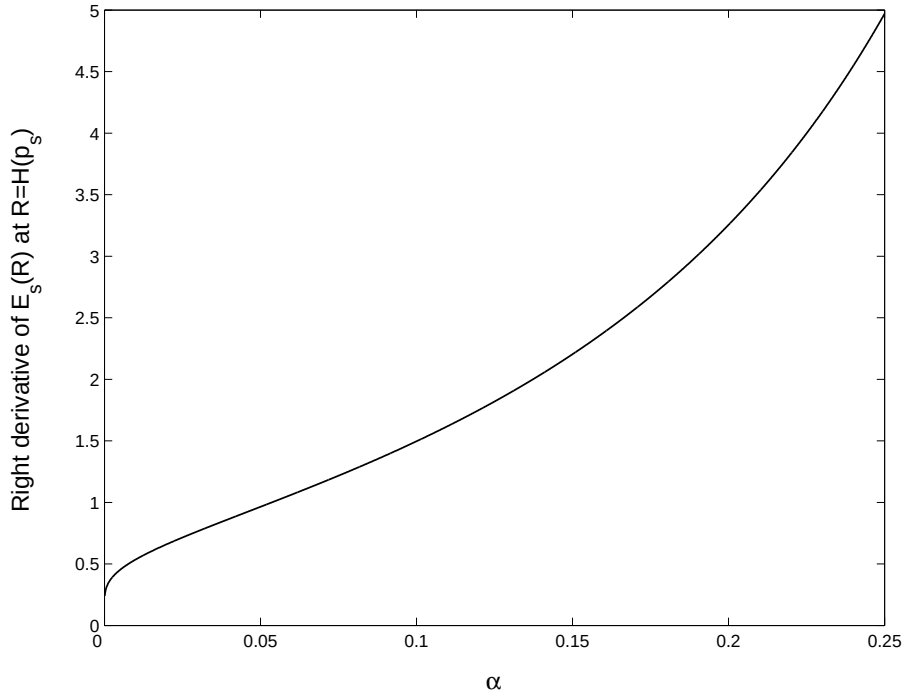

Figure 2.13. $\lim_{R \to H(p_x)^+} \frac{dE_s(R)}{dR}$ for Bernoulli (α) sources

The most striking difference between the block coding error exponent $E_{b,s}(R)$ and the delay constrained error exponent $E_s(R)$ is their behaviors in the high rate regime as illustrated in Figure 2.3, especially at $R = \log|\mathcal{X}|$. It is obvious that both error exponents are infinite (or mathematically strictly speaking: *not well defined*) when R is greater than $\log|\mathcal{X}|$ as shown in Proposition 1. The block coding error exponent $E_{b,s}R$ has a finite left limit at $R = \log|\mathcal{X}|$, the next proposition tells us that the delay constrained error exponent, however, has an infinite left limit at $R = \log|\mathcal{X}|$.

Proposition 10 *The left limit of the sequential error exponent is infinity as R approaches $\log|\mathcal{X}|$:* $\lim_{R \to \log|\mathcal{X}|} E_s(R) = \infty$

Proof: Notice that $E_b(R)$ is monotonically increasing for $R \in [0, \log|\mathcal{X}|)$ and $E_b(R) = \infty$ for $R > \log|\mathcal{X}|$. This implies that if $E_b((1+\alpha)R)$ is finite, α must be less than $\frac{\log|\mathcal{X}|}{R} - 1$. So we have for all $R \in [0, \log|\mathcal{X}|)$:

$$\begin{aligned} E_s(R) &= \inf_{\alpha > 0} \frac{1}{\alpha} E_b((1+\alpha)R) \\ &\geq \frac{1}{\frac{\log|\mathcal{X}|}{R} - 1} E_b(R) \end{aligned}$$

Thus the *left* limit of $E_s(R)$ at $R = \log|\mathcal{X}|$ is:

$$\begin{aligned}\lim_{R\to\log|\mathcal{X}|} E_s(R) &\geq \lim_{R\to\log|\mathcal{X}|} \frac{1}{\frac{\log|\mathcal{X}|}{R}-1} E_b(R)\\ &= \infty\end{aligned}$$

□

2.5.2 Conclusions and future work

In this chapter, we introduced the general setup of delay constrained source coding and the definition of delay-constrained error exponent. Unlike classical source coding, the source generates source symbols in a real time fashion and the performance of the coding system is measured by the probability of symbol error with a fixed-delay. The error exponent for lossless source coding with delay is completely characterized. The achievability part is proved by implementing a fixed-to-variable source encoder with a FIFO queue. Then the symbol error probability with delay is the same as the probability of an atypical queueing delay. The converse part is proved by constructing a block source coding system out of the delay-constrained source coding system. This simple idea can be applied to other source coding and channel coding problems and is called "focusing bound" in [67]. It was showed that this delay-constrained error exponent $E_s(R)$ has different properties than the classical block coding error exponent $E_{s,b}(R)$. We shown that $E_s(R)$ is strictly larger than $E_{s,b}(R)$ for all R. However, the delay constrained error exponent $E_s(R)$ is not completely understood. For example, we still do not know if $E_s(R)$ is convex $\cup$ on R. To further understand the properties of $E_s(R)$, one may find the rich literature in queueing theory [52] very useful. Particularly, the study of large deviations in queueing theory [42].

We also analyzed the delay constrained performance of the sequential random binning scheme. We proved that the standard random coding error exponent for block coding can be achieved using sequential random binning. Hence, sequential random binning is a suboptimal coding scheme for lossless source coding with delay. However, sequential random binning is a powerful technique which can be applied to other delay constrained coding problems such as distributed source coding [32, 12], source coding with side-information [16], joint source channel coding, multiple-access channel coding [14] and broadcast channel coding [15] under delay constraints. Some of these will be thoroughly studied in the next several chapters of this thesis. The decoding schemes have exponential complexity with *time*. However, [60] shows that the random coding error exponent is achievable using a

more computationally friendly stack-based decoding algorithm if the source distribution is known.

The most interesting result in this chapter is the "focusing" operator, which connects the delay constrained error exponent and the classical fixed-length block coding error exponent. A similar focusing bound was recently derived for delay constrained channel coding with feedback [67]. In order to understand for what problems this type of "focusing" bound applies to, we study a more general problem in the next chapter.

Chapter 3

Lossy Source Coding

Under the same streaming source model in Chapter 2, we study the delay constrained performance of lossy source coding where the decoder only needs to reconstruct the source to within a certain distortion. We derive the delay constrained error exponent through the "focusing " bound operator on the block error exponent for peak distortion measure. The proof in this chapter is more general than that in Chapter 2 and can be applied to delay constrained channel coding and joint source-channel coding.

3.1 Lossy source coding

As discussed in the review paper [6], lossy source coding is a source coding model where the estimate of the source at the decoder does not necessarily have to be exact. Unlike in Chapter 2, where the source and the reconstruction of the source are in the same alphabet set $\mathcal{X}$, for lossy source coding, the reconstruction of the source is in a different alphabet set $\mathcal{Y}$ where a distortion measure $d(\cdot,\cdot)$ is defined on $\mathcal{X}\times\mathcal{Y}$. We denote by $\mathcal{X}$ the source alphabet and $\mathcal{Y}$ the reconstruction alphabet. We also denote by y_i the reconstruction of x_i. This notational difference from Chapter 2 where we use $\widehat{x}_i$ as the reconstruction of x_i is to emphasize the fact that the reconstruction alphabet can be different from the source alphabet. Formally, a distortion measure is a non-negative function $d(x,y)$, $x\in\mathcal{X}$ and $y\in\mathcal{Y}$:

$$d:\mathcal{X}\times\mathcal{Y}\to[0,\infty)$$

The classical fixed length block coding results for lossy source coding are reviewed in

Section A.2 in the appendix. The core issue we are interested in for this chapter is the impact of causality and delay on lossy source coding. In [58], the rate distortion performance for a strictly zero delay *decoder* is studied, and it is shown that the optimal performance can be obtained by time-sharing between memoryless codes. Thus, it is in general strictly worse than the performance of classical fixed-block source coding that allows arbitrarily large delay. The large deviation performance of the zero delay decoder problem is studied in [57]. Allowing some finite end-to-end delay, [81, 80] shows that the average block coding rate distortion performance can still be approached exponentially with delay.

3.1.1 Lossy source coding with delay

As the common theme of this thesis, we consider a coding system for a streaming source, drawn iid from a distribution p_x on finite alphabet $\mathcal{X}$. The encoder, mapping source symbols into bits at fixed rate R, is causal and the decoder has to reconstruct the source symbols within a fixed end-to-end latency constraint Δ. The system is illustrated in Figure 5.3, which is the almost same as the lossless case in Figure 2.1 in Chapter 2. The only difference between the two figures is notational, we use y to indicate the reconstruction of the source x in this chapter, instead of $\widehat{x}$ in Chapter 2.

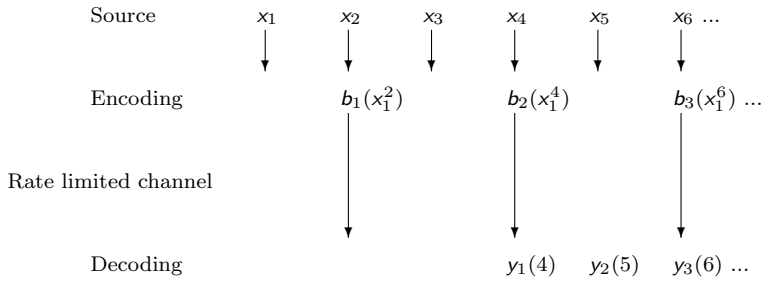

Figure 3.1. Time line of delay constrained source coding: rate $R = \frac{1}{2}$, delay $\Delta = 3$

Generalizing the notion of delay constrained error exponent (end-to-end delay performance) for lossless source coding in Definition 2, we have the following definition on delay constrained error exponent for lossy source coding.

Definition 4 *A rate R sequential source code shown in Figure 3.1 achieves error (distortion*

violation) exponent $E_D(R)$ with delay if for all $\epsilon > 0$, there exists $K < \infty$, s.t. $\forall i, \Delta > 0$

$$\Pr[d(x_i, y_i(i+\Delta)) > D] \leq K 2^{-\Delta(E_D(R)-\epsilon)}$$

The main goal of this chapter is to derive the error exponent defined above. The above lossy source coding system, in a way, has a symbol distortion constraint. This lead to an interesting relationship between the problem in Definition 4 and the lossy source coding under a peak distortion problem. We briefly introduce the relevant results of lossy source coding with a peak distortion in the next section.

3.1.2 Delay constrained lossy source coding error exponent

In this section, we present the main result of Chapter 3, the delay constrained error exponent for lossy source coding. This result first appeared in our paper [17]. We investigate the relation between delay Δ and the probability of distortion violation $\Pr[d(x_i, y_i(i+\Delta)) > D]$, where $y_i(i+\Delta)$ is the reconstruction of x_i at time $i+\Delta$ and D is the distortion constraint.

Theorem 2 *Consider fixed rate source coding of iid streaming data $x_i \sim p_x$, with a non-negative distortion measure d. For $D \in (\underline{D}, \overline{D})$, and rates $R \in (R(p_x, D), \overline{R}_D)$, the following error exponent with delay is optimal and achievable.*

$$E_D(R) \triangleq \inf_{\alpha > 0} \frac{1}{\alpha} E_D^b((\alpha+1)R) \tag{3.1}$$

where $E_D^b(R)$ is the block coding error exponent under peak distortion constraint, as defined in Lemma 6.

$\underline{D}$, $\overline{D}$ and $\overline{R}_D$ are defined in the next section.

3.2 A brief detour to peak distortion

In this section we introduce the peak distortion measure and present the relevant block coding result for lossy source coding with a peak distortion measure.

3.2.1 Peak distortion

Csiszár introduced the peak distortion measure [29] as an exercise problem (2.2.12), for a positive function d defined on finite alphabet set $\mathcal{X} \times \mathcal{Y}$, for $x_1^N \in \mathcal{X}^N$, $y_1^N \in \mathcal{Y}^N$, we define the distortion between x_1^N and y_1^N as $d(x_1^N, y_1^N)$ where

$$d(x_1^N, y_1^N) \triangleq \max_{1 \le i \le n} d(x_i, y_i) \tag{3.2}$$

This distortion measure reasonably reflects the reaction of human visual system [51]. Given some finite $D > 0$, we say y is a valid reconstruction of x if $d(x, y) \le D$.

The problem is only interesting if the target peak distortion D is higher than $\underline{D}$ and lower than $\overline{D}$, where

$$\underline{D} \triangleq \max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} d(x, y)$$

$$\overline{D} \triangleq \min_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} d(x, y).$$

If $D > \overline{D}$, then there exists a y^*, such that for all $x \in \mathcal{X}$, $d(x, y^*) \le D$, thus the decoder can reconstruct any source symbol x_i by y^* and the peak distortion requirement is still satisfied. On the other hand, if $D < \underline{D}$, then there exists an $x^* \in \mathcal{X}$, such that for all $y \in \mathcal{Y}$, $d(x^*, y) > D$. Now with the assumption that $p_X(x^*) > 0$, if N is big enough then with high probability that for some i, $x_i = x^*$. In this case, there is no reconstruction $y_1^N \in \mathcal{Y}^N$, such that $d(x_1^N, y_1^N) \le D$. Note that both $\underline{D}$ and $\overline{D}$ only depend on the distortion measure $d(\cdot, \cdot)$, not the source distribution p_X.

3.2.2 Rate distortion and error exponent for the peak distortion measure

cf. exercise (2.2.12) [29], the peak distortion problem can be treated as a special case of an average distortion problem. by defining a new distortion measure d_D on $\mathcal{X} \times \mathcal{Y}$, where

$$d_D(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{otherwise} \end{cases}$$

with average distortion level 0. And hence the rate distortion theorem and error exponent result follows the average distortion case. In [29], the rate distortion theorem for a peak distortion measure is derived.

Lemma 4 *The rate-distortion function $R(D)$ for peak distortion:*

$$R(p_X, D) \triangleq \min_{W \in \mathcal{W}_D} I(p_X, W) \tag{3.3}$$

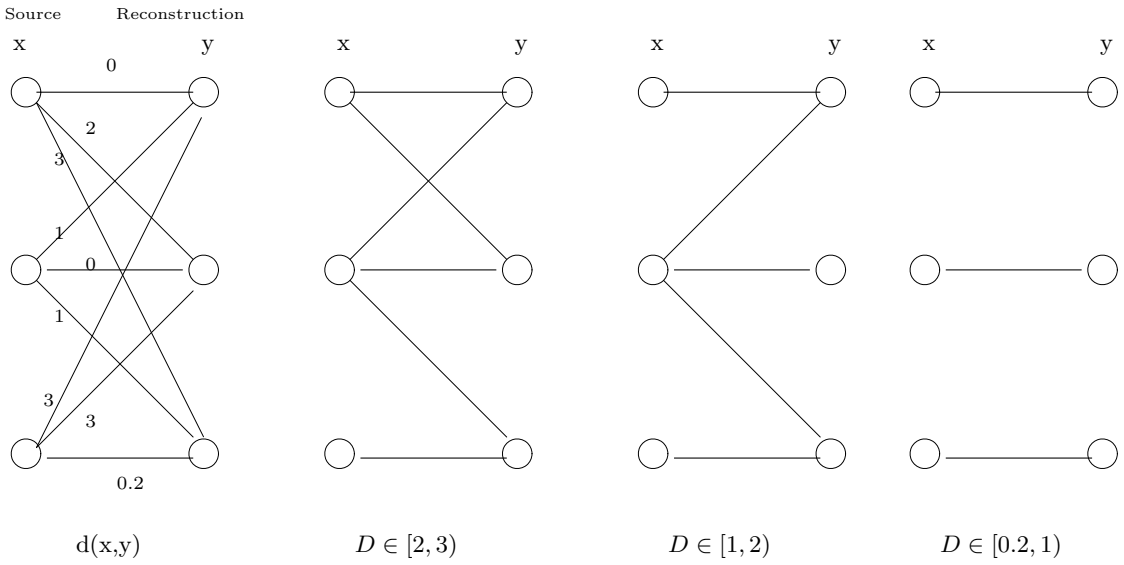

Figure 3.2. $d(x,y)$ and valid reconstructions under different peak distortion constraints D. (x,y) is linked if $d(x,y) \leq D$. $\underline{D} = 0.2$ and $\overline{D} = 3$. For $D \in [0.2,1)$, this is a lossless source coding problem.

where $\mathcal{W}_D$ is the set of all transition matrices that satisfy the peak distortion constraint, i.e. $\mathcal{W}_D = \{W : W(y|x) = 0, \textit{if } d(x,y) > D\}$.

Operationally, this lemma says, for any $\epsilon > 0$ and $\delta > 0$, for block length N big enough, there exists a code of rate $R(p_{\mathsf{x}}, D) + \epsilon$, such that the peak(maximum) distortion between the source string x_1^N and its reconstruction y_1^N is no bigger than D with probability at least $1 - \delta$.

This result can be viewed as a simple corollary of the rate distortion theorem in Theorem 10. Similar to the average distortion case, we have the following fixed-to-variable length coding result for peak distortion measure.

To have $\Pr[d(\mathsf{x}_1^N, \mathsf{y}_1^N) > D] = 0$, we can implement a universal variable length prefix-free code with code length $l_D(x_1^N)$ where

$$l_D(x_1^N) = n(R(p_{x_1^N}, D) + \delta_N) \tag{3.4}$$

where $p_{x_1^N}$ is the empirical distribution of x_1^N, and δ_N goes to 0 as N goes to infinity.

Like the average distortion case, this is a simple corollary of the type covering lemma [29, 28] which is derived from the Johnson SteinLovász theorem [23].

The rate distortion function $R(D)$ for average distortion measure is in general non-concave, non-$\cap$, in the source distribution p as pointed out in [55]. But for peak distortion, $R(p, D)$ *is* concave $\cap$ in p for a fixed distortion constraint D. The concavity of the rate distortion function under the peak distortion measure is an important property that that rate distortion functions of average distortion measure do not have.

Lemma 5 *$R(p, D)$ is concave $\cap$ in p for fixed D.*

Proof: The proof is in Appendix C. □

Now we study the large deviation properties of lossy source coding under the peak distortion measure. As a simple corollary of the block-coding error exponents for average distortion from [55], we have the following result.

Lemma 6 *Block coding error exponent under peak distortion:*

$$\begin{aligned} &\liminf_{n\to\infty} -\frac{1}{N}\log_2 \Pr[d(x_1^N, y_1^N) > D] = E_D^b(R) \\ &\text{where } E_D^b(R) \triangleq \min_{q_x: R(q_x, D) > R} D(q_x \| p_x) \end{aligned} \tag{3.5}$$

where y_1^N is the reconstruction of x_1^N using an optimal rate R code.

For lossless source coding, if $R > \log_2 |\mathcal{X}|$, the error probability is 0 and the error exponent is infinite. Similarly, for lossy source coding under peak distortion, the error exponent is infinite whenever

$$R > \overline{R}_D \triangleq \sup_{q_x} R(q_x, D)$$

where $\overline{R}_D$ only depends on $d(\cdot, \cdot)$ and D.

3.3 Numerical Results

Consider a distribution $p_x = \{0.7, 0.2, 0.1\}$ and a distortion measure on $\mathcal{X} \times \mathcal{Y}$ as shown in Figure 3.2. We plot the rate distortion $R - D$ curve under the peak distortion measure in Figure 3.3. The $R - D$ curve under peak distortion is a staircase function because the valid reconstruction is a $0 - 1$ function. The delay constrained lossy source coding error

exponents are shown in Figure 3.4. The delay constrained error exponent is higher than the block coding coding error exponent, which is also observed in the lossless case in Chapter 2.

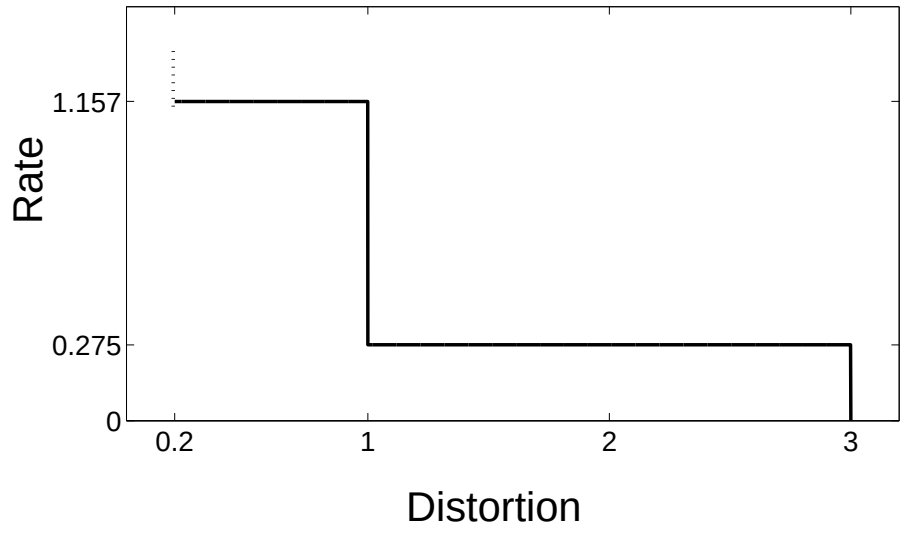

Figure 3.3. For $D_1 \in [1, 3)$, $R(p_x, D_1) = 0.275$ and $\overline{R}_{D_1} = 0.997$. For $D_2 \in [0.2, 1)$, the problem degenerates to lossless coding, so $R(p_x, D_1) = H(p_x) = 1.157$ and $\overline{R}_{D_1} = \log_2(3) = 1.585$

3.4 Proof of Theorem 2

In this section, we show that the error exponent in Theorem 2 is both achievable asymptotically with delay and that no better exponents are possible.

3.4.1 Converse

The proof of the converse is similar to the upper bound argument in Chapter 2 for lossless source coding with delay constraints. We leave the proof in Appendix D.

3.4.2 Achievability

We prove achievability by giving a universal coding scheme illustrated in Figure 3.5. This coding system looks almost identical to the delay constrained lossless source coding system in Section 2.4.1 as shown in Figure 2.11. The only difference is the fixed to variable

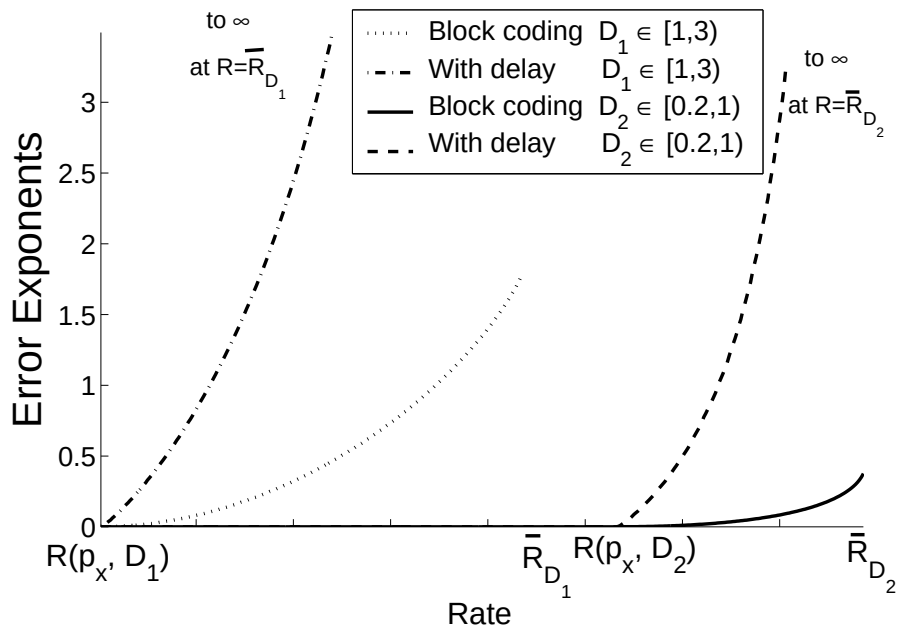

Figure 3.4. Error exponents of delay constrained lossy source coding and block source coding under peak distortion measure

length encoder. Instead of the universal optimal lossless encoder which is a one-to-one map from the source block space to the binary string space, we have a variable length lossy encoder in this chapter. So the average number of bits per symbol for an individual block is roughly the rate distortion function under peak distortion measure rather than the empirical entropy of that block.

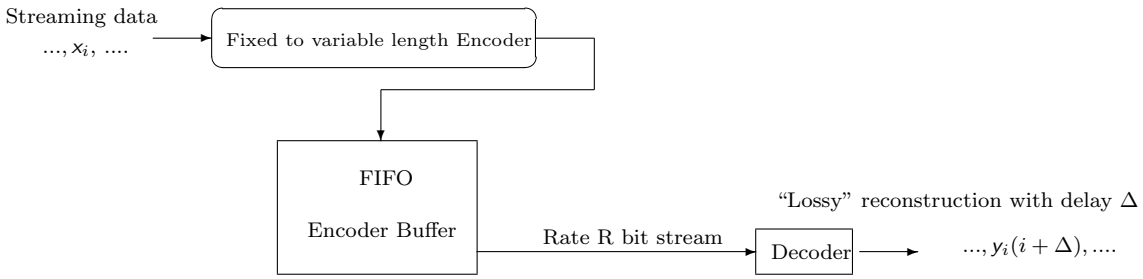

Figure 3.5. A delay optimal lossy source coding system.

A block-length N is chosen that is much smaller than the target end-to-end delays[1], while still being large enough. For a discrete memoryless source $\mathcal{X}$, distortion measure

[1]As always, we are interested in the performance with asymptotically large delays Δ.

$d(\cdot,\cdot)$, peak distortion constraint D, and large block-length N, we use the universal variable length prefix-free code in Proposition 4 to encode the i^{th} block $\vec{x}_i = x_{(i-1)N+1}^{iN} \in \mathcal{X}^N$. The code length $l_D(\vec{x}_i)$ is shown in (3.4),

$$NR(p_{\vec{x}_i}, D) \leq l_D(\vec{x}_i) \leq N(R(p_{\vec{x}_i}, D) + \delta_N) \tag{3.6}$$

The overhead δ_N is negligible for large N, since δ_N goes to 0 as N goes to infinity. The binary sequence describing the source is fed into a FIFO buffer described in Figure 3.5. The buffer is drained at a fixed rate R to obtain the encoded bits. Notice that if the buffer is empty, the output of the encoder buffer can be gibberish binary bits. The decoder simply discards these meaningless bits because it is aware that the buffer is empty. The decoder uses the bits it has received so far to get the reconstructions. If the relevant bits have not arrived by the time the reconstruction is due, it just guesses and we presume that a distortion-violation will occur.

As the following proposition indicates, the coding scheme is delay universal, i.e. the distortion-violation probability goes to 0 with exponent $E_D(R)$ for all source symbols and for all delays Δ big enough.

Proposition 11 *For the iid source $\sim p_x$, peak distortion constraint D, and large N, using the universal causal code described above, for all $\epsilon > 0$, there exists $K < \infty$, s.t. for all t, Δ:*

$$\Pr[d(\vec{x}_t, \vec{y}_t((t+\Delta)N)) > D] \leq K 2^{-\Delta N(E_D(R)-\epsilon)}$$

where $\vec{y}_t((t+\Delta)N)$ is the estimate of $\vec{x}_t$ at time $(t+\Delta)N$.

Before proving Proposition 11, we state the following lemma (proved in Appendix E) bounding the probability of atypical source behavior.

Lemma 7 *(Source atypicality) For all $\epsilon > 0$, block length N large enough, there exists $K < \infty$, s.t. for all n, for all $r < \overline{R}_D$:*

$$\Pr\left(\sum_{i=1}^{n} l_D(\vec{x}_i) > nNr\right) \leq K 2^{-nN(E_D^b(r)-\epsilon)} \tag{3.7}$$

Now we are ready to prove Proposition 11.

Proof: At time $(t+\Delta)N$, the decoder cannot decode $\vec{x}_t$ within peak distortion D only if the binary strings describing $\vec{x}_t$ are *not* all out of the buffer. Since the encoding buffer is FIFO, this means that the number of outgoing bits from some time t_1, where $t_1 \leq tN$ to $(t+\Delta)N$ is less than the number of the bits in the buffer at time t_1 plus the number of incoming bits from time t_1 to time tN. Suppose the buffer is last empty at time $tN - nN$ where $0 \leq n \leq t$, given this condition, the peak distortion is not satisfied only if $\sum_{i=0}^{n-1} l_D(\vec{x}_{t-i}) > (n+\Delta)NR$. Write $l_{D,\max}$ as the longest possible code length.

$$l_{D,max} \leq |\mathcal{X}| \log_2(N+1) + N \log_2 |\mathcal{X}|.$$

Then $\Pr[\sum_{i=0}^{n-1} l_D(\vec{x}_{t-i}) > (n+\Delta)NR] > 0$ only if $n > \frac{(n+\Delta)NR}{l_{D,max}} > \frac{\Delta NR}{l_{D,max}} \triangleq \beta\Delta$. So

$$\begin{aligned}
& \Pr\left(d(\vec{x}_t, \vec{y}_t((t+\Delta)N)) > D\right) \\
\leq & \sum_{n=\beta\Delta}^{t} \Pr[\sum_{i=0}^{n-1} l(\vec{x}_{t-i}) > (n+\Delta)NR] \qquad (3.8) \\
\leq_{(a)} & \sum_{n=\beta\Delta}^{t} K_1 2^{-nN(E_D^b(\frac{(n+\Delta)NR}{nN})-\epsilon_1)} \\
\leq_{(b)} & \sum_{n=\gamma\Delta}^{\infty} K_2 2^{-nN(E_D^b(R)-\epsilon_2)} + \sum_{n=\beta\Delta}^{\gamma\Delta} K_2 2^{-\Delta N(\min_{\alpha>1}\{\frac{E_D^b(\alpha R)}{\alpha-1}\}-\epsilon_2)} \\
\leq_{(c)} & K_3 2^{-\gamma\Delta N(E_D^b(R)-\epsilon_2)} + |\gamma\Delta - \beta\Delta| K_3 2^{-\Delta N(E_D(R)-\epsilon_2)} \\
\leq_{(d)} & K 2^{-\Delta N(E_D(R)-\epsilon)}
\end{aligned}$$

where $K_i's$ and $\epsilon_i's$ are properly chosen real numbers. (a) is true because of Lemma 7. Define $\gamma \triangleq \frac{E_D(R)}{E_D^b(R)}$. In the first part of (b), we only need the fact that $E_D^b(R)$ is non decreasing with R. In the second part of (b), we write $\alpha = \frac{n+\Delta}{n}$ and take the α to minimize the error exponents. The first term of (c) comes from the sum of a convergent geometric series and the second is by the definition of $E_D(R)$. (d) is by the definition of γ. ■

Combining the converse result in Proposition 12 in Appendix D and the achievability result in Proposition 11, we establish the desired results summarized in Theorem 2.

3.5 Discussions: why peak distortion measure?

For delay constrained lossy source coding, a "focusing" type bound is derived which is quite similar to its lossless source coding counterpart in the sense that they convert the block coding error exponent into a delay constrained error exponent through the same focusing

operator. As shown in Appendix E, the technical reason for the similarity is that the length of optimal variable-length codes, or equivalently the rate distortion functions, are concave $\cap$ in the empirical distribution for both lossless source coding and lossy source coding under peak distortion constraint. This is not the case for rate distortion functions under average distortion measures [29]. Thus the block error exponent for average distortion lossy source coding cannot be converted to delay constrained source lossy source coding by the same "focusing" operator. This leaves a great amount of future work to us.

The technical tools in this chapter are the same as those in Chapter 2, which are the large deviation principle and convex optimization methods. However, the block coding error exponent for lossy source coding under a peak distortion measure cannot be parameterized like the block lossless source coding error exponent. This poses a new challenge to our task. By only using the convexity and concavity of the relevant entities, but not the particular form of the error exponent, we develop a more general proof scheme for "focusing" type bounds. These techniques can be applied to other problems such as channel coding with perfect feedback, source coding with random arrivals and, undoubtedly, some other problems waiting to be discovered.

Chapter 4

Distributed Lossless Source Coding

In this chapter[1], we begin by reviewing classical results on the error exponents of the distributed source coding problem first studied by Slepian and Wolf [79]. Then we introduce the setup of delay constrained distributed source coding problem. We then present the main result of this chapter: achievable delay-constrained error exponents for delay constrained distributed source coding. The key techniques are sequential random binning which was introduced in Section 2.3.3 and a somewhat complicated way of partitioning error events. This is a "divide and conquer" proof. For each individual error event we use classical block code bounding techniques. We analyze both maximum likelihood decoding and universal decoding and show that the achieved exponents are equal. The proof of the equality of the two error exponents is in Appendix G. From the mathematical point of view, this proof is the most interesting and challenging part of this thesis. The key tools we use are the tilted-distribution from the statistics literature and the Lagrange dual from the convex optimization literature.

4.1 Distributed source coding with delay

In Section A.3 in the appendix, we review the distributed lossless source coding result in the block coding setup first discovered by David Slepian and Jack Wolf in their classical paper [79] which is now widely known as the Slepian-Wolf source coding problem. We then

[1]The result in this chapter is build upon a research project [12] jointly done by Cheng Chang, Stark Draper and Anant Sahai. I would like to thank Stark Draper for his great work in the project (especially the universal decoding part) and for allowing me to put this joint work in my thesis.

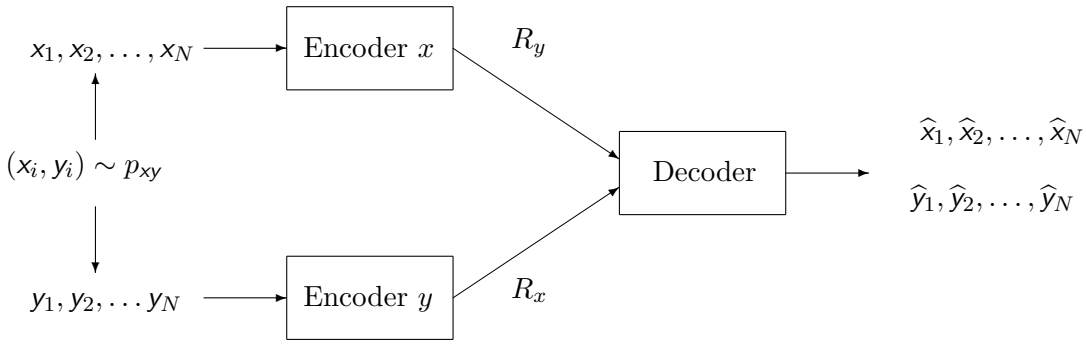

Figure 4.1. Slepian-Wolf distributed encoding and joint decoding of a pair of correlated sources.

introduce distributed source coding in the delay constrained framework. Then we give our result on error exponents with delay which first appeared in [12].

4.1.1 Lossless Distributed Source Coding with Delay Constraints

Parallelling to the point-to-point lossless source coding with delay problem introduced in Definition 1 in Section 2.1.1, we have the following setup for distributed source coding with delay constraints.

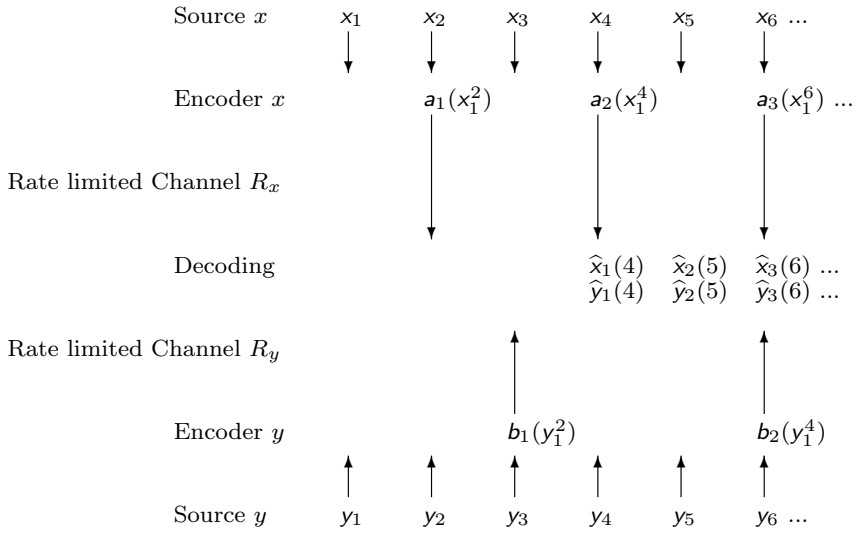

Figure 4.2. Delay constrained source coding with side-information: rates $R_x = \frac{1}{2}$, $R_y = \frac{1}{3}$, delay $\Delta = 3$

Definition 5 *A fixed-delay Δ sequential encoder-decoder triplet $(\mathcal{E}^x, \mathcal{E}^y, \mathcal{D})$ is a sequence of mappings, $\{\mathcal{E}_j^x\}, j = 1, 2, ...$, $\{\mathcal{E}_j^y\}, j = 1, 2, ...$ and $\{\mathcal{D}_j\}, j = 1, 2, ...$ The outputs of $\{\mathcal{E}_j^x\}$ and $\{\mathcal{E}_j^y\}$ from time $j-1$ to j.*

$$\begin{aligned} \mathcal{E}_j^x &: \mathcal{X}^j \longrightarrow \{0,1\}^{\lfloor jR_x \rfloor - \lfloor (j-1)R_x \rfloor}, \quad e.g., \quad \mathcal{E}_j^x(x_1^j) = a_{\lfloor (j-1)R_x \rfloor + 1}^{\lfloor jR_x \rfloor}, \\ \mathcal{E}_j^y &: \mathcal{Y}^j \longrightarrow \{0,1\}^{\lfloor jR_y \rfloor - \lfloor (j-1)R_y \rfloor}, \quad e.g., \quad \mathcal{E}_j^y(y_1^j) = b_{\lfloor (j-1)R_y \rfloor + 1}^{\lfloor jR_y \rfloor}. \end{aligned} \tag{4.1}$$

The output of the fixed-delay Δ decoder $\mathcal{D}_j$ is the decoding decision of x_j and y_j based on the received binary bits up to time $j + \Delta$.

$$\mathcal{D}_j : \{0,1\}^{\lfloor jR_x \rfloor} \times \{0,1\}^{\lfloor jR_y \rfloor} \longrightarrow \mathcal{X} \times \mathcal{Y}$$
$$\mathcal{D}_j(a^{\lfloor (j+\Delta)R_x \rfloor}, b^{\lfloor (j+\Delta)R_y \rfloor}) = (\widehat{x}_j(j+\Delta), \widehat{y}_j(j+\Delta))$$

Where $\widehat{x}_j(j+\Delta)$ and $\widehat{y}_j(j+\Delta)$ are the estimation of x_j and y_j at time $j+\Delta$ and thus has end-to-end delay of Δ seconds. A point-to-point delay constrained source coding system is illustrated in Figure 2.1.

In this chapter we study two symbol error probabilities. We define the pair of source estimates of the $(n-\Delta)^{th}$ source symbols $(x_{n-\Delta}, y_{n-\Delta})$ at time n as $(\widehat{x}_{n-\Delta}(n), \widehat{y}_{n-\Delta}(n)) = \mathcal{D}_n(\prod_{j=1}^n \mathcal{E}_j^x, \prod_{j=1}^n \mathcal{E}_j^y)$, where $\prod_{j=1}^n \mathcal{E}_j^x$ indicates the full $\lfloor nR_x \rfloor$ bit stream from encoder x up to time n. We use $(\widehat{x}_{n-\Delta}(n), \widehat{y}_{n-\Delta}(n))$ to indicate the estimate of the $(n-\Delta)^{th}$ symbol of each source stream at time n, the end to end delay is Δ. With these definitions the two error probabilities we study are

$$\Pr[\widehat{x}_{n-\Delta}(n) \neq x_{n-\Delta}] \text{ and } \Pr[\widehat{y}_{n-\Delta}(n) \neq y_{n-\Delta}].$$

Parallelling to the definition of delay constrained error exponent for point-to-point source coding in Definition 2, we have the following definition on delay constrained error exponents for distributed source coding.

Definition 6 *A family of rate R_x, R_y distributed sequential source codes $\{(\mathcal{E}^x, \mathcal{E}^y, \mathcal{D})\}$ are said to achieve delay error exponent $(E_x(R_x, R_y), E_y(R_x, R_y))$ if and only if for all $\epsilon > 0$, there exists $K < \infty$, s.t.$\forall i$, $\forall \Delta > 0$*

$$\Pr[\widehat{x}_{n-\Delta}(n) \neq x_{n-\Delta}] \leq K 2^{-\Delta(E_x(R_x,R_y)-\epsilon)} \text{ and } \Pr[\widehat{y}_{n-\Delta}(n) \neq y_{n-\Delta}] \leq K 2^{-\Delta(E_y(R_x,R_y)-\epsilon)}$$

Remark: The error exponents $E_x(R_x, R_y)$ and $E_y(R_x, R_y)$ are functions of both R_x and R_y. In contrast to (A.9) the error exponent we look at is in the delay, Δ, rather than total sequence length, n. Naturally, we define the delay constrained joint error exponent as follows, joint error exponent $E_{xy}(R_x, R_y)$ if and only if for all $\epsilon > 0$, there exists $K < \infty$, s.t.$\forall i$, $\forall \Delta > 0$

$$\Pr[(\widehat{\mathsf{x}}_{n-\Delta}(n), \widehat{\mathsf{y}}_{n-\Delta}(n)) \neq (\mathsf{x}_{n-\Delta}, \mathsf{y}_{n-\Delta})] \leq K 2^{-\Delta(E_{xy}(R_x,R_y)-\epsilon)}$$

By noticing the simple fact that:

$$\Pr[(\widehat{\mathsf{x}}_{n-\Delta}(n), \widehat{\mathsf{y}}_{n-\Delta}(n)) \neq (\mathsf{x}_{n-\Delta}, \mathsf{y}_{n-\Delta})] \leq 2 \max\{\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}], \Pr[\widehat{\mathsf{y}}_{n-\Delta}(n) \neq \mathsf{y}_{n-\Delta}]\}$$

and

$$\Pr[(\widehat{\mathsf{x}}_{n-\Delta}(n), \widehat{\mathsf{y}}_{n-\Delta}(n)) \neq (\mathsf{x}_{n-\Delta}, \mathsf{y}_{n-\Delta})] \geq \max\{\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}], \Pr[\widehat{\mathsf{y}}_{n-\Delta}(n) \neq \mathsf{y}_{n-\Delta}]\}$$

it should be clear that

$$E_{xy}(R_x, R_y) = \min\{E_x(R_x, R_y), E_y(R_x, R_y)\} \tag{4.2}$$

Following the point-to-point randomized sequential encoder-decoder pair defined in 3, we have the following definition of randomized sequential encoder-decoder triplets.

Definition 7 *A randomized sequential encoder-decoder triplets $(\mathcal{E}^x, \mathcal{E}^y, \mathcal{D})$ is a sequential encoder-decoder triplet defined in Definition 5 with common randomness, shared between encoder $\mathcal{E}^x$, $\mathcal{E}^y$ and decoder[2] $\mathcal{D}$. This allows us to randomize the mappings independent of the source sequences. We only need pair-wise independence, formally:*

$$\Pr[\mathcal{E}^x(x_1^i x_{i+1}^n) = \mathcal{E}^x(x_1^i \tilde{x}_{i+1}^n)] = 2^{-(\lfloor nR_x \rfloor - \lfloor iR_x \rfloor)} \leq 2 \times 2^{-(n-i)R_x} \tag{4.3}$$

where $x_{i+1} \neq \tilde{x}_{i+1}$, and

$$\Pr[\mathcal{E}^y(y_1^i y_{i+1}^n) = \mathcal{E}^y(y_1^i \tilde{y}_{i+1}^n)] = 2^{-(\lfloor nR_y \rfloor - \lfloor iR_y \rfloor)} \leq 2 \times 2^{-(n-i)R_y} \tag{4.4}$$

where $y_{i+1} \neq \tilde{y}_{i+1}$

[2]This common randomness is not shared between $\mathcal{E}^x$ and $\mathcal{E}^y$.

Recall the definition of bins in (2.11):

$$\mathcal{B}_x(x_1^n) = \{\tilde{x}_1^n \in \mathcal{X}^n : \mathcal{E}(\tilde{x}_1^n) = \mathcal{E}(x_1^n)\} \tag{4.5}$$

Hence (4.3) is equivalent to the following equality for $x_{i+1} \neq \tilde{x}_{i+1}$:

$$\Pr[\mathcal{E}(x_1^i \tilde{x}_{i+1}^n) \in \mathcal{B}_x(x_1^i x_{i+1}^n)] = 2^{-(\lfloor nR_x \rfloor - \lfloor iR_x \rfloor)} \leq 2 \times 2^{-(n-i)R_x} \tag{4.6}$$

Similar for source y:

$$\Pr[\mathcal{E}(y_1^i \tilde{y}_{i+1}^n) \in \mathcal{B}_y(y_1^i y_{i+1}^n)] = 2^{-(\lfloor nR_y \rfloor - \lfloor iR_y \rfloor)} \leq 2 \times 2^{-(n-i)R_y} \tag{4.7}$$

4.1.2 Main result of Chapter 4: Achievable error exponents

By using a randomized sequential encoder-decoder triplets in Definition 7, positive delay constrained error exponent pair (E_x, E_y) can be achieved if the rate pair (R_x, R_y) is in the classical Slepian-Wolf rate region [79], where both maximum likelihood decoding and universal decoding can achieve the positive error exponent pair. We summarize the main results of this chapter in the following two Theorems for maximum likelihood decoding and universal decoding respectively. Furthermore in Theorem 5, we show that the two decoding scheme achieve the same error exponent.

Theorem 3 *Let (R_x, R_y) be a rate pair such that $R_x > H(\mathsf{x}|\mathsf{y})$, $R_y > H(\mathsf{y}|\mathsf{x})$, $R_x + R_y > H(\mathsf{x}, \mathsf{y})$. Then, there exists a randomized encoder pair and maximum likelihood decoder triplet (per Definition 3) that satisfies the following three decoding criteria.*

(i) For all $\epsilon > 0$, there is a finite constant $K > 0$ such that
$\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}] \leq K 2^{-\Delta(E_{ML,SW,x}(R_x,R_y)-\epsilon)}$ *for all $n, \Delta \geq 0$ where*

$$E_{ML,SW,x}(R_x, R_y) = \min\left\{ \inf_{\gamma\in[0,1]} E_x^{ML}(R_x, R_y, \gamma), \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y^{ML}(R_x, R_y, \gamma) \right\}.$$

(ii) For all $\epsilon > 0$, there is a finite constant $K > 0$ such that
$\Pr[\widehat{\mathsf{y}}_{n-\Delta}(n) \neq \mathsf{y}_{n-\Delta}] \leq K 2^{-\Delta(E_{ML,SW,y}(R_x,R_y)-\epsilon)}$ *for all $n, \Delta \geq 0$ where*

$$E_{ML,SW,y}(R_x, R_y) = \min\left\{ \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_x^{ML}(R_x, R_y, \gamma), \inf_{\gamma\in[0,1]} E_y^{ML}(R_x, R_y, \gamma) \right\}.$$

(iii) For all $\epsilon > 0$ there is a finite constant $K > 0$ such that

$\Pr[(\widehat{\mathsf{x}}_{n-\Delta}(n), \widehat{\mathsf{y}}_{n-\Delta}(n)) \neq (\mathsf{x}_{n-\Delta}, \mathsf{y}_{n-\Delta})] \leq K2^{-\Delta(E_{ML,SW,xy}(R_x,R_y)-\epsilon)}$ *for all* $n, \Delta \geq 0$ *where*

$$E_{ML,SW,xy}(R_x, R_y) = \min\left\{ \inf_{\gamma\in[0,1]} E_x^{ML}(R_x, R_y, \gamma), \inf_{\gamma\in[0,1]} E_y^{ML}(R_x, R_y, \gamma) \right\}.$$

In definitions (i)–(iii),

$$\begin{aligned} E_x^{ML}(R_x, R_y, \gamma) &= \sup_{\rho\in[0,1]}[\gamma E_{x|y}(R_x, \rho) + (1-\gamma)E_{xy}(R_x, R_y, \rho)] \\ E_y^{ML}(R_x, R_y, \gamma) &= \sup_{\rho\in[0,1]}[\gamma E_{y|x}(R_x, \rho) + (1-\gamma)E_{xy}(R_x, R_y, \rho)] \end{aligned} \tag{4.8}$$

and

$$\begin{aligned} E_{xy}(R_x, R_y, \rho) &= \rho(R_x + R_y) - \log\Big[\textstyle\sum_{x,y} p_{\mathsf{xy}}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho} \\ E_{x|y}(R_x, \rho) &= \rho R_x - \log\Big[\textstyle\sum_y \Big[\sum_x p_{\mathsf{xy}}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big] \\ E_{y|x}(R_y, \rho) &= \rho R_y - \log\Big[\textstyle\sum_x \Big[\sum_y p_{\mathsf{xy}}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big] \end{aligned} \tag{4.9}$$

Theorem 4 *Let (R_x, R_y) be a rate pair such that $R_x > H(\mathsf{x}|\mathsf{y})$, $R_y > H(\mathsf{y}|\mathsf{x})$, $R_x + R_y > H(\mathsf{x},\mathsf{y})$. Then, there exists a randomized encoder pair and universal decoder triplet (per Definition 3) that satisfies the following three decoding criteria.*

(i) For all $\epsilon > 0$, there is a finite constant $K > 0$ such that

$\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}] \leq K2^{-\Delta(E_{UN,SW,x}(R_x,R_y)-\epsilon)}$ *for all* $n, \Delta \geq 0$ *where*

$$E_{UN,SW,x}(R_x, R_y) = \min\left\{ \inf_{\gamma\in[0,1]} E_x^{UN}(R_x, R_y, \gamma), \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y^{UN}(R_x, R_y, \gamma) \right\}. \tag{4.10}$$

(ii) For all $\epsilon > 0$, there is a finite constant $K > 0$ such that

$\Pr[\widehat{\mathsf{y}}_{n-\Delta}(n) \neq \mathsf{y}_{n-\Delta}] \leq K2^{-\Delta(E_{UN,SW,y}(R_x,R_y)-\epsilon)}$ *for all* $n, \Delta \geq 0$ *where*

$$E_{UN,SW,y}(R_x, R_y) = \min\left\{ \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_x^{UN}(R_x, R_y, \gamma), \inf_{\gamma\in[0,1]} E_y^{UN}(R_x, R_y, \gamma) \right\}. \tag{4.11}$$

(iii) For all $\epsilon > 0$, there is a finite constant $K > 0$ such that

$\Pr[(\widehat{\mathsf{x}}_{n-\Delta}(n), \widehat{\mathsf{y}}_{n-\Delta}(n)) \neq (\mathsf{x}_{n-\Delta}, \mathsf{y}_{n-\Delta})] \leq K2^{-\Delta(E_{UN,SW,xy}(R_x,R_y)-\epsilon)}$ *for all* $n, \Delta \geq 0$ *where*

$$E_{UN,SW,xy}(R_x, R_y) = \min\left\{ \inf_{\gamma\in[0,1]} E_x^{UN}(R_x, R_y, \gamma), \inf_{\gamma\in[0,1]} E_y^{UN}(R_x, R_y, \gamma) \right\}. \tag{4.12}$$

In definitions (i)–(iii),

$$\begin{aligned} E_x^{UN}(R_x, R_y, \gamma) &= \inf_{\tilde{x},\tilde{y},\bar{x},\bar{y}} \gamma D(p_{\tilde{x},\tilde{y}} \| p_{xy}) + (1-\gamma) D(p_{\bar{x},\bar{y}} \| p_{xy}) \\ &\quad + |\gamma[R_x - H(\tilde{x}|\tilde{y})] + (1-\gamma)[R_x + R_y - H(\bar{x},\bar{y})]|^+ \\ E_y^{UN}(R_x, R_y, \gamma) &= \inf_{\tilde{x},\tilde{y},\bar{x},\bar{y}} \gamma D(p_{\tilde{x},\tilde{y}} \| p_{xy}) + (1-\gamma) D(p_{\bar{x},\bar{y}} \| p_{xy}) \\ &\quad + |\gamma[R_y - H(\tilde{y}|\tilde{x})] + (1-\gamma)[R_x + R_y - H(\bar{x},\bar{y})]|^+ \end{aligned} \tag{4.13}$$

where the dummy random variables $(\tilde{x}, \tilde{y})$ *and* $(\bar{x}, \bar{y})$ *have joint distributions* $p_{\tilde{x},\tilde{y}}$ *and* $p_{\bar{x},\bar{y}}$*, respectively. Recall that the function* $|z|^+ = z$ *if* $z \geq 0$ *and* $|z|^+ = 0$ *if* $z < 0$*.*

Remark: We can compare the joint error event for block and streaming Slepian-Wolf coding, c.f. (4.12) with (A.10). The streaming exponent differs by the extra parameter γ that must be minimized over. If the minimizing $\gamma = 1$, then the block and streaming exponents are the same. The minimization over γ results from a fundamental difference in the types of error-causing events that can occur in streaming Slepian-Wolf as compared to block Slepian-Wolf.

Remark: The error exponents of maximum likelihood and universal decoding in Theorems 3 and 4 are the same. However, because there are new classes of error events possible in streaming, this needs proof. The equivalence is summarized in the following theorem.

Theorem 5 *Let* (R_x, R_x) *be a rate pair such that* $R_x > H(\mathsf{x}|\mathsf{y})$*,* $R_y > H(\mathsf{y}|\mathsf{x})$*, and* $R_x + R_y > H(\mathsf{x}, \mathsf{y})$*. Then,*

$$E_{ML,SW,x}(R_x, R_y) = E_{UN,SW,x}(R_x, R_y), \tag{4.14}$$

and

$$E_{ML,SW,x}(R_x, R_y) = E_{UN,SW,x}(R_x, R_y). \tag{4.15}$$

Theorem 5 follows directly from the following lemma, shown in the appendix.

Lemma 8 *For all* $\gamma \in [0, 1]$

$$E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma), \tag{4.16}$$

and

$$E_y^{ML}(R_x, R_y, \gamma) = E_y^{UN}(R_x, R_y, \gamma). \tag{4.17}$$

.

Proof: This is a very important Lemma. The proof is extremely laborious, so we put the details of the proof in Appendix G. The main tool used in the proof is convex optimization. There is a great deal of detailed analysis on tilted distribution in G.3 which is used throughout this thesis. ■

Remark: This theorem allows us to simplify notation. For example, we can define $E_x(R_x, R_y, \gamma)$ as $E_x(R_x, R_y, \gamma) = E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma)$, and can similarly define $E_y(R_x, R_y, \gamma)$. Further, since the ML and universal exponents are the same for the whole rate region we can define $E_{SW,x}(R_x, R_y)$ as $E_{SW,x}(R_x, R_y) = E_{ML,SW,x}(R_x, R_y) = E_{UN,SW,x}(R_x, R_y)$, and can similarly define $E_{SW,y}(R_x, R_y)$.

4.2 Numerical Results

To build insight into the differences between the sequential error exponents of Theorem 3 - 5 and block-coding error exponents, we give some examples of the exponents for binary sources.

For the point-to-point case, the error exponents of random sequential and block source coding are identical everywhere in the achievable rate region as can be seen by comparing Theorem 9 and Propositions 3 and 4. The same is true for source coding with decoder side-information which will be clear in Chapter 5. For distributed (Slepian-Wolf) source coding however, the sequential and block error exponents can be different. The reason for the discrepancy is that a new type of error event can be dominant in Slepian-Wolf source coding. This is reflected in Theorems 3 - 5 by the minimization over γ. Example 2 illustrates the impact of this γ term.

Given the sequential random source coding rule, for Slepian-Wolf source coding at very high rates, where $R_x > H(\mathsf{x})$, the decoder can ignore any information from encoder y and still decode x with with a positive error exponent. However, the decoder could also choose to decode source x and y jointly. Fig 4.6.a and 4.6.b illustrate that joint decoding may or surprisingly *may not* help decoding source x. This is seen by comparing the error exponent when the decoder ignores the side-information from encoder y (the dotted curves) to the joint error exponent (the lower solid curves). It seems that when the rate for source y is low, atypical behaviors of source y can cause joint decoding errors that end up corrupting x estimates. This holds for both block and sequential coding.

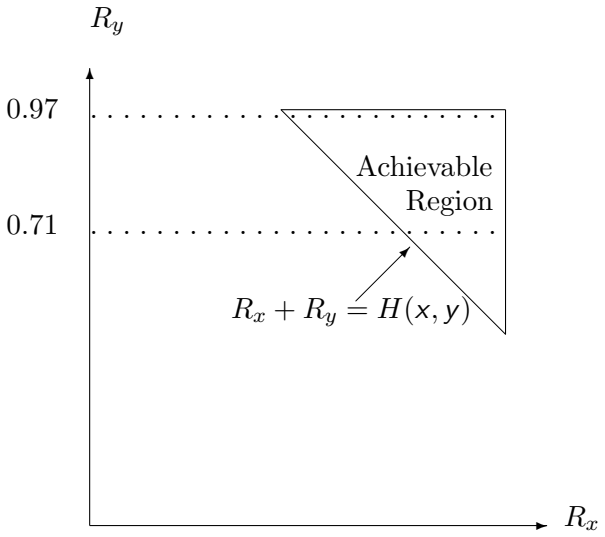

Figure 4.3. Rate region for the example 1 source, we focus on the error exponent on source x for fixed encoder y rates: $R_y = 0.71$ and $R_y = 0.97$

4.2.1 Example 1: symmetric source with uniform marginals

Consider a symmetric source where $|\mathcal{X}| = |\mathcal{Y}| = 2$, $p_{xy}(0,0) = 0.45$, $p_{xy}(0,1) = p_{xy}(1,0) = 0.05$ and $p_{xy}(1,1) = 0.45$. This is a marginally-uniform source: x is Bernoulli(1/2), y is the output from a BSC with input x, thus y is Bernoulli(1/2) as well. For this source $H(x) = H(y) = \log(2) = 1$, $H(x|y) = H(y|x) = 0.46$, $H(x,y) = 1.47$. The achievable rate region is the triangle shown in Figure(4.3).

For this source, as will be shown later, the dominant sequential error event is on the diagonal line in Fig 4.8. This is to say that:

$$E_{SW,x}(R_x, R_y) = E_{SW,x}^{BLOCK}(R_x, R_y) = E_x^{ML}(R_x, R_y, 0) = \sup_{\rho \in [0,1]} [E_{xy}(R_x, R_y, \rho)]. \quad (4.18)$$

Where $E_{SW,x}^{BLOCK}(R_x, R_y) = \min\{E_x^{ML}(R_x, R_y, 0), E_x^{ML}(R_x, R_y, 1)\}$ as shown in [39].

Similarly for source y:

$$E_{SW,y}(R_x, R_y) = E_{SW,y}^{BLOCK}(R_x, R_y) = E_y^{ML}(R_x, R_y, 0) = \sup_{\rho \in [0,1]} [E_{xy}(R_x, R_y, \rho)]. \quad (4.19)$$

We first show that for this source $\forall \rho \geq 0$, $E_{x|y}(R_x, \rho) \geq E_{xy}(R_x, R_y, \rho)$. By definition:

$$\begin{aligned}
&E_{x|y}(R_x, \rho) - E_{xy}(R_x, R_y, \rho) \\
&\quad = \rho R_x - \log\Big[\sum_y \Big[\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big] - \Big(\rho(R_x+R_y) - \log\Big[\sum_{x,y} p_{xy}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big) \\
&\quad = -\rho R_y - \log\Big[2\Big[\sum_x p_{xy}(x,0)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\Big] + \log\Big[2\sum_x p_{xy}(x,0)^{\frac{1}{1+\rho}}\Big]^{1+\rho} \\
&\quad = -\rho R_y - \log\Big[2\Big] + \log\Big[2\Big]^{1+\rho} \\
&\quad = \rho(\log[2] - R_y) \\
&\quad \geq 0
\end{aligned}$$

The last inequality is true because we only consider the problem when $R_y \leq \log|\mathcal{Y}|$. Otherwise, y is better viewed as perfectly known side-information. Now

$$\begin{aligned}
E_x^{ML}(R_x, R_y, \gamma) &= \sup_{\rho\in[0,1]} [\gamma E_{x|y}(R_x,\rho) + (1-\gamma)E_{xy}(R_x, R_y, \rho)] \\
&\geq \sup_{\rho\in[0,1]} [E_{xy}(R_x, R_y, \rho)] \\
&= E_x^{ML}(R_x, R_y, 0)
\end{aligned}$$

Similarly $E_y^{ML}(R_x, R_y, \gamma) \geq E_y^{ML}(R_x, R_y, 0) = E_x^{ML}(R_x, R_y, 0)$. Finally,

$$\begin{aligned}
E_{SW,x}(R_x, R_y) &= \min\left\{\inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma), \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x, R_y, \gamma)\right\} \\
&= E_x^{ML}(R_x, R_y, 0)
\end{aligned}$$

Particularly $E_x(R_x, R_y, 1) \geq E_x(R_x, R_y, 0)$, so

$$\begin{aligned}
E_{SW,x}^{BLOCK}(R_x, R_y) &= \min\{E_x^{ML}(R_x, R_y, 0), E_x^{ML}(R_x, R_y, 1)\} \\
&= E_x^{ML}(R_x, R_y, 0)
\end{aligned}$$

The same proof holds for source y.

In Fig 4.4 we plot the joint sequential/block coding error exponents $E_{SW,x}(R_x, R_y) = E_{SW,x}^{BLOCK}(R_x, R_y)$, the error exponents are positive iff $R_x > H(xy) - R_y = 1.47 - R_y$.

4.2.2 Example 2: non-symmetric source

Consider a non-symmetric source where $|\mathcal{X}| = |\mathcal{Y}| = 2$, $p_{xy}(0,0) = 0.1$, $p_{xy}(0,1) = p_{xy}(1,0) = 0.05$ and $p_{xy}(1,1) = 0.8$. For this source $H(x) = H(y) = 0.42$,

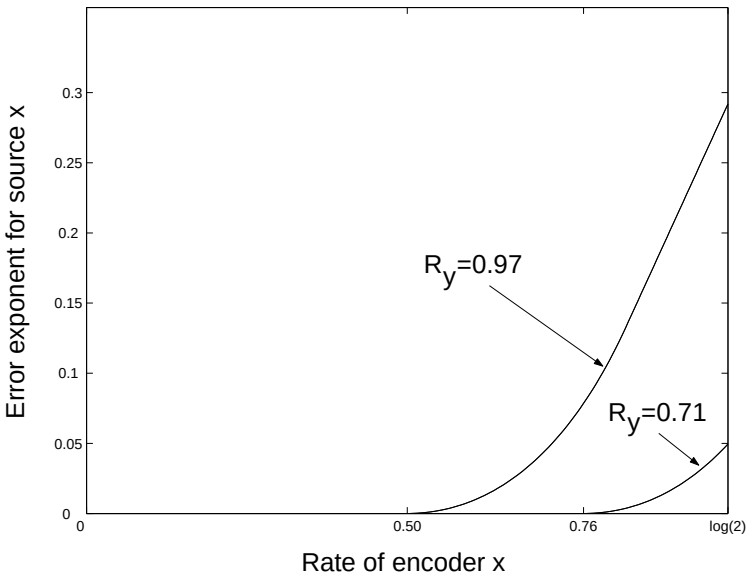

Figure 4.4. Error exponents plot: $E_{SW,x}(R_x, R_y)$ plotted for $R_y = 0.71$ and $R_y = 0.97$ $E_{SW,x}(R_x, R_y) = E_{SW,x}^{BLOCK}(R_x, R_y) = E_{SW,y}(R_x, R_y) = E_{SW,y}^{BLOCK}(R_x, R_y)$ and $E_x(R_x) = 0$

$H(x|y) = H(y|x) = 0.29$ and $H(x, y) = 0.71$. The achievable rate region is shown in Fig 4.5. In Fig 4.6.a, 4.6.b, 4.6.c and 4.6.d, we compare the joint sequential error exponent $E_{SW,x}(R_x, R_y)$ the joint block coding error exponent $E_{SW,x}^{BLOCK}(R_x, R_y) = \min\{E_x(R_x, R_y, 0), E_x(R_x, R_y, 1)\}$ as shown in [39] and the individual error exponent for source X, $E_x(R_x)$ as shown in Corollary 4. Notice that $E_x(R_x) > 0$ only if $R_x > H(x)$. In Fig 4.7, we compare the sequential error exponent for source y: $E_{SW,y}(R_x, R_y)$ and the block coding error exponent for source y: $E_{SW,y}^{BLOCK}(R_x, R_y) = \min\{E_y(R_x, R_y, 0), E_y(R_x, R_y, 1)\}$ and $E_y(R_y)$ which is a constant since we fix R_y.

For $R_y = 0.50$ as shown in Fig 4.6.a.b and 4.7.a.b, the difference between the block coding and sequential coding error exponents is very small for both source x and y. More interestingly, as shown in Fig 4.6.a, because the rate of source y is low, i.e. it is more likely to get a decoding error due to the atypical behavior of source y. So as R_x increases, it is sometimes better to ignore source y and decode x individually. This is evident as the dotted curve is above the solid curves.

For $R_y = 0.71$ as shown in Fig 4.6.c.d and 4.7.c.d, since the rate for source y is high enough, source y can be decoded with a positive error exponent individually as shown in Fig 4.7.c. But as the rate of source x increases, joint decoding gives a better error exponent. When R_x is very high, then we observe the saturation of the error exponent on y as if source

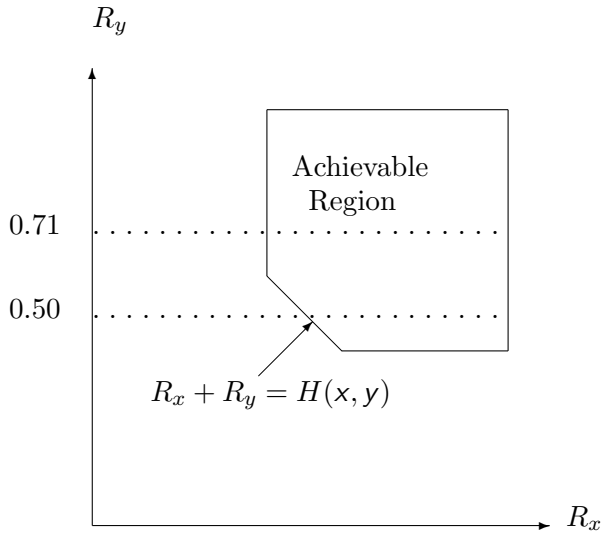

Figure 4.5. Rate region for the example 2 source, we focus on the error exponent on source x for fixed encoder y rates: $R_y = 0.50$ and $R_y = 0.71$

x is known perfectly to the decoder! This is illustrated by the flat part of the solid curves in Fig 4.7.c.

4.3 Proofs

In this section we provide the proofs of Theorems 3 and 4, which consider the delay constrained distributed source coding. As with the proofs for the point-to-point source coding result in Propositions 3 and 4 in Section 2.3.3, we start by proving the case for maximum likelihood decoding. The universal result follows.

4.3.1 ML decoding

ML decoding rule

First we explain the simple maximum likelihood decoding rule. Very similar to the ML decoding rule for the point-to-point source coding case in (2.16), the decoding rule is to simply pick the most likely source sequence pair under the joint distribution p_{xy}.

Denote by $(\widehat{x}_1^n(n), \widehat{y}_1^n(n))$ the estimate of the source sequence pair (x_1^n, y_1^n) at time n.

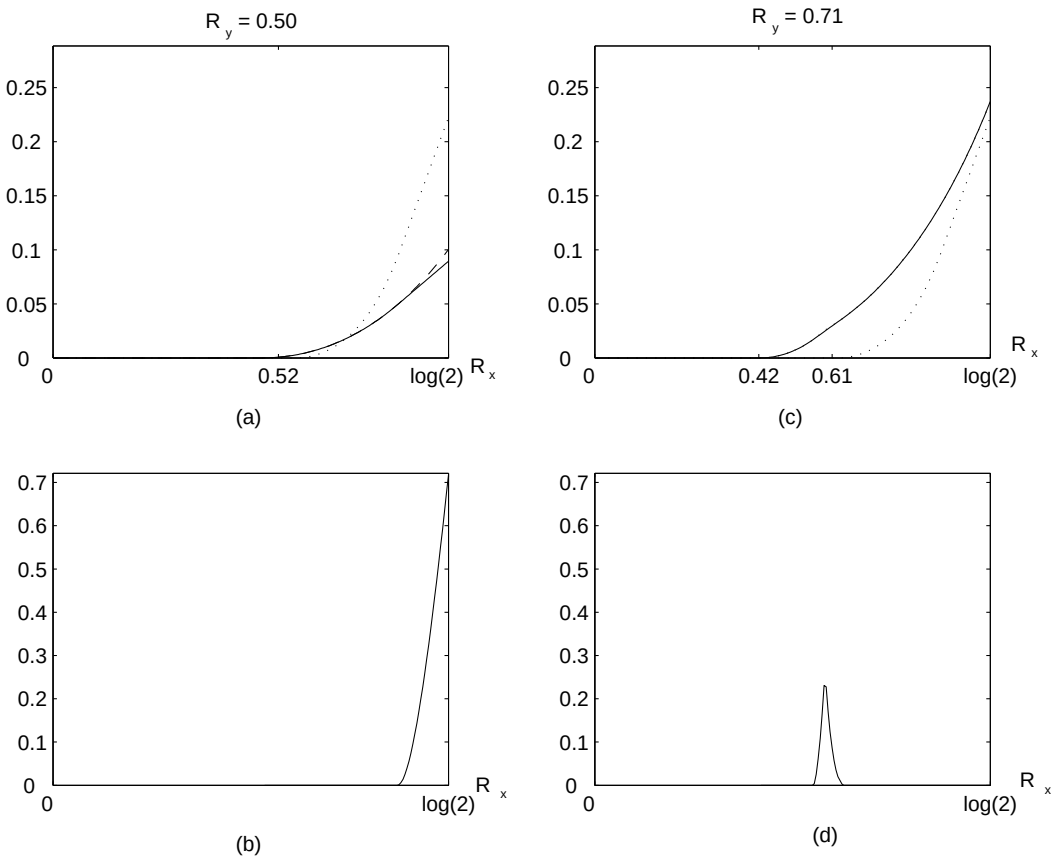

Figure 4.6. Error exponents plot for source x for fixed R_y as R_x varies:
$R_y = 0.50$:
(a) Solid curve: $E_{SW,x}(R_x, R_y)$, dashed curve $E_{SW,x}^{BLOCK}(R_x, R_y)$ and dotted curve: $E_x(R_x)$, notice that $E_{SW,x}(R_x, R_y) \leq E_{SW,x}^{BLOCK}(R_x, R_y)$ but the difference is small.
(b) $10\log_{10}(\frac{E_{SW,x}^{BLOCK}(R_x,R_y)}{E_{SW,x}(R_x,R_y)})$. This shows the difference is there at high rates.
$R_y = 0.71$:
(c) Solid curve $E_{SW,x}(R_x, R_y)$, dashed curve $E_{SW,x}^{BLOCK}(R_x, R_y)$ and dotted curve: $E_x(R_x)$, again $E_{SW,x}(R_x, R_y) \leq E_{SW,x}^{BLOCK}(R_x, R_y)$ but the difference is extremely small.
(d) $10\log_{10}(\frac{E_{SW,x}^{BLOCK}(R_x,R_y)}{E_{SW,x}(R_x,R_y)})$. This shows the difference is there at intermediate low rates.

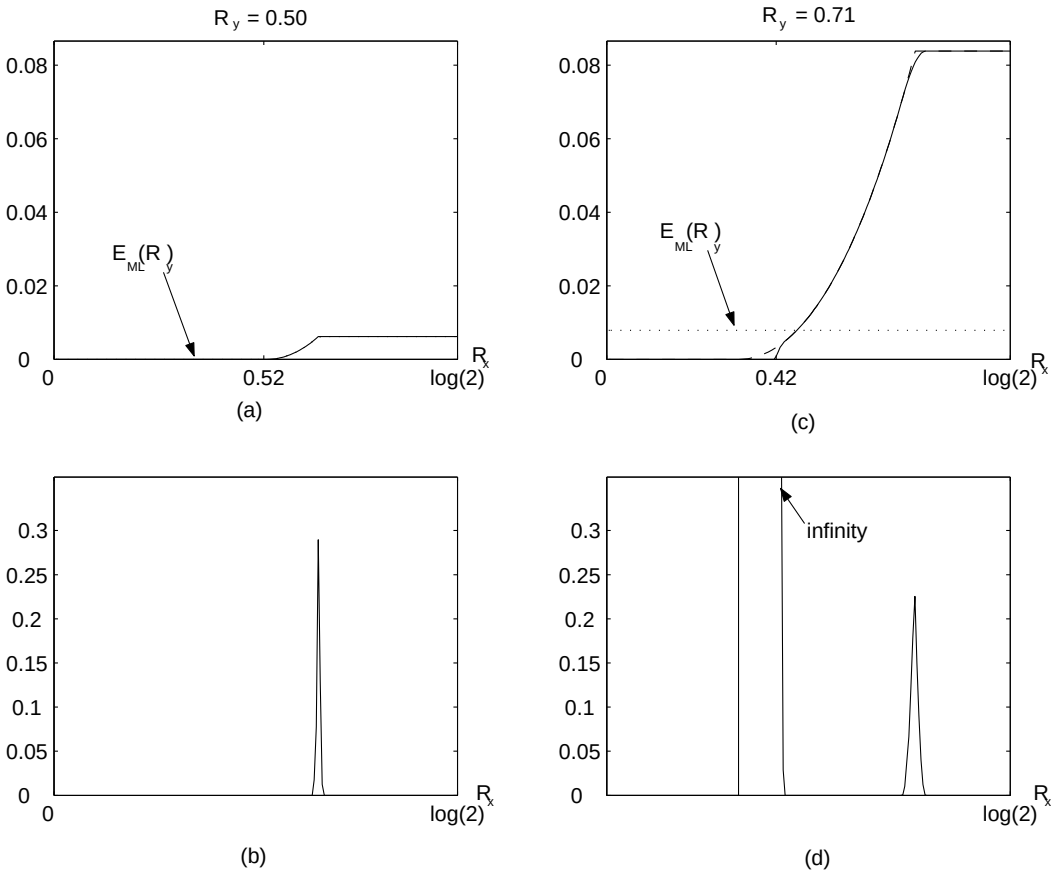

Figure 4.7. Error exponents plot for source y for fixed R_y as R_x varies:
$R_y = 0.50$:
(a) Solid curve: $E_{SW,y}(R_x, R_y)$ and dashed curve $E_{SW,y}^{BLOCK}(R_x, R_y)$, $E_{SW,y}(R_x, R_y) \leq E_{SW,y}^{BLOCK}(R_x, R_y)$, the difference is extremely small. $E_y(R_y)$ is 0 because $R_y = 0.50 < H(y)$.
(b) $10 \log_{10}(\frac{E_{SW,y}^{BLOCK}(R_x,R_y)}{E_{SW,y}(R_x,R_y)})$. This shows the two exponents are not identical everywhere.
$R_y = 0.71$:
(c) Solid curves: $E_{SW,y}(R_x, R_y)$, dashed curve $E_{SW,y}^{BLOCK}(R_x, R_y)$ and $E_{SW,y}(R_x, R_y) \leq E_{SW,y}^{BLOCK}(R_x, R_y)$ and $E_y(R_y)$ is constant shown in a dotted line.
(d) $10 \log_{10}(\frac{E_{SW,y}^{BLOCK}(R_x,R_y)}{E_{SW,y}(R_x,R_y)})$. Notice how the gap goes to infinity when we leave the Slepian-Wolf region.

$$(\widehat{x}_1^n(n), \widehat{y}_1^n(n)) = \underset{\tilde{x}_1^n \in \mathcal{B}_x(x_1^n), \tilde{y}_1^n \in \mathcal{B}_y(y_1^n)}{\arg\max} p_{xy}(\tilde{x}_1^n, \tilde{y}_1^n) = \underset{\tilde{s}_1^n \in \mathcal{B}_x(x_1^n), \tilde{y}_1^n \in \mathcal{B}_y(y_1^n)}{\arg\max} \prod_{i=1}^{n} p_{xy}(\tilde{s}_i, \tilde{y}_i) \quad (4.20)$$

At time n, the decoder simply picks the sequence pair $\widehat{x}_1^n(n)$ and $\widehat{y}_1^n(n)$ with the highest likelihood which is in the same bin as the true sequence x_1^n and y_1^n respectively. Now the estimate of source symbol $n-\Delta$ is simply the $(n-\Delta)^{th}$ symbol of $\widehat{x}_1^n(n)$ and $\widehat{y}_1^n(n)$, denoted by $(\widehat{x}_{n-\Delta}(n), \widehat{y}_{n-\Delta}(n))$.

Details of the proof of Theorem 3

In Theorems 3 and 4 three error events are considered: (i) $[\mathsf{x}_{n-\Delta} \neq \widehat{\mathsf{x}}_{n-\Delta}(n)]$, (ii) $[\mathsf{y}_{n-\Delta} \neq \widehat{\mathsf{y}}_{n-\Delta}(n)]$, and (iii) $[(\mathsf{x}_{n-\Delta}, \mathsf{y}_{n-\Delta}) \neq (\widehat{\mathsf{x}}_{n-\Delta}(n), \widehat{\mathsf{y}}_{n-\Delta}(n))]$. We develop the error exponent for case (i). The error exponent for case (ii) follows from a similar derivation, and that of case (iii) is the minimum of the exponents of cases (i) and (ii) by the simple union bound argument in 4.2.

To lead to the decoding error $[\mathsf{x}_{n-\Delta} \neq \widehat{\mathsf{x}}_{n-\Delta}(n)]$ there must be some spurious source pair $(\tilde{x}_1^n, \tilde{y}_1^n)$ that satisfies three conditions: (i) $\tilde{x}_1^n \in \mathcal{B}_x(x_1^n)$ and $\tilde{y}_1^n \in \mathcal{B}_y(y_1^n)$, (ii) it must be more likely than the true pair $p_{xy}(\tilde{x}_1^n, \tilde{y}_1^n) > p_{xy}(x_1^n, y_1^n)$, and (iii) $\tilde{x}_l \neq x_l$ for some $l \leq n-\Delta$.

The error probability is

$$\begin{aligned}
&\Pr[\mathsf{x}_{n-\Delta} \neq \widehat{\mathsf{x}}_{n-\Delta}(n)] \leq \Pr[\widehat{\mathsf{x}}_1^{n-\Delta}(n) \neq \mathsf{x}_1^{n-\Delta}] \\
&\quad = \sum_{x_1^n, y_1^n} \Pr[\widehat{\mathsf{x}}^{n-\Delta} \neq x^{n-\Delta} | \mathsf{x}_1^n = x_1^n, \mathsf{y}_1^n = y_1^n] p_{xy}(x_1^n, y_1^n) \\
&\quad \leq \sum_{x_1^n, y_1^n} p_{xy}(x_1^n, y_1^n) \Big\{ \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} \\
&\qquad \Pr\left[\exists\, (\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n) \cap \mathcal{F}_n(l, k, x_1^n, y_1^n) \text{ s.t. } p_{xy}(\tilde{x}_1^n, \tilde{y}_1^n) \geq p_{xy}(x_1^n, y_1^n)\right] \Big\} \qquad (4.21) \\
&\quad = \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} \Big\{ \sum_{x_1^n, y_1^n} p_{xy}(x_1^n, y_1^n) \\
&\qquad \Pr\left[\exists\, (\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n) \cap \mathcal{F}_n(l, k, x_1^n, y_1^n) \text{ s.t. } p_{xy}(\tilde{x}_1^n, \tilde{y}_1^n) \geq p_{xy}(x_1^n, y_1^n)\right] \Big\} \\
&= \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} p_n(l, k). \qquad (4.22)
\end{aligned}$$

In (4.21) we decompose the error event into a number of mutually exclusive events by partitioning all source pairs $(\tilde{x}_1^n, \tilde{y}_1^n)$ into sets $\mathcal{F}_n(l, k, x_1^n, y_1^n)$ defined by the times l and k

at which $\tilde{x}_1^n$ and $\tilde{y}_1^n$ diverge from the realized source sequences. The set $\mathcal{F}_n(l,k,x_1^n,y_1^n)$ is defined as

$$\mathcal{F}_n(l,k,x^n,y^n) = \{(\bar{x}_1^n,\bar{y}_1^n) \in \mathcal{X}^n \times \mathcal{Y}^n \text{ s.t. } \bar{x}_1^{l-1} = x_1^{l-1}, \bar{x}_l \neq x_l, \bar{y}_1^{k-1} = y_1^{k-1}, \bar{y}_k \neq y_k\}, \tag{4.23}$$

In contrast to streaming point-to-point or side-information coding (cf. (4.23) with (2.20)), the partition is now doubly-indexed. To find the dominant error event, we must search over both indices. Having two dimensions to search over results in an extra minimization when calculating the error exponent (and leads to the infimum over γ in Theorem 3).

Finally, to get (4.22) we define $p_n(l,k)$ as

$$\begin{aligned} p_n(l,k) &= \sum_{x_1^n,y_1^n} p_{xy}(x_1^n,y_1^n) \times \\ &\quad \Pr\left[\exists\, (\tilde{x}_1^n,\tilde{y}_1^n) \in \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n) \cap \mathcal{F}_n(l,k,x_1^n,y_1^n) \text{ s.t. } p_{xy}(\tilde{x}_1^n,\tilde{y}_1^n) \geq p_{xy}(x_1^n,y_1^n)\right]. \end{aligned}$$

The following lemma provides an upper bound on $p_n(l,k)$:

Lemma 9

$$\begin{aligned} p_n(l,k) &\leq 4 \times 2^{-(n-l+1)E_x(R_x,R_y,\frac{k-l}{n-l+1})} \quad \textit{if} \quad l \leq k, \\ p_n(l,k) &\leq 4 \times 2^{-(n-k+1)E_y(R_x,R_y,\frac{l-k}{n-k+1})} \quad \textit{if} \quad l \geq k, \end{aligned} \tag{4.24}$$

where $E_x(R_x,R_y,\gamma)$ and $E_y(R_x,R_y,\gamma)$ are defined in (4.8) and (4.9) respectively. Notice that $l,k \leq n$, for $l \leq k$: $\frac{k-l}{n-l+1} \in [0,1]$ serves as γ in the error exponent $E_x(R_x,R_y,\gamma)$. Similarly for $l \geq k$.

Proof: The proof is quite similar to the Chernoff bound [8] argument in the proof of Proposition 3 for the point to point source coding problem. We put the details of the proof in Appendix F.1. □

We use Lemma 9 together with (4.22) to bound $\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}]$ which is an upper bound on $\Pr[\widehat{x}_{n-\Delta}(n) \neq x_{n-\Delta}]$ for two distinct cases. The first, simpler case, is when $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma) > \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)$. To bound $\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}]$ in this case, we split the sum over the $p_n(l,k)$ into two terms, as visualized in Fig 4.8. There are $(n+1)\times(n-\Delta)$ such events to account for (those inside the box). The probability of the event within each oval are summed together to give an upper bound on $\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}]$. We add extra probabilities outside of the box but within the ovals to make the summation symmetric thus simpler. Those extra error events do not impact the error exponent because

$\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\rho,\gamma) \geq \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\rho,\gamma)$. The possible dominant error events are highlighted in Figure 4.8. Thus,

$$\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}] \leq \sum_{l=1}^{n-\Delta}\sum_{k=l}^{n+1} p_n(l,k) + \sum_{k=1}^{n-\Delta}\sum_{l=k}^{n+1} p_n(l,k) \tag{4.25}$$

$$\leq 4\times\sum_{l=1}^{n-\Delta}\sum_{k=l}^{n+1} 2^{-(n-l+1)\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)} + 4\times\sum_{k=1}^{n-\Delta}\sum_{l=k}^{n+1} 2^{-(n-k+1)\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)} \tag{4.26}$$

$$= 4\times\sum_{l=1}^{n-\Delta}(n-l+2)2^{-(n-l+1)\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)} + 4\times\sum_{k=1}^{n-\Delta}(n-k+2)2^{-(n-k+1)\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)}$$

$$\leq 8\sum_{l=1}^{n-\Delta}(n-l+2)2^{-(n-l+1)\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)} \tag{4.27}$$

$$\leq \sum_{l=1}^{n-\Delta} C_1 2^{-(n-l+2)[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)-\epsilon]} \tag{4.28}$$

$$\leq C_2 2^{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)-\epsilon]} \tag{4.29}$$

(4.25) follows directly from (4.22), in the first term $l \leq k$, in the second term $l \geq k$. In (4.26), we use Lemma 9. In (4.27) we use the assumption that $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma) > \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)$. In (4.28) the $\epsilon > 0$ results from incorporating the polynomial into the first exponent, and can be chosen as small as desired. Combining terms and summing out the decaying exponential yield the bound (4.29).

The second, more involved case, is when $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\rho,\gamma) < \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\rho,\gamma)$. To bound $\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}]$, we could use the same bounding technique used in the first case. This gives the error exponent $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)$ which is generally smaller than what we can get by dividing the error events in a new scheme as shown in Figure 4.9. In this situation we split (4.22) into three terms, as visualized in Fig 4.9. Just as in the first case shown in Fig 4.8, there are $(n+1)\times(n-\Delta)$ such events to account for (those inside the box). The error events are partitioned into 3 regions. Region 2 and 3 are separated by $k^*(l)$ using a dotted line. In region 3, we add extra probabilities outside of the box but within the ovals to make the summation simpler. Those extra error events do not affect the error exponent as shown in

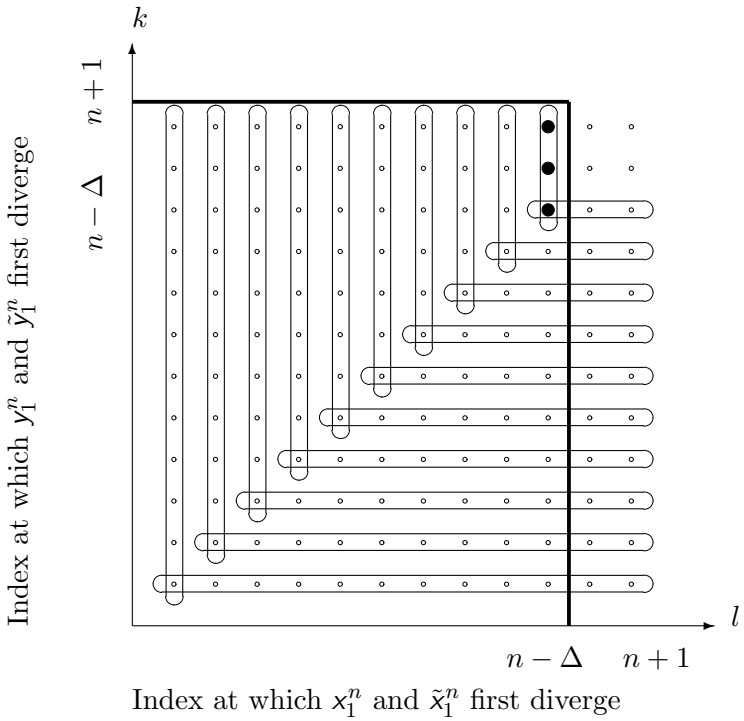

Figure 4.8. Two dimensional plot of the error probabilities $p_n(l,k)$, corresponding to error events (l,k), contributing to $\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}]$ in the situation where $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\rho,\gamma) \geq \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\rho,\gamma)$.

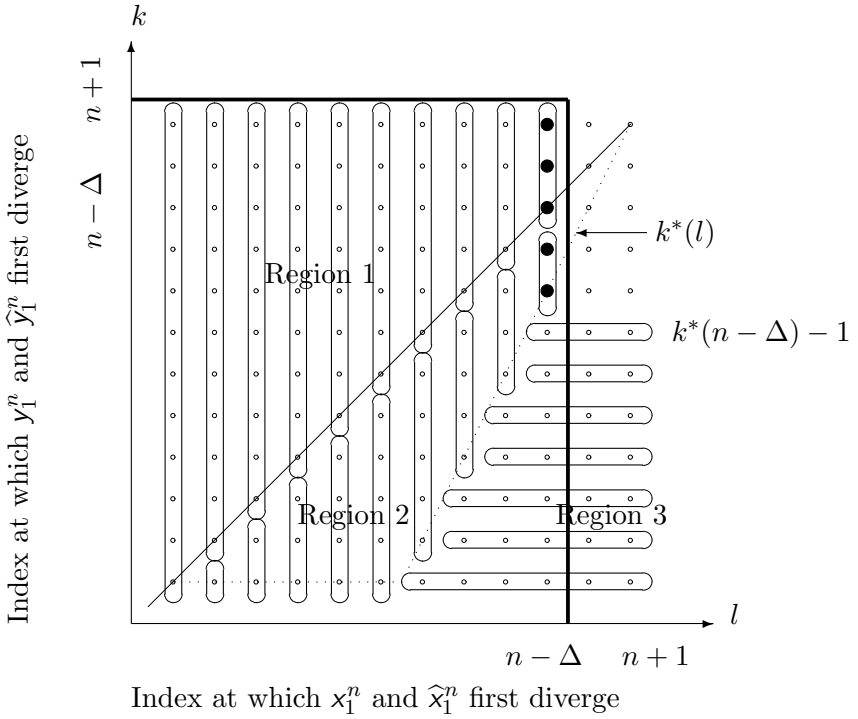

Figure 4.9. Two dimensional plot of the error probabilities $p_n(l,k)$, corresponding to error events (l,k), contributing to $\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}]$ in the situation where $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma) < \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)$.

the proof. The possible dominant error events are highlighted shown in Fig 4.9. Thus,

$$\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}] \leq \sum_{l=1}^{n-\Delta}\sum_{k=l}^{n+1} p_n(l,k) + \sum_{l=1}^{n-\Delta}\sum_{k=k^*(l)}^{l-1} p_n(l,k) + \sum_{l=1}^{n-\Delta}\sum_{k=1}^{k^*(l)-1} p_n(l,k) \quad (4.30)$$

Where $\sum_{k=1}^{0} p_k = 0$. The lower boundary of Region 2 is $k^*(l) \geq 1$ as a function of n and l:

$$\begin{aligned} k^*(l) &= \max\left\{1, n+1-\lceil\frac{\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)}{\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)}\rceil(n+1-l)\right\} \\ &= \max\left\{1, n+1-G(n+1-l)\right\} \end{aligned} \quad (4.31)$$

where we use G to denote the ceiling of the ratio of exponents. Note that when $\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma) > \inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)$ then $G = 1$ and region two of Fig. 4.9 disappears. In other words, the middle term of (4.30) equals zero. This is the first case considered. We now consider the cases when $G \geq 2$ (because of the ceiling function G is a positive integer).

The first term of (4.30), i.e., region one in Fig. 4.9 where $l \leq k$, is bounded in the same

way that the first term of (4.25) is, giving

$$\sum_{l=1}^{n-\Delta}\sum_{k=l}^{n+1} p_n(l,k) \le C_2 2^{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)-\epsilon]}. \tag{4.32}$$

In Fig. 4.9, region two is upper bounded by the 45-degree line, and lower bounded by $k^*(l)$. The second term of (4.30), corresponding to this region where $l \ge k$,

$$\begin{aligned}
\sum_{l=1}^{n-\Delta}\sum_{k=k^*(l)}^{l-1} p_n(l,k) &\le 4 \times \sum_{l=1}^{n-\Delta}\sum_{k=k^*(l)}^{l-1} 2^{-(n-k+1)E_y(R_x,R_y,\frac{l-k}{n-k+1})} \\
&= 4 \times \sum_{l=1}^{n-\Delta}\sum_{k=k^*(l)}^{l-1} 2^{-(n-k+1)\frac{n-l+1}{n-l+1}E_y(R_x,R_y,\frac{l-k}{n-k+1})} && (4.33)\\
&\le 4 \times \sum_{l=1}^{n-\Delta}\sum_{k=k^*(l)}^{l-1} 2^{-(n-l+1)\inf_{\gamma\in[0,1]}\frac{1}{1-\gamma}E_y(R_x,R_y,\gamma)} && (4.34)\\
&= 4 \times \sum_{l=1}^{n-\Delta}(l-k^*(l))2^{-(n-l+1)\inf_{\gamma\in[0,1]}\frac{1}{1-\gamma}E_y(R_x,R_y,\gamma)} && (4.35)
\end{aligned}$$

In (4.33) we note that $l \ge k$, so define $\frac{l-k}{n-k+1} = \gamma$ as in (4.34). Then $\frac{n-k+1}{n-l+1} = \frac{1}{1-\gamma}$.

The third term of (4.30), i.e., the intersection of region three and the "box" in Fig. 4.9 where $l \ge k$, can be bounded as,

$$\begin{aligned}
\sum_{l=1}^{n-\Delta}\sum_{k=1}^{k^*(l)-1} p_n(l,k) &\le \sum_{l=1}^{n+1}\sum_{k=1}^{\min\{l,k^*(n-\Delta)-1\}} p_n(l,k) && (4.36)\\
&= \sum_{k=1}^{k^*(n-\Delta)-1}\sum_{l=k}^{n+1} p_n(l,k) && (4.37)\\
&\le 4 \times \sum_{k=1}^{k^*(n-\Delta)-1}\sum_{l=k}^{n+1} 2^{-(n-k+1)E_y(R_x,R_y,\frac{l-k}{n-k+1})} \\
&\le 4 \times \sum_{k=1}^{k^*(n-\Delta)-1}\sum_{l=k}^{n+1} 2^{-(n-k+1)\inf_{\gamma\in[0,1]}E_y(R_x,R_y,\gamma)} \\
&\le 4 \times \sum_{k=1}^{k^*(n-\Delta)-1}(n-k+2)2^{-(n-k+1)\inf_{\gamma\in[0,1]}E_y(R_x,R_y,\gamma)} && (4.38)
\end{aligned}$$

In (4.36) we note that $l \le n-\Delta$ thus $k^*(n-\Delta)-1 \ge k^*(l)-1$, also $l \ge 1$, so $l \ge k^*(l)-1$. This can be visualized in Fig 4.9 as we extend the summation from the intersection of the "box" and region 3 to the whole region under the diagonal line and the horizontal line $k = k^*(n-\Delta)-1$. In (4.37) we simply switch the order of the summation.

Finally when $G \geq 2$, we substitute (4.32), (4.35), and (4.38) into (4.30) to give

$$\begin{aligned}
\Pr[\widehat{x}_1^{n-\Delta}(n) \neq x_1^{n-\Delta}] &\leq C_2 2^{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)-\epsilon]} \\
&\quad + 4 \times \sum_{l=1}^{n-\Delta} (l - k^*(l)) 2^{-(n-l+1)\inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x,R_y,\gamma)} \qquad (4.39) \\
&\quad + 4 \times \sum_{k=1}^{k^*(n-\Delta)-1} (n-k+2) 2^{-(n-k+1)\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)} \\
&\leq C_2 2^{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)-\epsilon]} \\
&\quad + 4 \times \sum_{l=1}^{n-\Delta} (l - n - 1 + G(n+1-l)) 2^{-(n-l+1)\inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x,R_y,\gamma)} \\
&\quad + 4 \times \sum_{k=1}^{n+1-G(\Delta+1)} (n-k+2) 2^{-(n-k+1)\inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)} \qquad (4.40) \\
&\leq C_2 2^{-\Delta[\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma)-\epsilon]} \\
&\quad + (G-1) C_3 2^{-\Delta\left[\inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x,R_y,\gamma)-\epsilon\right]} \\
&\quad + C_4 2^{-\left[\Delta G \inf_{\gamma\in[0,1]} E_y(R_x,R_y,\gamma)-\epsilon\right]} \\
&\leq C_5 2^{-\Delta\left[\min\left\{\inf_{\gamma\in[0,1]} E_x(R_x,R_y,\gamma), \inf_{\gamma\in[0,1]} \frac{1}{1-\gamma} E_y(R_x,R_y,\gamma)\right\}-\epsilon\right]}. \qquad (4.41)
\end{aligned}$$

To get (4.40), we use the fact that $k^*(l) \geq n + 1 - G(n + 1 - l)$ from the definition of $k^*(l)$ in (4.31) to upper bound the second term. We exploit the definition of G to convert the exponent in the third term to $\inf_{\gamma\in[0,1]} E_x(R_x, R_y, \gamma)$. Finally, to get (4.41) we gather the constants together, sum out over the decaying exponentials, and are limited by the smaller of the two exponents.

Note: in the proof of Theorem 3, we regularly double count the error events or add smaller extra probabilities to make the summations simpler. But it should be clear that the error exponent is not affected. ■

4.3.2 Universal Decoding

Universal decoding rule

As discussed in Section 2.3.3, we do not use a pairwise minimum joint-entropy decoder because of polynomial term in n would multiply the exponential decay in Δ. Analogous to the sequential decoder used there, we use a "weighted suffix entropy" decoder. The decoding starts by first identifying candidate sequence pairs as those that agree with the

encoding bit streams up to time n, i.e., $\bar{x}_1^n \in \mathcal{B}_x(x_1^n), \bar{y}_1^n \in \mathcal{B}_y(y_1^n)$. For any one of the $|\mathcal{B}_x(x_1^n)||\mathcal{B}_y(y_1^n)|$ sequence pairs in the candidate set, i.e., $(\bar{x}_1^n, \bar{y}_1^n) \in \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n)$ we compute $(n+1) \times (n+1)$ weighted entropies:

$$
\begin{aligned}
H_S(l,k,\bar{x}_1^n,\bar{y}_1^n) &= H(\bar{x}_l^{(n+1-l)}, \bar{y}_l^{(n+1-l)}), \quad l = k \\
H_S(l,k,\bar{x}_1^n,\bar{y}_1^n) &= \frac{k-l}{n+1-l} H(\bar{x}_l^{k-1}|\bar{y}_l^{k-1}) + \frac{n+1-k}{n+1-l} H(\bar{x}_k^n, \bar{y}_k^n), \quad l < k \\
H_S(l,k,\bar{x}_1^n,\bar{y}_1^n) &= \frac{l-k}{n+1-k} H(\bar{y}_k^{l-1}|\bar{x}_k^{l-1}) + \frac{n+1-l}{n+1-k} H(\bar{x}_l^n, \bar{y}_l^n), \quad l > k.
\end{aligned}
$$

We define the *score* of $(\bar{x}_1^n, \bar{y}_1^n)$ as the pair of integers $i_x(\bar{x}_1^n, \bar{y}_1^n)$, $i_y(\bar{x}_1^n, \bar{y}_1^n)$ s.t.,

$$
\begin{aligned}
i_x(\bar{x}_1^n, \bar{y}_1^n) &= \max\{i : H_S(l,k,(\bar{x}_1^n,\bar{y}_1^n)) < H_S(l,k,\tilde{x}_1^n,\tilde{y}_1^n) \forall k = 1,2,...n+1, \forall l = 1,2,...i, \\
&\quad \forall (\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n) \cap \mathcal{F}_n(l,k,\bar{x}_1^n,\bar{y}_1^n)\} \qquad (4.42) \\
i_y(\bar{x}_1^n, \bar{y}_1^n) &= \max\{i : H_S(l,k,(\bar{x}_1^n,\bar{y}_1^n)) < H_S(l,k,\tilde{x}_1^n,\tilde{y}_1^n) \forall l = 1,2,...n+1, \forall k = 1,2,...i, \\
&\quad \forall (\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n) \cap \mathcal{F}_n(l,k,\bar{x}_1^n,\bar{y}_1^n)\} \qquad (4.43)
\end{aligned}
$$

While $\mathcal{F}_n(l,k,x_1^n,y_1^n)$ is the same set as defined in (4.23), we repeat the definition here for convenience,

$$
\mathcal{F}_n(l,k,x_1^n,y_1^n) = \{(\bar{x}_1^n, \tilde{y}_1^n) \in \mathcal{X}^n \times \mathcal{Y}^n \text{ s.t. } \bar{x}_1^{l-1} = x_1^{l-1}, \bar{x}_l \neq x_l, \bar{y}_1^{k-1} = y_1^{k-1}, \bar{y}_k \neq y_k\}.
$$

The definition of $(i_x(\bar{x}_1^n, \bar{y}_1^n), i_y(\bar{x}_1^n, \bar{y}_1^n))$ can be visualized in the following procedure. As shown in Fig. 4.10, for all $1 \leq l, k \leq n+1$, if there exists $(\bar{\bar{x}}_1^n, \bar{\bar{y}}_1^n) \in \mathcal{F}_n(l,k,(\bar{x}_1^n,\bar{y}_1^n)) \cap \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n)$ s.t. $H_S(l,k,\bar{x}_1^n,\bar{y}_1^n) \geq H_S(l,k,\bar{\bar{x}}_1^n,\bar{\bar{y}}_1^n)$, then we mark (l,k) on the plane as shown in Fig.4.10. Eventually we pick the maximum integer which is smaller than all marked x-coordinates as $i_x(\bar{x}_1^n, \bar{y}_1^n)$ and the maximum integer which is smaller than all marked y-coordinates as $i_y(\bar{x}_1^n, \bar{y}_1^n)$. The score of $(\bar{x}_1^n, \bar{y}_1^n)$ tells us the first branch(either x or y) point where a "better sequence pair" (with a smaller weighted entropy) exists.

Define the set of the winners as the sequences (not sequence pair) with the maximum score:

$$
\mathcal{W}_n^x = \{\bar{x}_1^n \in \mathcal{B}_x(x_1^n) : \exists \bar{y}_1^n \in \mathcal{B}_y(y_1^n), s.t. i_x(\bar{x}_1^n, \bar{y}_1^n) \geq i_x(\tilde{x}_1^n, \tilde{y}_1^n), \forall (\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n)\}
$$

$$
\mathcal{W}_n^y = \{\bar{y}_1^n \in \mathcal{B}_y(y_1^n) : \exists \bar{x}_1^n \in \mathcal{B}_x(x_1^n), s.t. i_y(\bar{x}_1^n, \bar{y}_1^n) \geq i_y(\tilde{x}_1^n, \tilde{y}_1^n), \forall (\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n)\}
$$

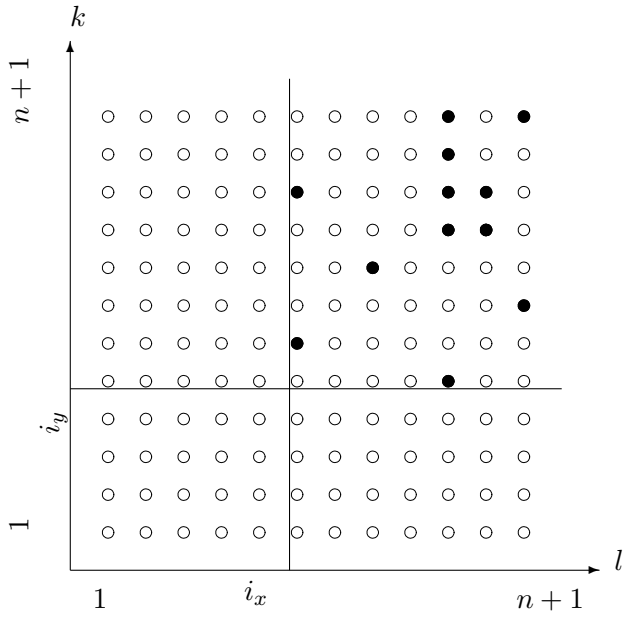

Figure 4.10. 2D interpretation of the *score*, $(i_x(\bar{x}_1^n, \bar{y}_1^n), i_y(\bar{x}_1^n, \bar{y}_1^n))$, of a sequence pair $(\bar{x}_1^n, \bar{y}_1^n)$. If there exists a sequence pair in $\mathcal{F}_n(l, k, \bar{x}_1^n, \bar{y}_1^n)$ with less or the same score, then (l, k) is marked with a solid dot. The *score* $i_x(\bar{x}_1^n, \bar{y}_1^n)$ is the largest integer which is smaller than all the x-coordinates of the marked points. Similarly for $i_y(\bar{x}_1^n, \bar{y}_1^n)$,

Then arbitrarily pick one sequence from $\mathcal{W}_n^x$ and one from $\mathcal{W}_n^y$ as the decision $(\widehat{x}_1^n(n), \widehat{y}_1^n(n))$ at decision time n.

Details of the proof of Theorem 4

Using the above universal decoding rule, we give an upper bound on the decoding error $\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}]$, and hence derive an achievable error exponent.

We bound the probability that there exists a sequence pair in $\mathcal{F}_n(l, k, (\mathsf{x}_1^n, \mathsf{y}_1^n)) \cap \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n)$ with smaller weighted minimum-entropy suffix score as:

$$\begin{aligned} p_n(l,k) &= \sum_{x_1^n} \sum_{y_1^n} p_{\mathsf{xy}}(x_1^n, y_1^n) \Pr[\exists (\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n) \cap \mathcal{F}_n(l, k, x_1^n, y_1^n), \\ &\quad s.t. H_S(l, k, \tilde{x}_1^n, \tilde{y}_1^n) \leq H_S(l, k, (x_1^n, y_1^n))] \end{aligned}$$

Note that the $p_n(l, k)$ here differs from the $p_n(l, k)$ defined in the ML decoding by replacing $p_{\mathsf{xy}}(x_1^n, y_1^n) \leq p_{\mathsf{xy}}(\tilde{x}_1^n, \tilde{y}_1^n)$ with $H_S(l, k, \tilde{x}_1^n, \tilde{y}_1^n) \leq H_S(l, k, (x_1^n, y_1^n))$.

The following lemma, analogous to (4.22) for ML decoding, tells us that the "suffix weighted entropy" decoding rule is a good one.

Lemma 10 *Upper bound on symbol-wise decoding error* $\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}]$ *:*

$$\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}] \leq \Pr[\widehat{\mathsf{x}}_1^{n-\Delta}(n) \neq \mathsf{x}_1^{n-\Delta}] \leq \sum_{l=1}^{n-\Delta} \sum_{k=1}^{n+1} p_n(l,k)$$

Proof: The first inequality is trivial. We only need to show the second inequality. According to the decoding rule, $\widehat{x}_1^{n-\Delta} \neq x_1^{n-\Delta}$ implies that there exists a sequence $\tilde{x}_1^n \in \mathcal{W}_n^x$ s.t.$\tilde{x}_1^{n-\Delta} \neq x_1^{n-\Delta}$. This means that there exists a sequence $\tilde{y}_1^n \in \mathcal{B}_y(y_1^n)$, s.t. $i_x(\tilde{x}_1^n, \tilde{y}_1^n) \geq i_x(x_1^n, y_1^n)$. Suppose that $(\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{F}_n(l,k,x_1^n,y_1^n)$, then $l \leq n-\Delta$ because $\tilde{x}_1^{n-\Delta} \neq x_1^{n-\Delta}$. By the definition of i_x, we know that $H_S(l,k,\tilde{x}_1^n,\tilde{y}_1^n) \leq H_S(l,k,x_1^n,y_1^n)$. And using the union bound argument we get the desired inequality. □

We only need to bound each single error probability $p_n(l,k)$ to finish the proof.

Lemma 11 *Upper bound on* $p_n(l,k)$*,* $l \leq k$*:* $\forall \epsilon > 0$*,* $\exists K_1 < \infty$*, s.t.*

$$p_n(l,k) \leq 2^{-(n-l+1)[E_x(R_x,R_y,\lambda)-\epsilon]}$$

where $\lambda = (k-l)/(n-l+1) \in [0,1]$*.*

Proof: The proof bears some similarity to the that of the universal decoding problem of the point-to-point source coding problem in Proposition 4. The details of the proof is in Appendix F.2 □

A similar derivation yields a bound on $p_n(l,k)$ for $l \geq k$.

Combining Lemmas 11 and 10, and then following the same derivation for ML decoding yields Theorem 4. ■

4.4 Discussions

We derived the achievable delay constrained error exponents for distributed source coding. The key technique is "divide and conquer". We divide the delay constrained error event

into individual error events. For each individual error events, we apply the classical block coding analysis to get an individual error exponent for that specific error event. Then by a union bound argument we determine the *dominant* error event. This "divide and conquer" scheme not only works for the Slepian-Wolf source coding in this chapter but also works for other information-theoretic problems illustrated in [13, 14]. While the encoder is the same sequential random binning shown in Section 2.3.3, the decoding is quite complicated due to the nature of the problem. Similar to that in Section 2.3.3, we derived the error exponent for both ML and universal decoding. To show the equivalence of the two different error exponents, we apply Lagrange duality and tilted distributions in Appendix G. There are several other important results in Appendix G. The derivations are extremely laborious but the reward is quite fulfilling. These results in Appendix G essentially follow the I-projection theory in the statistics literature [30] and were recently discussed in the context of channel coding error exponents [9].

We only showed that some positive delay constrained error exponent is achievable as long as the rate pair is in the interior of the classical Slepian-Wolf region in Figure A.3. In general, the delay constrained error exponent for distributed source coding is smaller or equal to its block coding counterpart. This is different from what we observed in Chapters 2 and 3. To further understand the difference, we need a *tight* upper bound on the error exponent. A trivial example tells us that this error exponent is nowhere tight. Consider the special case where the two sources are independent, the random coding scheme in this chapter gives the standard random coding error exponent for each single source as shown in (2.13) in Section 2.3.3. This error exponent is strictly smaller than the optimal coding scheme in Chapter 2 as shown in Proposition 8 in Section 2.5. The optimal error exponent problem could be an extremely difficult one. It would be interesting to give *an* upper bound on it. However, we do not have any general non-trivial upper bounds either. In Chapter 5, we will study the upper bound on the delay constrained error exponents for source coding with decoder side-information problem. This is a special case of what we study in this chapter since the decoder side-information can be treated as an encoder with rate higher than the logarithm of the alphabet. However, this upper bound is not tight in general. These open questions are left future research.

Chapter 5

Lossless Source Coding with Decoder Side-Information

In this chapter, we study the delay constrained source coding with decoder side-information problem. As a special case of the Slepian-Wolf problem studied in Chapter 4, the sequential random binning scheme's delay performance is derived as a corollary of those results in Chapters 2 and 4. An upper bound on the delay constrained error exponent is derived by using the feed-forward decoding scheme from the channel coding literature. The results in this chapter are also summarized in our paper [18], especially the implications of these results on compression of encrypted data [50].

5.1 Delay constrained source coding with decoder side-information

In [79], Slepian and Wolf studied the distributed lossless source coding problems. One of the problems studied, the source coding with only decoder side-information problem is shown in Figure 4.1, where the encoder has access to the source x only, but not the side-information y. It is clear that if the side-information y is also available to the encoder, to achieve arbitrarily small decoding error for fixed block coding, rate at the conditional entropy $H(p_{x|y})$ is necessary and sufficient. Somewhat surprisingly, even without the *encoder*

side-information, Slepian and Wolf showed that rate at conditional entropy $H(p_{x|y})$ is still sufficient.

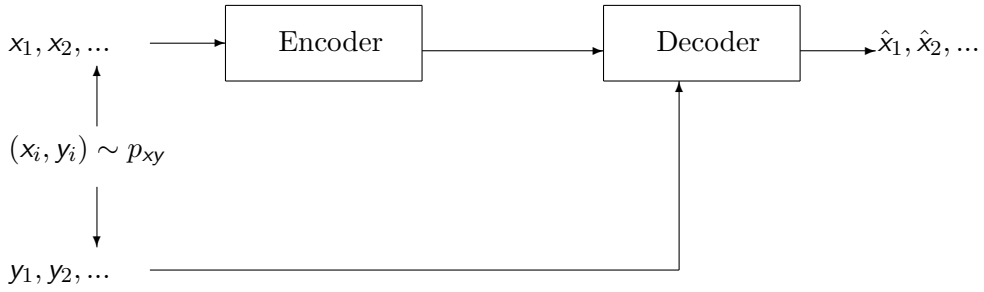

Figure 5.1. Lossless source coding with decoder side-information

We can also factor the joint probability to treat the source as a random variable x and consider the side-information y as the output of a discrete memoryless channel (DMC) $p_{y|x}$ with x as input. This model is shown in Figure 5.2.

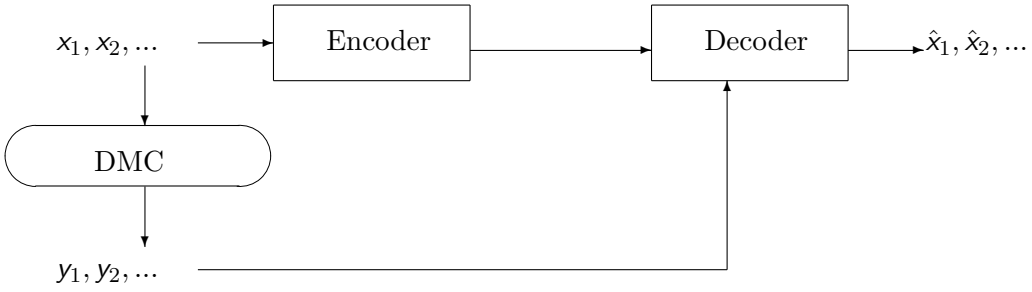

Figure 5.2. Lossless source coding with side-information: DMC

We first review the error exponent result for source coding with decoder side-information in the fixed-length block coding setup in Section A.4 in the appendix. Then we formally define the delay constrained source coding with side-information problem. In Section 5.1.2 we give a lower a bound and an upper bound on the error exponent on this problem. These two bounds in general do not match, which leaves space for future explorations.

5.1.1 Delay Constrained Source Coding with Side-Information

Rather than being known in advance, the source symbols stream into the encoder in a real-time fashion. We assume that the source generates a pair of source symbols (x, y) per second from the finite alphabet $\mathcal{X} \times \mathcal{Y}$. The j^{th} source symbol x_j is not known at the encoder until time j and similarly for y_j at the decoder. Rate R operation means that the encoder sends 1 binary bit to the decoder every $\frac{1}{R}$ seconds. For obvious reasons (cf. Proposition 1 and 2), we focus on cases with $H_{x|y} < R < \log_2 |\mathcal{X}|$.

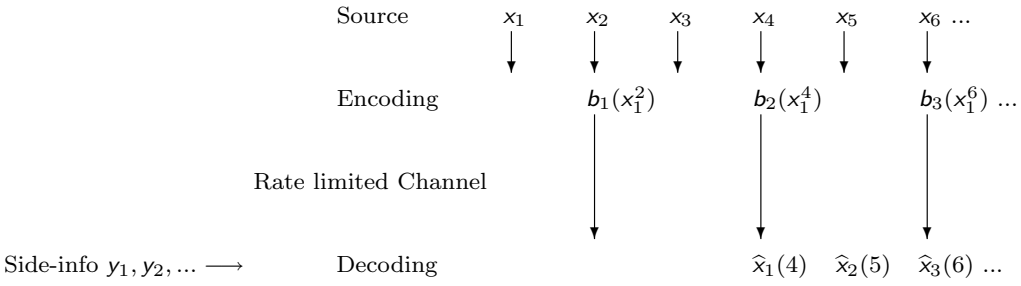

Figure 5.3. Delay constrained source coding with side-information: rate $R = \frac{1}{2}$, delay $\Delta = 3$

Definition 8 *A sequential encoder-decoder pair $\mathcal{E}, \mathcal{D}$ are sequence of maps. $\{\mathcal{E}_j\}, j = 1, 2, ...$ and $\{\mathcal{D}_j\}, j = 1, 2, ...$ The outputs of $\mathcal{E}_j$ are the outputs of the encoder $\mathcal{E}$ from time $j-1$ to j.*

$$\mathcal{E}_j : \mathcal{X}^j \longrightarrow \{0,1\}^{\lfloor jR \rfloor - \lfloor (j-1)R \rfloor}$$

$$\mathcal{E}_j(x_1^j) = b_{\lfloor (j-1)R \rfloor + 1}^{\lfloor jR \rfloor}$$

The outputs of $\mathcal{D}_j$ are the decoding decisions of all the arrived source symbols by time j based on the received binary bits up to time j as well as the side-information.

$$\mathcal{D}_j : \{0,1\}^{\lfloor jR \rfloor} \times \mathcal{Y}^j \longrightarrow \mathcal{X}$$

$$\mathcal{D}_j(b_1^{\lfloor jR \rfloor}, y_1^j) = \widehat{x}_{j-\Delta}(j)$$

Where $\widehat{x}_{j-\Delta}(j)$ is the estimation of $x_{j-\Delta}$ at time j and thus has end-to-end delay of Δ seconds. A rate $R = \frac{1}{2}$ sequential source coding system is illustrated in Figure 5.3.

For sequential source coding, it is important to study the symbol by symbol decoding error probability instead of the block coding error probability.

Definition 9 *A family of rate R sequential source codes $\{(\mathcal{E}, \mathcal{D}^{\Delta})\}$ are said to achieve delay-reliability $E_{si}(R)$ if and only if for all $\epsilon > 0$, there exists $K < \infty$, s.t. $\forall i$, $\Delta > 0$*

$$\Pr[x_i \neq \widehat{x}_i(i+\Delta)] \leq K 2^{-\Delta(E_{si}(R)-\epsilon)}$$

Following this definition, we have both lower and upper bound on delay constrained error exponent for source coding with side-information.

5.1.2 Main results of Chapter 5: lower and upper bound on the error exponents

There are two parts of the main result. First, we give an achievable lower bound on the delay constrained error exponent $E_{si}(R)$. This part is realized by using the same sequential random binning scheme in Definition 3 and a Maximum-likelihood decoder in Section 4.3.1 or a universal decoder in Section 4.3.2. The lower bound can be treated as a simple corollary of the more general theorems in the distributed source coding setup in Theorems 3 and 4. The second part is an upper bound on $E_{si}(R)$. We apply a modified version of the feed-forward decoder used by Pinsker [62] and recently clarified by Sahai [67]. This feed-forward decoder technique is a powerful tool in the upper bound analysis for delay constrained error exponents. Using this technique, we derived an upper bound on the joint source-channel coding error exponent in [15].

A lower bound on $E_{si}(R)$

We state the relevant lower bound (achievability) results which comes as a simpler result to the more general result proved in Theorems 3 and 4.

Theorem 6 *Delay constrained random source coding theorem: Using a random sequential coding scheme and the ML decoding rule using side-information, for all i, Δ:*

$$\Pr[\widehat{x}_i(i+\Delta) \neq x_i] \leq K 2^{-\Delta E_{si}^{lower}(R)} \tag{5.1}$$

Where K is a constant, and

$$E_{si}^{lower}(R) = E_{si,b}^{lower}(R) = \min_{q_{xy}}\{D(q_{xy}\|p_{xy}) + |0, R - H(q_{x|y})|^+\}$$

where $E_{si,b}^{lower}(R)$ is defined in Theorem 13. Another definition is

$$E_{si}^{lower}(R) = E_{si,b}^{lower}(R) = \max_{\rho\in[0,1]} \rho R - \log[\sum_y [\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}]^{1+\rho}]$$

These two expressions can be shown to be equivalent following the same Lagrange multiplier argument in [39] and Appendix G.

Similarly for the universal decoding rule, for all $\epsilon > 0$, there exist a finite constant K, s.t. for all n and Δ:

$$\Pr[\widehat{x}_i(i+\Delta) \neq x_i] \leq K 2^{-\Delta(E_{si}^{lower}(R)-\epsilon)} \tag{5.2}$$

The encoder is the same random sequential encoding scheme described in Definition 3 which can be realized using an infinite constraint-length time-varying random convolutional code. Common randomness between the encoder and the decoder is assumed. The decoder can use a maximum likelihood rule or minimum empirical joint entropy decoding rule which are similar to that in the point-to-point source coding setup in Chapter 2. Details of the proof are in Section 5.5.

An upper bound on $E_{si}(R)$

We give an upper bound on the delay constrained error exponent for source coding with side-information defined in Definition 9. This bound is for any generic joint distribution p_{xy}. Some special cases will be discussed in Section 5.2.

Theorem 7 *For the source coding with side-information problem in Figure 5.2, if the source is iid $\sim p_{xy}$ from a finite alphabet, then the error exponents $E_{si}(R)$ with fixed delay must satisfy $E_{si}(R) \leq E_{si}^{upper}(R)$, where*

$$\begin{aligned} E_{si}^{upper}(R) = \quad \min \quad \{ & \inf_{q_{xy},\alpha\geq 1: H(q_{x|y})>(1+\alpha)R} \{\frac{1}{\alpha} D(q_{xy}\|p_{xy})\}, \\ & \inf_{q_{xy},1\geq\alpha\geq 0: H(q_{x|y})>(1+\alpha)R} \{\frac{1-\alpha}{\alpha} D(q_x\|p_x) + D(q_{xy}\|p_{xy})\}\} \end{aligned}$$

5.2 Two special cases

It is clear that the upper bound and the lower bound on the error exponent $E_{si}(R)$are not the same thing. There seems to be a gap between the two bounds. But how big or how small can the gap be? In this section, we answer that question by showing two *extreme* cases.

5.2.1 Independent side-information

Consider the case that x and y are independent, or there is no side-information. Then clearly this source coding with the irrelevant "side-information" at the decoder problem is the same as the single source coding problem discussed in Chapter 2. So the upper bound $E_{si}^{upper}(R) \geq E_s(R)$, where $E_s(R)$ is the delay constrained source coding error exponent defined in Theorem 1 for source x. So we only need to show that $E_{si}^{upper}(R) \leq E_s(R)$ to establish the tightness of our upper bound for this extreme case where no side-information is presented. The proof simply follows the following argument:

$$\begin{aligned}
E_{si}^{upper}(R) &=_{(a)} \min\{\inf_{q_{\mathsf{xy}},\alpha\geq 1: H(q_{\mathsf{x}|\mathsf{y}})>(1+\alpha)R}\{\frac{1}{\alpha}D(q_{\mathsf{xy}}\|p_{\mathsf{xy}})\}, \\
&\qquad \inf_{q_{\mathsf{xy}},1\geq\alpha\geq 0: H(q_{\mathsf{x}|\mathsf{y}})>(1+\alpha)R}\{\frac{1-\alpha}{\alpha}D(q_{\mathsf{x}}\|p_{\mathsf{x}})+D(q_{\mathsf{xy}}\|p_{\mathsf{xy}})\}\} \\
&\leq_{(b)} \inf_{q_{\mathsf{xy}},\alpha\geq 0: H(q_{\mathsf{x}|\mathsf{y}})>(1+\alpha)R}\{\frac{1}{\alpha}D(q_{\mathsf{xy}}\|p_{\mathsf{xy}})\} \\
&\leq_{(c)} \inf_{q_{\mathsf{x}}\times p_{\mathsf{y}},\alpha\geq 1: H(q_{\mathsf{x}|\mathsf{y}})>(1+\alpha)R}\{\frac{1}{\alpha}D(q_{\mathsf{x}}\times p_{\mathsf{y}}\|p_{\mathsf{x}}\times p_{\mathsf{y}})\} \\
&=_{(d)} \inf_{q_{\mathsf{x}},\alpha\geq 1: H(q_{\mathsf{x}})>(1+\alpha)R}\{\frac{1}{\alpha}D(q_{\mathsf{x}}\|p_{\mathsf{x}})\} \\
&=_{(e)} E_s(R)
\end{aligned} \tag{5.3}$$

(a) is by definition. (b) is because $D(q_{\mathsf{xy}}\|p_{\mathsf{xy}}) \geq D(q_{\mathsf{x}}\|p_{\mathsf{x}})$. (c) is true because of the following two observations. First, x and y are independent, so $p_{\mathsf{xy}} = p_{\mathsf{x}} \times p_{\mathsf{y}}$. Second, we take the inf over a subset of all q_{xy}: $q_{\mathsf{x}} \times p_{\mathsf{y}}$, i.e. the distributions such that x and y are independent and the marginal $q_{\mathsf{y}} = p_{\mathsf{y}}$. (d) is true because under p and q, the two marginals x and y are independent with $\mathsf{y} \sim p_{\mathsf{y}}$. (e) is by definition.

The lower bound on the error exponent $E_{si}^{lower}(R)$ has the following form for independent

side-information.

$$
\begin{aligned}
E_{si}^{lower}(R) &=_{(a)} \max_{\rho\in[0,1]} \rho R - \log[\sum_y [\sum_x p_{\mathsf{x}\mathsf{y}}(x,y)^{\frac{1}{1+\rho}}]^{1+\rho}] \\
&=_{(b)} \max_{\rho\in[0,1]} \rho R - \log[\sum_y [\sum_x p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}} p_{\mathsf{y}}(y)^{\frac{1}{1+\rho}}]^{1+\rho}] \\
&=_{(c)} \max_{\rho\in[0,1]} \rho R - \log[\sum_y p_{\mathsf{y}}(y) [\sum_x p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}}]^{1+\rho}] \\
&=_{(d)} \max_{\rho\in[0,1]} \rho R - (1+\rho)\log[\sum_x p_{\mathsf{x}}(x)^{\frac{1}{1+\rho}}] \\
&=_{(e)} E_r(R)
\end{aligned} \tag{5.4}
$$

(a) is by definition. (b) is because the side-information y is independent of the source x. (c) and (d) are trivial. (e) is by definition in (A.5).

From the above analysis, we know that our upper bound $E_{si}^{upper}(R)$ is tight which can be achieved as shown in Chapter 2 and our lower bound is the same as the random coding error exponent for source x only in Section 2.3.3. Which validates our results in Section 5.1.2. Comparing the upper bound in (5.3) and lower bound in (5.4), we know that the lower bound is strictly lower than the upper bound as shown in Proposition 8 in Chapter 2 and illustrated in Figure 2.3.

5.2.2 Delay constrained encryption of compressed data

The traditional view of encryption of redundant source is in the block coding context in which all the source symbols are compressed first then encrypted[1]. In the receiver end, decompression of the source is after the decryption. As the common theme in this thesis, we consider the end-to-emd delay between the realization of the source and the decoding/decryption at the sink. The delay constrained compression first then encryption system is illustrated in Figure 5.5. Without getting into the details of the information-theoretic security, we briefly introduce the system. The compression and decompression part is exactly the delay constrained source coding shown in Figure 2.1 in Chapter 2 where the source is iid $\sim p_{\mathsf{s}}$ on a finite alphabet $\mathcal{S}$. The secret keys $y_1, ...$ are iid Bernoulli 0.5 random variables, so the output of the encrypter $\tilde{b} \oplus y$ is independent with the output of the compressor $\tilde{b}$, thus the name "perfect secrecy". Here the $\oplus$ operator is sum mod two, or more generally speaking, the addition operator in the finite field of size 2. Since the

[1]We consider the type I Shannon-sense security (perfect secrecy) [73, 46].

encryption and decryption are instantaneous, the overall system has a delay constrained error exponent $E_s(R)$ defined in Theorem 1, Chapter 2.

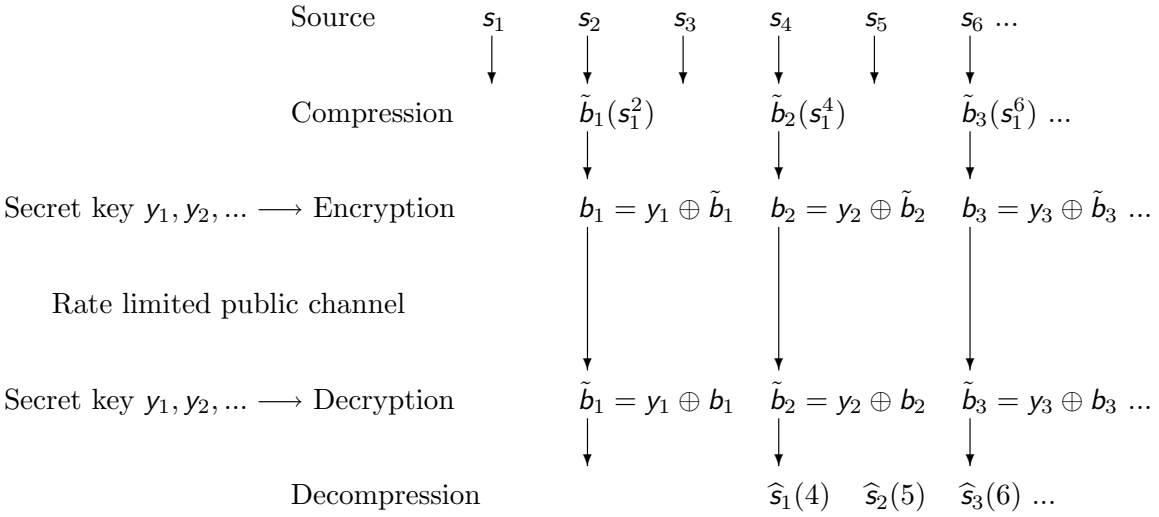

Figure 5.4. Encryption of compressed streaming data with delay constraints rate $R = \frac{1}{2}$, delay $\Delta = 3$

In the thought provoking paper [50], Johnson shows that the same compression rate (entropy rate of the source s) can be achieved by first encrypting the source then compressing the encrypted data without knowing the key, then decompress/decrypt the encoded data by using the key as the decoder side-information. Recently practical system is build by implementing LDPC [36, 65] codes for source coding with side-information [71]. The delay constrained system of the encryption first, then compression system is shown in Figure 5.5.

The source is iid $\sim p_s$ on a finite alphabet $\mathcal{S}$, to achieve perfect secrecy, the secret keys y's are iid uniform random variables on $\mathcal{S}$. And hence the encrypted data x is also uniformly distributed in $\mathcal{S}$ where

$$x = s \oplus y \tag{5.5}$$

where the $\oplus$ operator is the addition in the finite field of size $|\mathcal{S}|$. For uniform sources, no compression can be done. However, since y is known to the decoder and y is correlated to x, the source x can be compressed to the conditional entropy $H(x|y)$ which is equal to $H(s)$ because of the relations of x, y and s through (5.5). The emphasis of this section is the fundamental information-theoretical model of the problem in Figure 5.5, as shown in

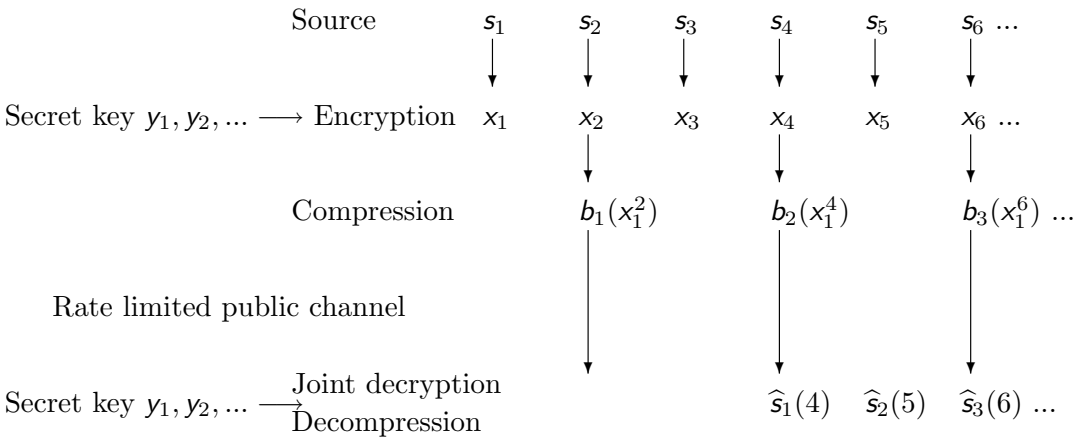

Figure 5.5. Compression of encrypted streaming data with delay constraints: rate $R = \frac{1}{2}$, delay $\Delta = 3$

Figure 5.6. A detailed discussion of the delay constrained encryption of compressed data problem is in [18].

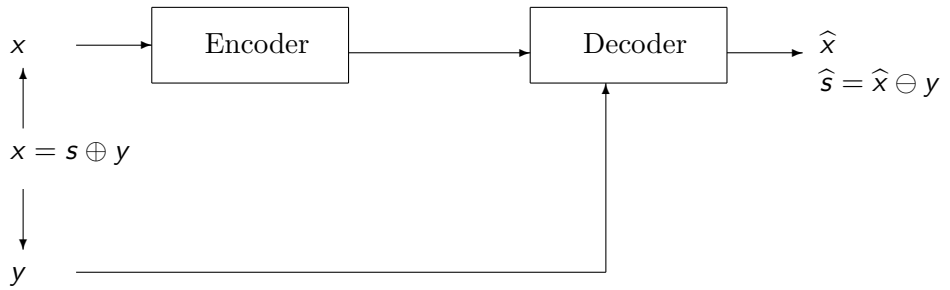

Figure 5.6. Information-theoretic model of compression of encrypted data, s is the source, y is uniformly distributed in $\mathcal{S}$ is the secret key. The encoder is facing a uniformly distributed x and the decoder has side-information y.

Notice that the estimate of x and s are related in the following way:

$$\widehat{s} = \widehat{x} \ominus y \tag{5.6}$$

so the estimation problems are equivalent for s and x. The lower bound and the upper bound on the delay constrained error exponent for x are shown in Theorem 6 and Theorem 7 respectively for general sources x and side information $y \sim p_{xy}$. For the compression of

encrypted data problem, as illustrated in Figure 5.6, we have the following corollary for both lower bound and upper bound.

Corollary 1 *Bounding the delay constrained error exponent for compression of encrypted data: for the compression of encrypted data system in Figure 5.5 and hence the information-theoretic interpretation as a decoder side-information problem of source x given side-information y in Figure 5.6, the delay constrained error exponent $E_{si}(R)$ is bounded by the following two error exponents:*

$$E_r(R, p_s) \leq E_{si}(R) \leq E_{s,b}(R, p_s) \tag{5.7}$$

where $E_r(R, p_s)$ is the random coding error exponent for source p_s defined in (A.5), $E_{s,b}(R, p_s)$ is the block coding error exponent defined in (A.4), both definitions are in Chapter 2.

It should be clear that this corollary is true for any source side-information pair shown in Figure 5.6, where $x = s \oplus y$ and y is uniform on $\mathcal{S}$. The proof is in Appendix I.

5.3 Delay constrained Source Coding with Encoder Side-Information

In this section, we study the source coding problem for x from a joint distribution $(x, y) \sim p_{xy}$. Suppose that the side information y is known at both encoder and decoder shown in Figure 5.7.

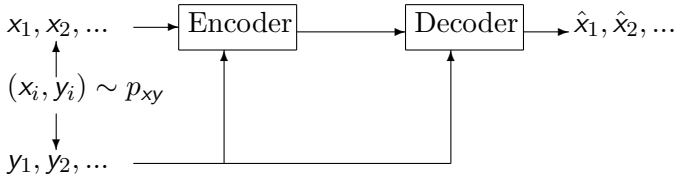

Figure 5.7. Lossless source coding with both encoder and decoder side-information

We first review the error exponent result for source coding with encoder side information problem. As shown in Figure 5.7, the sources are iid random variables x_1^n, y_1^n from a finite

alphabet $\mathcal{X}\times\mathcal{Y}$. Without loss of generality, $p_x(x)>0$, $\forall x\in\mathcal{X}$ and $p_y(y)>0$, $\forall y\in\mathcal{Y}$. x_1^n is the source known to the encoder and y_1^n is the side-information known to both the decoder and the encoder. A rate R block source coding system for n source symbols consists of an encoder-decoder pair $(\mathcal{E}_n,\mathcal{D}_n)$. Where

$$\begin{aligned}&\mathcal{E}_n:\mathcal{X}^n\times\mathcal{Y}^n\to\{0,1\}^{\lfloor nR\rfloor}, && \mathcal{E}_n(x_1^n,y_1^n)=b_1^{\lfloor nR\rfloor}\\ &\mathcal{D}_n:\{0,1\}^{\lfloor nR\rfloor}\times\mathcal{Y}^n\to\mathcal{X}^n, && \mathcal{D}_n(b_1^{\lfloor nR\rfloor},y_1^n)=\widehat{x}_1^n\end{aligned}$$

The error probability is $\Pr(x_1^n\neq\widehat{x}_1^n)=\Pr(x_1^n\neq\mathcal{D}_n(\mathcal{E}_n(x_1^n)))$. The exponent $E_{ei,b}(R)$ is achievable if $\exists$ a family of $\{(\mathcal{E}_n,\mathcal{D}_n)\}$, s.t.

$$\lim_{n\to\infty}-\frac{1}{n}\log_2\Pr(x_1^n\neq\widehat{x}_1^n)=E_{ei,b}(R) \tag{5.8}$$

As a simple corollary of the relevant results of [29, 39], we have the following lemma.

Lemma 12 *$E_{ei,b}(R)=E_{si,b}^{upper}(R)$ where $E_{si,b}^{upper}(R)$ is the upper bound on the source coding with decoder only side-information defined in Theorem 13.*

$$\begin{aligned}E_{si,b}^{upper}(R) &= \min_{q_{xy}:H(q_{x|y})\geq R}\{D(q_{xy}\|p_{xy})\}\\ &= \sup_{\rho\geq 0}\rho R-\log\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{(1+\rho)}\end{aligned}$$

Note: this error exponent is both achievable and tight in the usual block coding way [29].

5.3.1 Delay constrained error exponent

Similar to the previous cases, we have the following notion on delay constrained source coding with both encoder and decoder side-information.

Definition 10 *A sequential encoder-decoder pair $\mathcal{E},\mathcal{D}$ are sequence of maps. $\{\mathcal{E}_j\}, j=1,2,...$ and $\{\mathcal{D}_j\}, j=1,2,....$ The outputs of $\mathcal{E}_j$ are the outputs of the encoder $\mathcal{E}$ from time $j-1$ to j.*

$$\begin{aligned}&\mathcal{E}_j:\mathcal{X}^j\times\mathcal{Y}^j\longrightarrow\{0,1\}^{\lfloor jR\rfloor-\lfloor(j-1)R\rfloor}\\ &\mathcal{E}_j(x_1^j,y_1^j)=b_{\lfloor(j-1)R\rfloor+1}^{\lfloor jR\rfloor}\end{aligned}$$

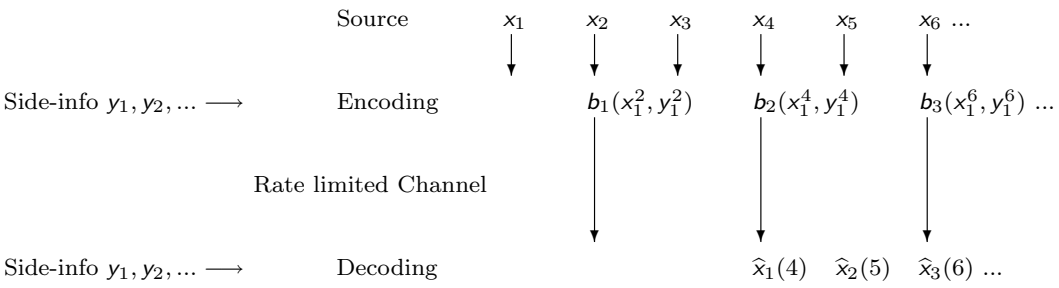

Figure 5.8. Delay constrained source coding with encoder side-information: rate $R = \frac{1}{2}$, delay $\Delta = 3$

The outputs of $\mathcal{D}_j$ are the decoding decisions of all the arrived source symbols by time j based on the received binary bits up to time j as well as the side-information.

$$\mathcal{D}_j : \{0,1\}^{\lfloor jR \rfloor} \times \mathcal{Y}^j \longrightarrow \mathcal{X}$$

$$\mathcal{D}_j(b_1^{\lfloor jR \rfloor}, y_1^j) = \widehat{x}_{j-\Delta}(j)$$

Where $\widehat{x}_{j-\Delta}(j)$ is the estimation of $x_{j-\Delta}$ at time j and thus has end-to-end delay of Δ seconds. A rate $R = \frac{1}{2}$ sequential source coding system is illustrated in Figure 5.8.

The delay constrained error exponent is defined in Definition 11. This is parallel to previous definitions on delay constrained error exponents.

Definition 11 *A family of rate R sequential source codes $\{(\mathcal{E}^\Delta, \mathcal{D}^\Delta)\}$ are said to achieve delay-reliability $E_{ei}(R)$ if and only if for all $\epsilon > 0$, there exists $K < \infty$, s.t. $\forall i$, $\Delta > 0$*

$$\Pr(x_i \neq \widehat{x}_i(i + \Delta)) \leq K 2^{-\Delta(E_{ei}(R) - \epsilon)}$$

Following the definition of delay constrained source coding error exponent $E_{ei}(R)$ in Definition 11, we have the following result in Theorem 8.

Theorem 8 *Delay constrained error exponent with both encoder and decoder side-information*

$$E_{ei}(R) = \inf_{\alpha > 0} \frac{1}{\alpha} E_{ei,b}((\alpha + 1)R)$$

Where $E_{ei,b}(R)$ is the block source coding error exponent defined in Lemma 12.

Similar to the lossless source coding in Chapter 2 and lossy source coding in Chapter 3, the delay constrained source coding error exponent and the block coding error exponent are connected by the focusing operator.

In Appendix J, we first show the achievability of $E_{ei}(R)$ by a simple variable length universal code and a FIFO queue coding scheme which is very similar to that for lossless source coding in Chapter 2. Then we show that $E_{ei}(R)$ is the upper bound on the delay constrained error exponent by the same argument used in the proof for lossless source coding error exponent in Chapter 2. Indeed, encoder side-information eliminates all the *future* randomness in the source and the side-information, so the coding system should have the same nature as the lossless source coding system discussed in Chapter 2. Although in different forms, it should be conceptually clear that $E_{ei}(R)$ and $E_s(R)$ share many characteristics. Without proof, we claim that the properties in Section 2.5 for $E_s(R)$ can be also found in $E_{ei}(R)$.

5.3.2 Price of ignorance

In the block coding setup, with or without encoder side-information does not change the error exponent in a dramatic way. As shown in [29], the difference is only between the random coding error exponent $E_{si,b}^{lower}(R)$ and the error exponent $E_{si,b}^{upper}(R)$. These two are the same in the low rate regime, this is similar to the well-known channel coding error exponent where random coding and sphere packing bounds are the same in the high rate regime [41]. Furthermore, the gap between with encoder side-information and without encoder side-information can be further reduced by using expurgation as shown in [2].

However for the delay constrained case, we show that without the encoder side-information, the error exponent with only decoder side-information is generally strictly smaller than the error exponent when the encoder side-information is *also* presented.

Corollary 2 *Price of ignorance:*

$$E_{ei}(R) > E_{si}^{upper}(R)$$

Proof: $D(q_{XY}\|p_{XY}) > D(q_X\|p_X)$ in general. ■

5.4 Numerical results

In this section we show three examples to illustrate the nature of the upper bound and the lower bound on the delay constrained error exponent with decoder side-information.

5.4.1 Special case 1: independent side-information

As shown in Section 5.2.1, if the side-information y is independent with the source x, the upper bound agrees with the delay constrained source coding error exponent $E_s(R)$ for x and the lower bound agrees with the random coding error exponent $E_r(R)$ for source x. This shows that our bounding technique for the upper bound is tight in this case. To bring the gap between the upper bound and the lower bound, the coding scheme should be the optimal delay coding scheme in Chapter 2. The sequential random binning scheme is clearly suboptimal. Because it is proved in Section 2.5 that the delay constrained error exponent $E_s(R)$ is strictly higher than the random coding error exponent $E_r(R)$. This is clearly shown in Figure 5.9.

The source is the same as it in Section 2.2. Source x with alphabet size 3, $\mathcal{X} = \{A, B, C\}$ and the following distribution

$$p_x(A) = 0.65 \quad p_x(B) = 0.175 \quad p_x(C) = 0.175$$

The side-information is independent with the source, so its distribution does not matter. We arbitrarily let the marginal $p_y = \{0.420, 0.580\}$.

5.4.2 Special case 2: compression of encrypted data

In Section 5.2.2, we show that the delay constrained error exponent for compression of encrypted data for source p_s is sandwiched by the block coding error exponent $E_{s,b}(R, p_s)$ and the random coding error exponent $E_r(R, p_s)$, in Corollary 1. For source s the stream cipher is $x = s \oplus y$. For the binary case, this stream cipher is illustrated in Figure 5.10. Where the source $p_s(0) = 1 - \epsilon$ and $p_s(1) = \epsilon$, both the key y and the output of the cipher x are uniform on $\{0, 1\}$. The bounds on the delay constrained error exponents are plotted in Figure 5.11. In the example in Figure 5.11, $\epsilon = 0.1$. For this problem the entropy of the source $H(s)$ is equal to the conditional entropy $H(x|y)$.

For uniform source x and the side-information that is the output of a symmetric channel with x as the input, it can be shown that the upper bound and lower bound agree at low

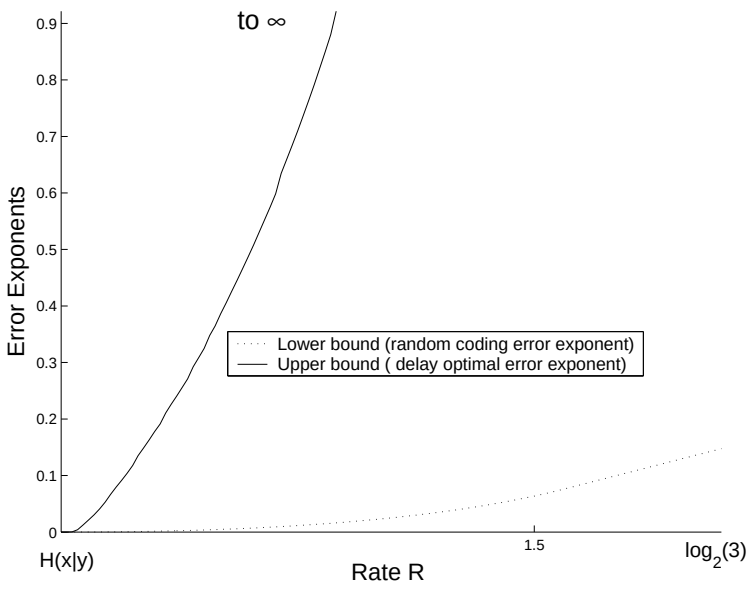

Figure 5.9. Delay constrained Error exponents for source coding with independent decoder side-information: Upper bound $E_{si}^{upper}(R) = E_s(R)$, Lower bound $E_{si}^{lower}(R) = E_r(R)$

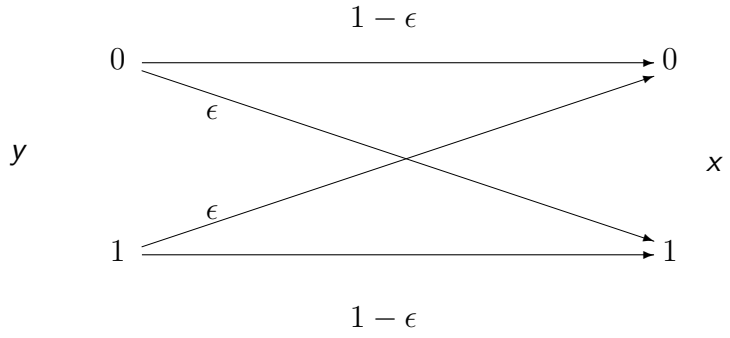

Figure 5.10. A stream cipher $x = s \oplus y$ can be modeled as a discrete memoryless channel, where key y is uniform and independent with source s. Key y is the input, encryption x is the output of the channel.

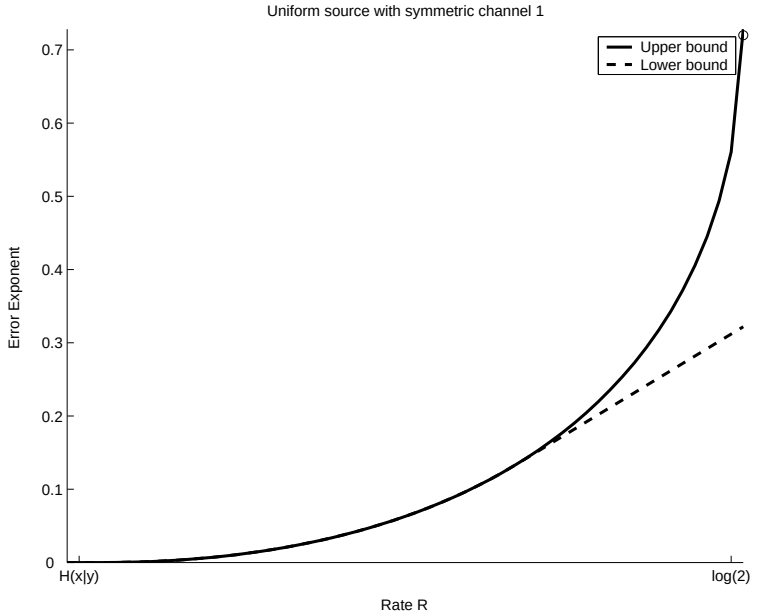

Figure 5.11. Delay constrained Error exponents for uniform source coding with symmetric decoder side-information $x = y \oplus s$: Upper bound $E_{si}^{upper}(R) = E_s(R, p_s)$, Lower bound $E_{si}^{lower}(R) = E_r(R, p_s)$. These two bounds agree in the low rate regime

rate regime just as shown in Figure 5.11. *Note:*this is a more general problem than the compression of encrypted data problem. A uniform source with erasure channel between the source and the side-information is illustrated in 5.12. The bounds on error exponents are shown in Figure 5.13.

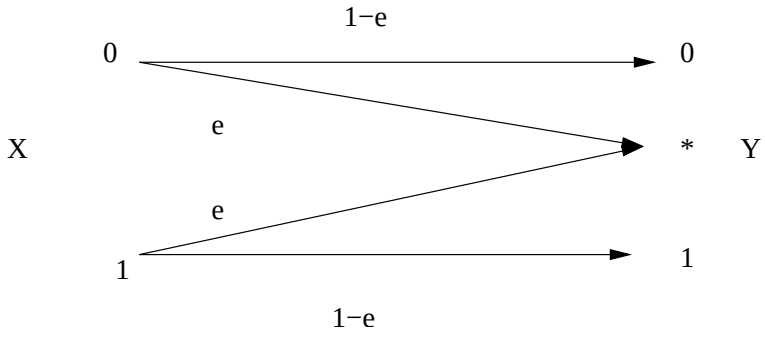

Figure 5.12. uniform source and side-information connected by a symmetric erasure channel

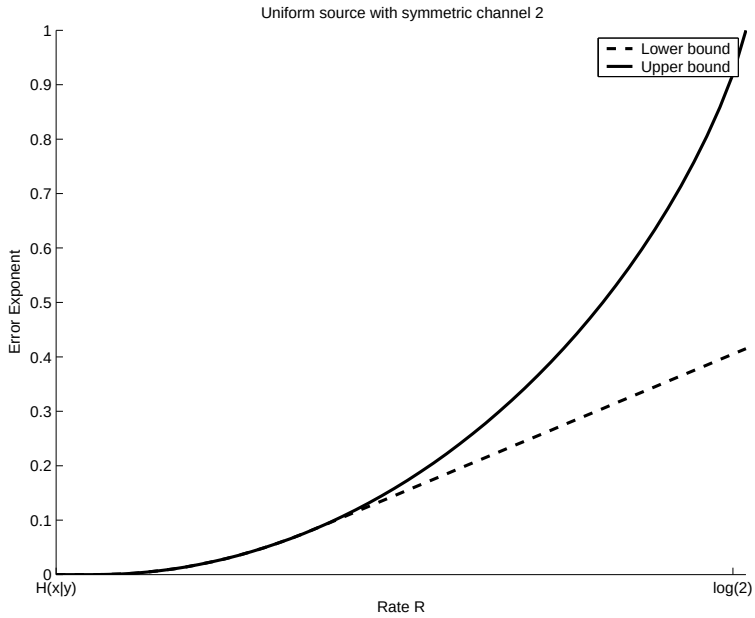

Figure 5.13. Delay constrained Error exponents for uniform source coding with symmetric decoder side-information 2 x and y are connected by an erasure channel: Upper bound $E_{si}^{upper}(R) = E_r(R, p_s)$ in the low rate regime

5.4.3 General cases

In this section, we show the upper bound and the lower bound for a *general* distribution of the source and side-information. That is, the side-information is dependent of the source and from the encoder point of view the source is not uniform. The uniformity here includes the marginal distribution of the source and the side-information. For example the following distribution is not *uniform* although the marginals are uniform:

$p_{xy}(x, y)$	x=1	x=2	x=3
y=1	a	b	$\frac{1}{3} - a - b$
y=2	c	d	$\frac{1}{3} - c - d$
y=3	$\frac{1}{3} - a - c$	$\frac{1}{3} - b - d$	$-\frac{1}{3} + a + b + c + d$

Table 5.1. A non-uniform source with uniform marignals

Where $a, b, c, d, a+b, a+c, d+c, d+b \in [0, \frac{1}{3}]$, and the encoder can do more than just sequential random binning. An extreme case where the encoder only need to deal with $x = 2, 3$ is as follows: let $a = \frac{1}{3}$ and $b = c = 0$, then the side information at the decoder can

be used to tell when $x = 1$. Without proof, we claim that for this marginal-uniform source, random coding is suboptimal. Now consider the following 2×2 source:

$$p_{xy} = \begin{pmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{pmatrix} \tag{5.9}$$

In Figure 5.14, we plot the upper bound, the lower bound on the delay constrained error exponent with decoder only side-information. To illustrate the "price of ignorance" phenomenon, we also plot the delay constrained error exponent with both encoder and decoder side-information: $E_{ei}(R)$. As the lossless source coding with delay constraints error exponent $E_s(R)$, $E_{ei}(R)$ is related to its block coding error exponent with the focusing operator.

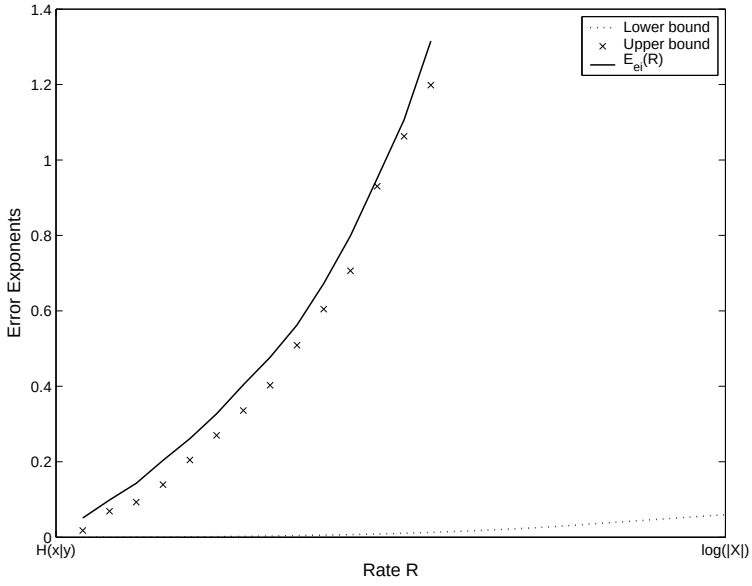

Figure 5.14. Delay constrained Error exponents for general source and decoder side-information: both the upper bound $E_{si}^{upper}(R)$ and the lower bound $E_{si}^{lower}(R)$ are plotted, the delay constrained error exponent with both encoder and decoder side information $E_{ei}(R)$ is plotted in dotted lines

Note: we do not know how to parameterize the upper bound $E_{si}^{upper}(R)$ as what we did in Proposition 7 in Section 2. Instead, we brutal forcefully minimize

$$\begin{aligned} E_{si}^{upper}(R) = \min \quad & \{ \inf_{q_{xy}, \alpha \geq 1 : H(q_{x|y}) > (1+\alpha)R} \{ \frac{1}{\alpha} D(q_{xy} \| p_{xy}) \}, \\ & \inf_{q_{xy}, 1 \geq \alpha \geq 0 : H(q_{x|y}) > (1+\alpha)R} \{ \frac{1-\alpha}{\alpha} D(q_x \| p_x) + D(q_{xy} \| p_{xy}) \} \} \end{aligned}$$

on a 5 dimensional space (q_{xy}, α). This gives a not so smooth plot as shown in Figure 5.14.

5.5 Proofs

The achievability part of the proof is a special case for that of the Slepian-Wolf source coding shown in Chapter 4. The upper bound is derived by a feedforward decoding scheme which was first developed in channel coding context. There is no surprise that we can borrow the channel coding technique to source coding with decoder side-information since it has been long known the duality between the two. The upper bound and the achievable lower bound are not identical in general.

5.5.1 Proof of Theorem 6: random binning

We prove Theorem 6 by following the proofs for point-to-point source coding in Propositions 3 and 4. The encoder is the same sequential random binning encoder in Definition 3 with common randomness shared by the encoder and decoder. Recall that the key property of this random binning scheme is the pair-wise independence, formally, for all i, n:

$$\Pr[\mathcal{E}(x_1^i x_{i+1}^n) = \mathcal{E}(x_1^i \tilde{x}_{i+1}^n)] = 2^{-(\lfloor nR \rfloor - \lfloor iR \rfloor)} \leq 2 \times 2^{-(n-i)R} \tag{5.10}$$

First we show (5.1) in Theorem 6.

ML decoding

ML decoding rule:

Denote by $\widehat{x}_1^n(n)$ the estimate of the source sequence x_1^n at time n.

$$\widehat{x}_1^n(n) = \underset{\tilde{x}_1^n \in \mathcal{B}_x(x_1^n)}{\arg\max}\, p_{\mathsf{xy}}(\tilde{x}_1^n, y_1^n) = \underset{\tilde{x}_1^n \in \mathcal{B}_x(x_1^n)}{\arg\max} \prod_{i=1}^{n} p_{\mathsf{xy}}(\tilde{x}_i, y_i) \tag{5.11}$$

The ML decoding rule in (5.11) is very simple. At time n, the decoder simply picks the sequence $\widehat{x}_1^n(n)$ with the highest *joint* likelihood with the side-information y_1^n while $\widehat{x}_1^n(n)$ is in the same bin as the true sequence x_1^n. Now the estimate of source symbol $n - \Delta$ is simply the $(n - delay)^{th}$ symbol of $\widehat{x}_1^n(n)$, denoted by $\widehat{x}_{n-\Delta}(n)$.

Details of the proof:

The proof is quite similar to that of Proposition 3, we will omit some of the details of the proof in this thesis. To lead to a decoding error, there must be some false source sequence $\tilde{x}_1^n$ that satisfies three conditions: (i) it must be in the same bin (share the same parities)

as x_1^n, i.e., $\tilde{x}_1^n \in \mathcal{B}_x(x_1^n)$, (ii) it must be more likely than the true sequence given the same side-information y_1^n, i.e., $p_{xy}(\tilde{x}_1^n, y_1^n) > p_{xy}(x_1^n, y_1^n)$, and (iii) $\tilde{x}_l \neq x_l$ for some $l \leq n - \Delta$.

The error probability again can be union bounded as:

$$\begin{aligned}
&\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}] \\
\leq &\Pr[\widehat{\mathsf{x}}_1^{n-\Delta}(n) \neq \mathsf{x}_1^{n-\Delta}] \\
= &\sum_{x_1^n, y_1^n} \Pr[\widehat{\mathsf{x}}_1^{n-\Delta}(n) \neq x_1^{n-\Delta} | \mathsf{x}_1^n = x_1^n] p_{xy}(x_1^n, y_1^n) \\
= &\sum_{x_1^n, y_1^n} \sum_{l=1}^{n-\Delta} \Pr\left[\exists\, \tilde{x}_1^n \in \mathcal{B}_x(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } p_{xy}(\tilde{x}_1^n, y_1^n) \geq p_{xy}(x_1^n, y_1^n)\right] p_{xy}(x_1^n, y_1^n) \\
= &\sum_{l=1}^{n-\Delta} \Big\{ \sum_{x_1^n, y_1^n} \Pr\left[\exists\, \tilde{x}_1^n \in \mathcal{B}_x(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } p_{xy}(\tilde{x}_1^n, y_1^n) \geq p_{xy}(x_1^n, y_1^n)\right] p_{xy}(x_1^n, y_1^n) \Big\} \\
= &\sum_{l=1}^{n-\Delta} p_n(l) \qquad (5.12)
\end{aligned}$$

Recall the definition of $\mathcal{F}_n(l, x_1^n)$ in (2.20)

$$\mathcal{F}_n(l, x_1^n) = \{\tilde{x}_1^n \in \mathcal{X}^n | \tilde{x}_1^{l-1} = x_1^{l-1}, \tilde{x}_l \neq x_l\}$$

and we define

$$p_n(l) = \sum_{x_1^n, y_1^n} \Pr\left[\exists\, \tilde{x}_1^n \in \mathcal{B}_x(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } p_{xy}(\tilde{x}_1^n, y_1^n) \geq p_{xy}(x_1^n, y_1^n)\right] p_{xy}(x_1^n, y_1^n) \quad (5.13)$$

We now upper bound $p_n(l)$ using a Chernoff bound argument similar to [39] and in the proof of Lemma 1. The details of the proof of the following Lemma 13 is in Appendix H.1.

Lemma 13 $p_n(l) \leq 2 \times 2^{-(n-l+1)E_{si}^{lower}(R)}$.

Using the above lemma, substitute the bound on $p_n(l)$ into (5.12), we prove the ML decoding part, (5.1), in Theorem 6. ■

Universal decoding (sequential minimum empirical joint entropy decoding)

In this section we prove (5.2) in Theorem 6.

Universal decoding rule:

$$\widehat{x}_l(n) = w[l]_l \quad \text{where} \quad w[l]_1^n = \underset{\bar{x}^n \in \mathcal{B}_x(x_1^n) \text{ s.t. } \bar{x}_1^{l-1} = \widehat{x}_1^{l-1}(n)}{\arg\min} H(\bar{x}_l^n, y_l^n). \qquad (5.14)$$

We term this a sequential minimum joint empirical-entropy decoder which tightly follows the sequential minimum empirical-entropy decoder for point-to-point source coding in (2.30).

Details of the proof: With this decoder, errors can only occur if there is some sequence $\tilde{x}_1^n$ such that (i) $\tilde{x}_1^n \in \mathcal{B}_x(x_1^n)$, (ii) $\tilde{x}_1^{l-1} = x^{l-1}$, and $\tilde{x}_l \neq x_l$, for some $l \leq n - \Delta$, and (iii) the joint empirical entropy of $\tilde{x}_l^n$ and y_l^n is such that $H(\tilde{x}_l^n, y_l^n) < H(x_l^n, y_l^n)$. Building on the common core of the achievability (5.12) with the substitution of universal decoding in the place of maximum likelihood results in the following definition of $p_n(l)$:

$$p_n(l) = \sum_{x_1^n, y_1^n} \Pr\left[\exists\, \tilde{x}_1^n \in \mathcal{B}_x(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } H(\tilde{x}_l^n, y_l^n) \leq H(x_l^n, y_l^n)\right] p_{xy}(x_1^n, y_1^n) \quad (5.15)$$

The following lemma gives a bound on $p_n(l)$. The proof is in Appendix H.2.

Lemma 14 *For sequential joint minimum empirical entropy decoding,*

$$p_n(l) \leq 2 \times (n - l + 2)^{2|\mathcal{X}||\mathcal{Y}|} 2^{-(n-l+1)E_{si}^{lower}(R)}.$$

Lemma 14 and $\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}] \leq \sum_{l=1}^{n-\Delta} p_n(l)$ imply that:

$$\begin{aligned}
\Pr[\widehat{\mathsf{x}}_{n-\Delta}(n) \neq \mathsf{x}_{n-\Delta}] &\leq \sum_{l=1}^{n-\Delta} (n - l + 2)^{2|\mathcal{X}||\mathcal{Y}|} 2^{-(n-l+1)E_{si}^{lower}(R)} \\
&\leq \sum_{l=1}^{n-\Delta} K_1 2^{-(n-l+1)[E_{si}^{lower}(R) - \epsilon]} \\
&\leq K 2^{-\Delta[E_{si}^{lower}(R) - \epsilon]}
\end{aligned}$$

where K and K_1 are finite constants. The above analysis follows the same argument of that in the proof of Proposition 4. This concludes the proof of the universal decoding part, (5.2), in Theorem 6. ■

5.5.2 Proof of Theorem 7: Feed-forward decoding

The theorem is proved by applying a variation of the bounding technique used in [67] (and originating in [62]) for the fixed-delay channel coding problem. Lemmas 15-20 are the source coding counterparts to Lemmas 4.1-4.5 in [67]. The idea of the proof is to first build a feed-forward sequential source decoder which has access to the previous source symbols in addition to the encoded bits and the side-information. The second step is to construct a block source-coding scheme from the optimal feed-forward sequential decoder and showing

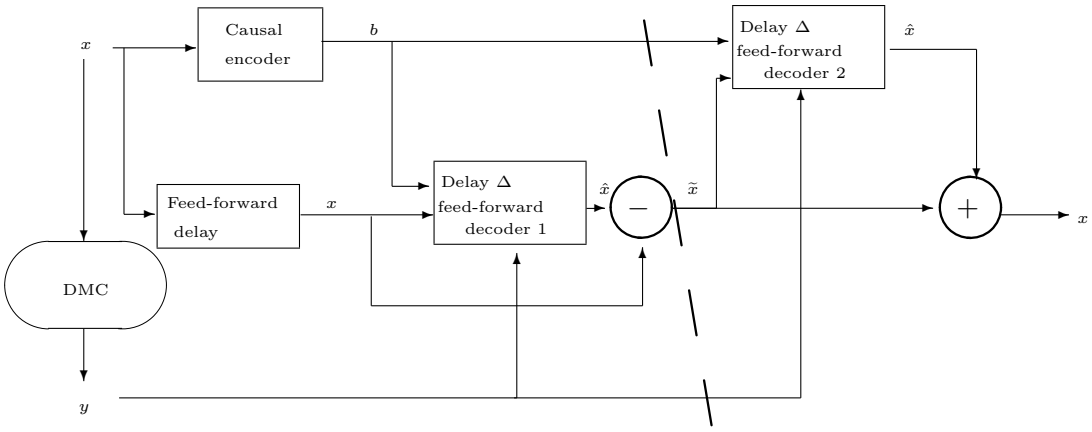

Figure 5.15. A cutset illustration of the Markov Chain $x_1^n - (\tilde{x}_1^n, b_1^{\lfloor (n+\Delta)R \rfloor}, y_1^{n+\Delta}) - x_1^n$. Decoder 1 and decoder 2 are type I and II delay Δ rate R feed-forward decoder respectively. They are equivalent.

that if the side-information behaves atypically enough, then the decoding error probability will be large for at least one of the source symbols. The next step is to prove that the atypicality of the side-information before that particular source symbol does not cause the error because of the feed-forward information. Thus, cause of the decoding error for that particular symbol is the atypical behavior of the future side-information only. The last step is to lower bound the probability of the atypical behavior and upper bound the error exponents. The proof spans into the next several subsections.

Feed-forward decoders

Definition 12 *A delay Δ rate R decoder $\mathcal{D}^{\Delta,R}$ with feed-forward is a decoder $\mathcal{D}_j^{\Delta,R}$ that also has access to the past source symbols x_1^{j-1} in addition to the encoded bits $b_1^{\lfloor (j+\Delta)R \rfloor}$ and side-information $y_1^{j+\Delta}$.*

Using this feed-forward decoder, the estimate of x_j at time $j+\Delta$ is :

$$\widehat{x}_j(j+\Delta) = \mathcal{D}_j^{\Delta,R}(b_1^{\lfloor (j+\Delta)R \rfloor}, y_1^{j+\Delta}, x_1^{j-1}) \tag{5.16}$$

Lemma 15 *For any rate R encoder $\mathcal{E}$, the optimal delay Δ rate R decoder $\mathcal{D}^{\Delta,R}$ with feed-forward only needs to depend on $b_1^{\lfloor (j+\Delta)R \rfloor}, y_j^{j+\Delta}, x_1^{j-1}$*

Proof: The source and side-information (x_i, y_i) is an iid random process and the encoded bits $b_1^{\lfloor (j+\Delta)R \rfloor}$ are functions of $x_1^{j+\Delta}$ so obeys the Markov chain: y_1^{j-1} –

$(x_1^{j-1}, b_1^{\lfloor (j+\Delta)R \rfloor}, y_j^{j+\Delta}) \;-\; x_j^{j+\Delta}$. Conditioned on the past source symbols, the past side-information is completely irrelevant for estimation. □

Notice that for any finite alphabet set $\mathcal{X}$, we can always define a group $Z_{|\mathcal{X}|}$ on $\mathcal{X}$, where the operators $-$ and $+$ are indeed $-, +$ mod $|\mathcal{X}|$. So we write the error sequence of the feed-forward decoder as $\tilde{x}_i = x_i - \widehat{x}_i$. Then we have the following property for the feed-forward decoders.

Lemma 16 *Given a rate R encoder $\mathcal{E}$, the optimal delay Δ rate R decoder $\mathcal{D}^{\Delta,R}$ with feed-forward for symbol j only needs to depend on $b_1^{\lfloor (j+\Delta)R \rfloor}, y_1^{j+\Delta}, \tilde{x}_1^{j-1}$*

Proof: Proceed by induction. It holds for $j = 1$ since there are no prior source symbols. Suppose that it holds for all $j < k$ and consider $j = k$. By the induction hypothesis, the action of all the prior decoders j can be simulated using $(b_1^{\lfloor (j+\Delta)R \rfloor}, y_1^{j+\Delta}, \tilde{x}_1^{j-1})$ giving $\widehat{x}_1^{k-1}$. This in turn allows the recovery of x_1^{k-1} since we also know $\tilde{x}_1^{k-1}$. Thus the decoder is equivalent. □

We call the feed-forward decoders in Lemmas 15 and 16 type I and II delay Δ rate R feed-forward decoders respectively. Lemma 15 and 16 tell us that feed-forward decoders can be thought in three ways: having access to all encoded bits, all side-information and all past source symbols, $(b_1^{\lfloor (j+\Delta)R \rfloor}, y_1^{j+\Delta}, x_1^{j-1})$, having access to all encoded bits, a recent window of side information and all past source symbols, $(b_1^{\lfloor (j+\Delta)R \rfloor}, y_j^{j+\Delta}, x_1^{j-1})$, or having access to all encoded bits, all side-information and all past decoding errors, $(b_1^{\lfloor (j+\Delta)R \rfloor}, y_1^{j+\Delta}, \tilde{x}_1^{j-1})$.

Constructing a block code

To encode a block of n source symbols, just run the rate R encoder $\mathcal{E}$ and terminate with the encoder run using some random source symbols drawn according to the distribution of p_{x} with matching side-information on the other side. To decode the block, just use the delay Δ rate R decoder $\mathcal{D}^{\Delta,R}$ with feed-forward, and then use the fedforward error signals to correct any mistakes that might have occurred. As a block coding system, this hypothetical system never makes an error from end to end. As shown in Figure 5.15, the data processing inequality implies:

Lemma 17 *If n is the block-length, the block rate is $R(1 + \frac{\Delta}{n})$, then*

$$H(\tilde{\mathsf{x}}_1^n) \geq -(n + \Delta)R + nH(\mathsf{x}|\mathsf{y}) \tag{5.17}$$

Proof:

$$\begin{aligned} nH(x) &=_{(a)} H(x_1^n) \\ &= I(x_1^n; x_1^n) \\ &=_{(b)} I(x_1^n; \tilde{x}_1^n, b_1^{\lfloor (n+\Delta)R \rfloor}, y_1^{n+\Delta}) \\ &=_{(c)} I(x_1^n; y_1^{n+\Delta}) + I(x_1^n; \tilde{x}_1^n | y_1^{n+\Delta}) + I(x_1^n; b_1^{\lfloor (n+\Delta)R \rfloor} | y_1^{n+\Delta}, \tilde{x}_1^n) \\ &\leq_{(d)} nI(x,y) + H(\tilde{x}_1^n) + H(b_1^{\lfloor (n+\Delta)R \rfloor}) \\ &\leq nH(x) - nH(x|y) + H(\tilde{x}_1^n) + (n+\Delta)R \end{aligned}$$

(a) is true because the source is iid. (b) is true because of the data processing inequality considering the following Markov chain: $x_1^n - (\tilde{x}_1^n, b_1^{\lfloor (n+\Delta)R \rfloor}, y_1^n) - x_1^n$, thus $I(x_1^n; x_1^n) \leq I(x_1^n; \tilde{x}_1^n, b_1^{\lfloor (n+\Delta)R \rfloor}, y_1^{n+\Delta})$. And the fact that $I(x_1^n; x_1^n) = H(x_1^n) \geq I(x_1^n; \tilde{x}_1^n, b_1^{\lfloor (n+\Delta)R \rfloor}, y_1^{n+\Delta})$. Combining the two equalities we get (b). (c) is the chain rule for mutual information. In (d), first notice that (x, y) are iid across time, thus $I(x_1^n; y_1^{n+\Delta}) = I(x_1^n; y_1^n) = nI(x,y)$. Secondly entropy of a random variable is never less than the mutual information of that random variable with another one, condition on other random variable or not. Others are obvious. □

Lower bound the symbol-wise error probability

Now suppose this block-code were to be run with the distribution q_{xy}, s.t. $H(q_{x|y}) > (1 + \frac{\Delta}{n})R$, from time 1 to n, and were to be run with the distribution p_{xy} from time $n+1$ to $n+\Delta$. Write the hybrid distribution as Q_{xy}. Then the block coding scheme constructed in the previous section will with probability 1 make a block error. Moreover, many individual symbols will also be in error often:

Lemma 18 *If the source and side-information is coming from q_{xy}, then there exists a $\delta > 0$ so that for n large enough, the feed-forward decoder will make at least $\frac{H(q_{x|y}) - \frac{n+\Delta}{n}R}{2\log_2|\mathcal{X}| - (H(q_{x|y}) - \frac{n+\Delta}{n}R)} n$ symbol errors with probability δ or above. δ satisfies*

$$h_\delta + \delta \log_2(|\mathcal{X}| - 1) = \frac{1}{2}(H(q_{x|y}) - \frac{n+\Delta}{n}R),$$

where $h_\delta = -\delta \log_2 \delta - (1-\delta)\log_2(1-\delta)$.

Proof: Lemma 17 implies:

$$\sum_{i=1}^{n} H(\tilde{x}_i) \quad \geq \quad H(\tilde{x}_1^n) \geq -(n+\Delta)R + nH(q_{x|y}) \tag{5.18}$$

The average entropy per source symbol for $\tilde{x}$ is at least $H(q_{x|y}) - \frac{n+\Delta}{n}R$. Now suppose that $H(\tilde{x}_i) \geq \frac{1}{2}(H(q_{x|y}) - \frac{n+\Delta}{n}R)$ for A positions. By noticing that $H(\tilde{x}_i) \leq \log_2 |\mathcal{X}|$, we have

$$\sum_{i=1}^{n} H(\tilde{x}_i) \leq A \log_2 |\mathcal{X}| + (n-A)\frac{1}{2}(H(q_{x|y}) - \frac{n+\Delta}{n}R)$$

With (5.18), we derive the desired result:

$$A \geq \frac{(H(q_{x|y}) - \frac{n+\Delta}{n}R)}{2\log_2 |\mathcal{X}| - (H(q_{x|y}) - \frac{n+\Delta}{n}R)}n \tag{5.19}$$

Where $2\log_2 |\mathcal{X}| - (H(q_{x|y}) - \frac{n+\Delta}{n}R) \geq 2\log_2 |\mathcal{X}| - H(q_{x|y}) \geq 2\log_2 |\mathcal{X}| - \log_2 |\mathcal{X}| > 0$

Now for A positions $1 \leq j_1 < j_2 < ... < j_A \leq n$ the individual entropy $H(\tilde{x}_j) \geq \frac{1}{2}(H(q_{x|y}) - \frac{n+\Delta}{n}R)$. By the property of the binary entropy function,

$$\Pr[\tilde{x}_j \neq x_0] = \Pr[x_j \neq \hat{x}_j] \geq \delta,$$

where x_0 is the zero element in the finite group $Z_{|\mathcal{X}|}$. □

We can pick $j^* = j_{\frac{A}{2}}$, by Lemma 18, we know that $\min\{j^*, n-j^*\} \geq \frac{1}{2}\frac{(H(q_{x|y}) - \frac{n+\Delta}{n}R)}{2\log_2 |\mathcal{X}| - (H(q_{x|y}) - \frac{n+\Delta}{n}R)}n$, so if we fix $\frac{\Delta}{n}$ and let n go to infinity, then $\min\{j^*, n-j^*\}$ goes to infinity as well.

At this point, Lemma 15 and 18 together imply that even if the source and side-information only behaves like it came from the hybrid distribution Q_{xy} from time j^* to $j^*+\Delta$ *and* the source behaves like it came from a distribution q_x from time 1 to j^*-1, the same minimum error probability δ still holds. Now define the "bad sequence" set E_{j^*} as the set of source and side-information sequence pairs so the type I delay Δ rate R decoder makes an decoding error at j^*. Formally

$$E_{j^*} = \{(\vec{x}, \bar{\bar{y}}) | x_{j^*} \neq \mathcal{D}_{j^*}^{\Delta,R}(\mathcal{E}(\vec{x}), y_j^{j^*+\Delta}, \bar{x})\},$$

where to simplify the notation, we write: $\vec{x} = x_1^{j^*+\Delta}$, $\bar{x} = x_1^{j^*-1}$, $\bar{\bar{x}} = x_{j^*}^{j^*+\Delta}$, $\bar{\bar{y}} = y_{j^*}^{j^*+\Delta}$. By Lemma 18, $Q_{xy}(E_{j^*}) \geq \delta$. Notice that E_{j^*} does not depend on the distribution of the source but only on the encoder-decoder pair. Define $J = \min\{n, j^*+\Delta\}$, and $\bar{\bar{\bar{x}}} = x_{j^*}^J$,

$\bar{\bar{y}} = y_{j^*}^J$. Now we write the strongly typical set

$$\begin{aligned} A_J^\epsilon(q_{xy}) = \{ \quad & (\vec{x}: \quad \bar{\bar{y}}) \in \mathcal{X}^{j^*+\Delta} \times \mathcal{Y}^{\Delta+1} | \forall x, r_{\bar{x}}(x) \in (q_x(x) - \epsilon, q_x(x) + \epsilon) \\ & \forall (x,y) \in \mathcal{X} \times \mathcal{Y}, \quad q_{xy}(x,y) > 0 : r_{\bar{\bar{x}},\bar{\bar{y}}}(x,y) \in (q_{xy}(x,y) - \epsilon, q_{xy}(x,y) + \epsilon), \\ & \forall (x,y) \in \mathcal{X} \times \mathcal{Y}, \quad q_{xy}(x,y) = 0 : r_{\bar{\bar{x}},\bar{\bar{y}}}(x,y) = 0 \quad \} \end{aligned}$$

where the empirical distribution of $(\bar{\bar{x}}, \bar{\bar{y}})$ is denoted by $r_{\bar{\bar{x}},\bar{\bar{y}}}(x,y) = \frac{n_{x,y}(\bar{\bar{x}},\bar{\bar{y}})}{\Delta+1}$, the empirical distribution of $\bar{x}$ by $r_{\bar{x}}(x) = \frac{n_x(\bar{x})}{j^*-1}$.

Lemma 19 *$Q_{xy}(E_{j^*} \cap A_J^\epsilon(q_{xy})) \geq \frac{\delta}{2}$ for large n and Δ.*

Proof: Fix $\frac{\Delta}{n}$, let n go to infinity, then $\min\{j^*, n - j^*\}$ goes to infinity. By the definition of J, $\min\{j^*, J - j^*\}$ goes to infinity as while. By Lemma 13.6.1 in [26], we know that $\forall \epsilon > 0$, if $J - j^*$ and j^* are large enough, then $Q_{xy}(A_J^\epsilon(q_{xy})^C) \leq \frac{\delta}{2}$. By Lemma 18, $Q_{xy}(E_{j^*}) \geq \delta$. So

$$Q_{xy}(E_{j^*} \cap A_J^\epsilon(q_{xy})) \geq Q_{xy}(E_{j^*}) - Q_{xy}(A_J^\epsilon(q_{xy})^C) \geq \frac{\delta}{2}$$

□

Lemma 20 *For all $\epsilon < \min_{x,y:p_{xy}(x,y)>0}\{p_{xy}(x,y)\}$, $\forall (\vec{x}, \bar{\bar{y}}) \in A_J^\epsilon(q_{xy})$,*

$$\frac{p_{xy}(\vec{x}, \bar{\bar{y}})}{Q_{xy}(\vec{x}, \bar{\bar{y}})} \geq 2^{-(J-j^*+1)D(q_{xy}\|p_{xy}) - (j^*-1)D(q_x\|p_x) - JG\epsilon}$$

where $G = \max\{|\mathcal{X}||\mathcal{Y}| + \sum_{x,y:p_{xy}(x,y)>0} \log_2(\frac{q_{xy}(x,y)}{p_{xy}(x,y)} + 1), |\mathcal{X}| + \sum_x \log_2(\frac{q_x(x)}{p_x(x)} + 1)\}$

Proof: For $(\vec{x}, \bar{\bar{y}}) \in A_J^\epsilon(q_{xy})$, by definition of the strong typical set, it can be easily shown by algebra: $D(r_{\bar{\bar{x}},\bar{\bar{y}}}\|p_{xy}) \leq D(q_{xy}\|p_{xy}) + G\epsilon$ and $D(r_{\bar{x}}\|p_x) \leq D(q_x\|p_x) + G\epsilon$.

$$\begin{aligned} \frac{p_{xy}(\vec{x}, \bar{\bar{y}})}{Q_{xy}(\vec{x}, \bar{\bar{y}})} &= \frac{p_{xy}(\bar{x})}{q_{xy}(\bar{x})} \frac{p_{xy}(\bar{\bar{x}}, \bar{\bar{x}})}{q_{xy}(\bar{\bar{x}}, \bar{\bar{y}})} \frac{p_{xy}(x_{J+1}^{j^*+\Delta}, y_{J+1}^{j^*+\Delta})}{p_{xy}(x_{J+1}^{j^*+\Delta}, y_{J+1}^{j^*+\Delta})} \\ &= \frac{2^{-(J-j^*+1)(D(r_{\bar{\bar{x}},\bar{\bar{y}}}\|p_{xy}) + H(r_{\bar{\bar{x}},\bar{\bar{y}}}))}}{2^{-(J-j^*+1)(D(r_{\bar{\bar{x}},\bar{\bar{y}}}\|q_{xy}) + H(r_{\bar{\bar{x}},\bar{\bar{y}}}))}} \frac{2^{-(j^*-1)(D(r_{\bar{x}}\|p_x) + H(r_{\bar{x}}))}}{2^{-(j^*-1)(D(r_{\bar{x}}\|q_x) + H(r_{\bar{x}}))}} \\ &\geq_{(a)} 2^{-(J-j^*+1)(D(q_{xy}\|p_{xy}) + G\epsilon) - (j^*-1)(D(q_x\|p_x) + G\epsilon)} \\ &= 2^{-(J-j^*+1)D(q_{xy}\|p_{xy}) - (j^*-1)D(q_x\|p_x) - JG\epsilon} \end{aligned}$$

(a) is true by Equation 12.60 in [26]. □

Lemma 21 *For all* $\epsilon < \min_{x,y}\{p_{xy}(x,y)\}$*, and large* Δ*,* n*:*

$$p_{xy}(E_{j^*}) \geq \frac{\delta}{2} 2^{-(J-j^*+1)D(q_{xy}\|p_{xy})-(j^*-1)D(q_x\|p_x)-JG\epsilon}$$

Proof: Combining Lemma 19 and 20:

$$\begin{aligned} p_{xy}(E_{j^*}) &\geq p_{xy}(E_{j^*} \cap A_J^\epsilon(q_{xy})) \\ &\geq q_{xy}(E_{j^*} \cap A_J^\epsilon(q_{xy}))2^{-(J-j^*+1)D(q_{xy}\|p_{xy})-(j^*-1)D(q_x\|p_x)-JG\epsilon} \\ &\geq \frac{\delta}{2} 2^{-(J-j^*+1)D(q_{xy}\|p_{xy})-(j^*-1)D(q_x\|p_x)-JG\epsilon} \end{aligned}$$

□

Final touch of the proof of Theorem 7

Now we are finally ready to prove Theorem 7. Notice that as long as $H(q_{x|y}) > \frac{n+\Delta}{n}R$, we know $\delta > 0$ by letting ϵ go to 0, Δ and n go to infinity proportionally. We have: $\Pr[\widehat{x}_{j^*}(j^*+\Delta) \neq x_{j^*}] = p_{xy}(E_{j^*}) \geq K2^{-(J-j^*+1)D(q_{xy}\|p_{xy})-(j^*-1)D(q_x\|p_x)}$.

Notice that $D(q_{xy}\|p_{xy}) \geq D(q_x\|p_x)$ and $J = \min\{n, j^*+\Delta\}$, then for all possible $j^* \in [1,n]$, we have: for $n \geq \Delta$

$$\begin{aligned} (J-j^*+1)D(q_{xy}\|p_{xy}) + (j^*-1)D(q_x\|p_x) &\leq (\Delta+1)D(q_{xy}\|p_{xy}) + (n-\Delta-1)D(q_x\|p_x) \\ &\approx \Delta(D(q_{xy}\|p_{xy}) + \frac{n-\Delta}{\Delta}D(q_x\|p_x)) \end{aligned}$$

For $n < \Delta$

$$(J-j^*+1)D(q_{xy}\|p_{xy}) + (j^*-1)D(q_x\|p_x) \leq nD(q_{xy}\|p_{xy}) = \Delta(\frac{n}{\Delta}D(q_{xy}\|p_{xy}))$$

Write $\alpha = \frac{\Delta}{n}$, then the upper bound on the error exponent is the minimum of the above error exponents over all $\alpha > 0$, i.e:

$$\begin{aligned} E_{si}^{upper}(R) = \min \quad & \{\inf_{q_{xy}, \alpha \geq 1: H(q_{x|y}) > (1+\alpha)R} \{\frac{1}{\alpha}D(q_{xy}\|p_{xy})\}, \\ & \inf_{q_{xy}, 1 \geq \alpha \geq 0: H(q_{x|y}) > (1+\alpha)R} \{\frac{1-\alpha}{\alpha}D(q_x\|p_x) + D(q_{xy}\|p_{xy})\}\} \end{aligned}$$

This finalizes the proof. ■

5.6 Discussions

In this chapter, we attempted to derive a tight upper bound on the delay constrained distributed source coding error exponent. For source coding with decoder only side-information, we give *an* upper bound by using a feed-forward coding argument borrowed from the channel coding literature [62, 67]. The upper bound is shown to be tight in two special cases, namely independent side-information and the low rate regime in the compression of encrypted data problem. We gave a generic achievability result by sequential random binning. The random coding based lower bound agrees with the upper bound in the compression of encrypted data problem. In the independent side-information case, the random coding error exponent is strictly suboptimal. For general cases, there is a gap between the upper and lower bounds in the whole rate region. We believe the lower bound can be improved by using a variable length random binning scheme. This is a difficult problem and should be further studied.

We also studied delay constrained source coding with both encoder and decoder side-information. With encoder side-information, this problem resembles lossless source coding, thus delay constrained coding has a focusing type bound. The exponent with both encoder and decoder side-information is strictly higher than the upper bound of the decoder only case. This phenomenon is called the "price of ignorance" [18] and is not observed in block coding. This is another example that block length is not the same as delay.

Chapter 6

Future Work

In this thesis, we studied the *asymptotic* performance bound for several delay constrained streaming source coding problems where the source is assumed to be *iid* with *constant* arrival rate. What are the performance bounds if the sources are not iid, what if the arrivals are random, what if the distortion is not defined on a symbol by symbol basis? What is the non asymptotic performance bounds for these problems? More importantly, the main theme of this thesis is to figure out the *dominant error event* on a symbol by symbol basis. For a specific source symbol with a finite end-to-end decoding delay constraint, what's the most likely atypical event that causes error? This is a fundamental question that is not answered yet for several cases.

6.1 Past, present and future

"This duality can be pursued further and is related to the duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it."

– Claude Shannon [74]

What is the past, what is the present and what is the future? In delay constrained streaming coding problems, for a source symbol that enters the encoder at time t, we define past as time prior to time t, present as time t, future as time between t and time $t + \Delta$, where Δ is the finite delay constraint. Time after $t + \Delta$ is irrelevant. With this definition, past, present and future is only relevant to a particular time t as shown in Figure 6.1

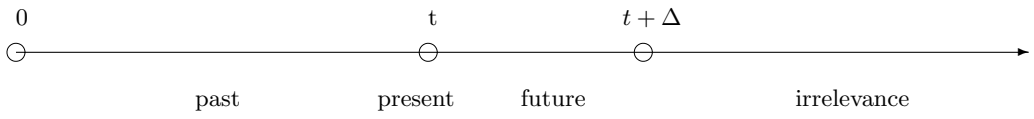

Figure 6.1. Past, present and future at time t

6.1.1 Dominant error event

For delay constrained streaming coding, what is the dominant error event (atypical event) for time t? For a particular coding scheme, we define the dominant error event as the event with the highest probability that makes a decoding error for the source symbol t. For delay constrained lossless source coding in Chapter 2, we discovered that the dominant error event for *optimal* coding is the atypical behavior of the source in the past, the future does not matter! This is shown in Figure 6.2. The same goes for the delay constrained lossy source coding problem in Chapter 3. However, the dominant error event for the sequential random binning scheme is the *future* atypicality of the binning and/or the source behavior as shown in Section 2.3.3. This illustrates that dominant error events are coding scheme dependent.

We summarize what we know about the dominant error events for the *optimal* coding schemes in Table 6.1. In th table, we put a ✓ mark if we completely characterized the nature of the dominant error event, and a ? mark if we have a conjecture on the nature of the dominant error event. *Note:*for source coding with decoder only side-information and

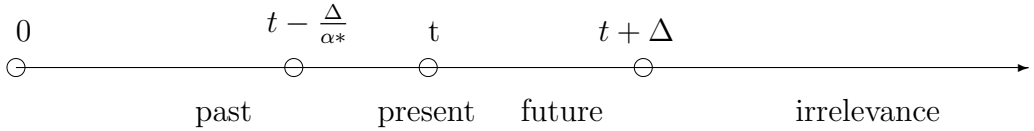

Figure 6.2. Dominant error event for delay constrained lossless source coding is the atypicality of the source between time $t - \frac{\Delta}{\alpha*}$ and time t, α^* is the optimizer in (2.7)

	Past	Future	Both
Point-to-point Lossless (Chapter 2)	✓		
Point-to-point Lossy (Chapter 3)	✓		
Slepian-Wolf coding (Chapter 4)			?
Source coding with decoder only side-info (Chapter 5)			?
Source coding with both side-information (Chapter 5)	✓		
Channel coding without feedback [67]		✓	
Erasure-like channel coding wit feedback [67]			✓
Joint source channel coding [15]			?
MAC [14], Broadcast channel coding [13]		?	

Table 6.1. Type of dominant error events for different coding problems

joint source channel coding, we derive the upper bound by using the feed-forward decoder scheme and the the dominant error event spans across past and future. However, the lower bounds (achievability) are derived by a suboptimal sequential random binning scheme whose dominant error event is in the future.

6.1.2 How to deal with both past and future?

In this thesis, we develop several tools to deal with different delay constrained streaming source coding problems. For the lower bounds on the error exponent, we use sequential random binning to deal with future dominant error events and we see that the scheme is often optimal as shown in [67]. We use variable length coding with FIFO queue to deal with past dominant error events as shown in Chapters 2 and 3. For the upper bounds, we use the feedforward decoding scheme and a simple focusing type argument to translate the delay constrained error event into a block coding error event and then derive the upper bounds. Obviously, more tools are needed to deal with more complex problems. At what point can we claim that we have developed all the needed tools for delay constrained streaming coding problems in Figure 1.1?

For those problems whose dominant error event spans across past and future, we do not have a concrete idea of what the optimal coding scheme is. The only exception is the

erasure-like channel with feedback problem in [67], this channel coding problem is a special case of source coding with delay. For more general cases, source coding with decoder side-information and joint source channel coding, we believe that the "variable length random binning" scheme should be studied. This is a difficult problem and more serious research needs to be done in this area.

6.2 Source model and common randomness

In this thesis, the sources are always modeled as iid random variables on a finite alphabet set. This assumption simplifies the analysis and captures the main challenges in streaming source coding with delay. Our analysis for lossless source coding in Chapter 2 can be easily generalized to finite alphabet stationary ergodic processes. The infinite alphabet case [59] is very challenging and our current universal proof technique does not apply. For lossy source coding in Chapter 3, an important question is: what if the sources are continuous random variables instead discrete and the distortion measures are L_2 distances? Is a variable length vector quantizer with FIFO queueing system still optimal? Another interesting case is that we do not have any assumption on the statistics of the source at all. This is the individual sequence problem [11]. It seems that our universal variable length code and FIFO queue would work just fine for individual sequences for the same reason that Lempel-Ziv coding is optimal for individual sequences. A rigorous proof is needed. In our study of lossy source coding, we focus on the per symbol loss case. What if the loss function is defined as an average over a period of time? What can we learn from the classical sliding-block source coding literature [45, 44, 7, 77]?

Another very interesting problem is when multiple sources have to share the same bandwidth and encoder/decoder. We recently studied the error exponent tradeoff in the block coding setup. The delay constrained performance for multiple streaming sources is a more challenging problem. Also, as mentioned in Section 1.2, random arrivals of the source symbols poses another dimension of challenges. Other than several special distributions of the inter-symbol random arrivals, we do not have a general result on the delay constrained error exponent. For the above problems, techniques from queueing theory [42] might be useful.

In the proofs of the achievabilities for delay constrained distributed source coding in Chapters 4 and 5 and channel coding [13, 14], we use a sequential random coding scheme. This assumes that the encoder(s) and the decoder(s) share some common randomness. The amount of randomness is unbounded since our coding system has an infinite horizon. A natural question is if there *exists a* encoder decoder pair that achieve the random coding error exponents without the presence of common randomness. For the block coding cases, the existence is proved by a simple argument [26, 41]. However, this argument does not apply to the delay constrained coding problems because we are facing infinite horizons.

6.3 Small but interesting problems

In Chapter 2, we showed several properties of the delay constrained source coding error exponent $E_s(R)$, including the positive derivative of the error exponent at the entropy rate etc. One important question is whether the exponent $E_s(R)$ is convex $\cup$. We conjecture that the answer is yes although the channel coding counter part can be neither convex nor concave [67]. We also conjecture that the delay constrained lossy source coding error exponent $E_D(R)$ is convex $\cup$. This is a more difficult problem, because there is no obvious parametrization of $E_D(R)$, thus the only possible way to show the concavity of $E_D(R)$ is through the definition of $E_D(R)$ and show a general result for all focusing type bounds. An important question is: what is the sufficient condition for $F(R)$ such that the following focusing function $E(R)$ is convex $\cup$?

$$E(R) = \inf_{\alpha>0} \frac{1}{\alpha} F((1+\alpha)R) \tag{6.1}$$

We do not have a parametrization result for the upper bound on the delay constrained source coding with decoder side-information as shown in Theorem 7. A parametrization result like that in Proposition 7 can greatly simplify the calculation of the upper bound.

We use the feed-forward coding scheme in upper bounding the error exponents. All the problems we have studied using this technique are point to point coding with one encoder and one decoder. An important future direction is to generalize our current technique to distributed source coding and channel coding. This should be a reasonably easy problem for delay constrained Slepian-Wolf coding by modifying the feed-forward diagram in Figure 5.15.

Lastly, we study the error exponents in the asymptotic regime, i.e. the error exponent tells how fast the error probability decays to *zero* when the delay is long as shown in (1.2) with $\approx$ instead of $=$. In order to accurately bound the error probability in the short delay regime, we cannot ignore the often ignored polynomial terms in the expressions of error probabilities. This could be an extremely laborious problem that lacks the mathematical neatness. We leave this problem to more practically minded researchers.

I do not expect my thesis be error free. Please send your comments, suggestions, questions, worries, concerns, solutions to the open problems to **chechang@ocf.berkeley.edu**

Bibliography

[1] Rudolf Ahlswede and Imre Csiszár. Common randomness in information theory and cryptography part I: Secret sharing. *IEEE Transactions on Information Theory*, 39:1121– 1132, 1993.

[2] Rudolf Ahlswede and Gunter Dueck. Good codes can be produced by a few permutations. *IEEE Transactions on Information Theory*, 28:430– 443, 1982.

[3] Venkat Anantharam and Sergio Verdu. Bits through queues. *IEEE Transactions on Information Theory*, 42:4– 18, 1996.

[4] Andrew Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998.

[5] Toby Berger. *Rate Distortion Theory: A mathematical basis for data compression.* Prentice-Hall, 1971.

[6] Toby Berger and Jerry D. Gibson. Lossy source coding. *IEEE Transactions on Information Theory*, 44:2693 – 2723, 1998.

[7] Toby Berger and Joseph KaYin Lau. On binary sliding block codes. *IEEE Transactions on Information Theory*, 23:343 – 353, 1977.

[8] Patrick Billingsley. *Probability and Measure.* Cambridge University Press, Wiley-Interscience, 1995.

[9] Shashi Borade and Lizhong Zheng. I-projection and the geometry of error exponents. *Allerton Conference*, 2006.

[10] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[11] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[12] Cheng Chang, Stark Draper, and Anant Sahai. Sequential random binning for streaming distributed source coding. *submitted to IEEE Transactions on Information Theory*, 2006.

[13] Cheng Chang and Anant Sahai. Sequential random coding error exponents for degraded broadcast channels. *Allerton Conference*, 2005.

[14] Cheng Chang and Anant Sahai. Sequential random coding error exponents for multiple access channels. *WirelessCom 05 Symposium on Information Theory*, 2005.

[15] Cheng Chang and Anant Sahai. Error exponents for joint source-channel coding with delay-constraints. *Allerton Conference*, 2006.

[16] Cheng Chang and Anant Sahai. Upper bound on error exponents with delay for lossless source coding with side-information. *ISIT*, 2006.

[17] Cheng Chang and Anant Sahai. Delay-constrained source coding for a peak distortion measure. *ISIT*, 2007.

[18] Cheng Chang and Anant Sahai. The price of ignorance: the impact on side-information for delay in lossless source coding. *submitted to IEEE Transactions on Information Theory*, 2007.

[19] Cheng Chang and Anant Sahai. Universal quadratic lower bounds on source coding error exponents. *Conference on Information Sciences and Systems*, 2007.

[20] Cheng Chang and Anant Sahai. The error exponent with delay for lossless source coding. *Information Theory Workshop, Punta del Este, Uruguay*, March 2006.

[21] Cheng Chang and Anant Sahai. Universal fixed-length coding redundancy. *Information Theory Workshop, Lake Tahoe, CA*, September 2007.

[22] Jun Chen, Dake He, Ashish Jagmohan, and Luis A. Lastras-Montano. On the duality and difference between Slepian-Wolf coding and channel coding. *Information Theory Workshop, Lake Tahoe, CA, USA*, September 2007.

[23] Gérard Cohen, Iiro Honkala, Simon Litsyn, and Antoine Lobstein. *Covering Codes*. Elsevier, 1997.

[24] Thomas M. Cover. Broadcast channels. *IEEE Transactions on Information Theory*, 18:2–14, 1972.

[25] Thomas M. Cover and Abbas El Gamal. Capacity theorems for the relay channel. *IEEE Transactions on Information Theory*, pages 572–584, 1979.

[26] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley and Sons Inc., New York, 1991.

[27] Imre Csiszár. Joint source-channel error exponent. *Problem of Control and Information Theory*, 9:315–328, 1980.

[28] Imre Csiszár. The method of types. *IEEE Transactions on Information Theory*, 44:2505–2523, 1998.

[29] Imre Csiszár and János Körner. *Information Theory*. Akadémiai Kiadó, Budapest, 1986.

[30] Imre Csiszár and Paul C. Shields. *Information Theory and Statistics: A Tutorial*. `http://www.nowpublishers.com/getpdf.aspx?doi=0100000004&product=CIT`.

[31] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 1998.

[32] Stark Draper, Cheng Chang, and Anant Sahai. Sequential random binning for streaming distributed source coding. *ISIT*, 2005.

[33] Richard Durrett. *Probability: Theory and Examples*. Duxbury Press, 1995.

[34] David Forney. Convolutional codes III. sequential decoding. *Information and Control*, 25:267–297, 1974.

[35] Robert Gallager. Fixed composition arguments and lower bounds to error probability. `http://web.mit.edu/gallager/www/notes/notes5.pdf`.

[36] Robert Gallager. *Low Density Parity Check Codes, Sc.D. thesis*. Massachusetts Institute Technology, 1960.

[37] Robert Gallager. Capacity and coding for degraded broadcast channels. *Problemy Peredachi Informatsii*, 10(3):3–14, 1974.

[38] Robert Gallager. Basic limits on protocol information in data communication networks. *IEEE Transactions on Information Theory*, 22:385–398, 1976.

[39] Robert Gallager. Source coding with side information and universal coding. Technical Report LIDS-P-937, Mass. Instit. Tech., 1976.

[40] Robert Gallager. A perspective on multiaccess channels. *IEEE Transactions on Information Theory*, 31, 1985.

[41] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley, New York, NY, 1971.

[42] Ayalvadi Ganesh, Neil O´Connell, and Damon Wischik. *Big queues*. Springer, Berlin/Heidelberg, 2004.

[43] Michael Gastpar. *To code or not to code, Phd Thesis*. École Polytechnique Fédérale de Lausanne, 2002.

[44] Robert M. Gray. Sliding-block source coding. *IEEE Transactions on Information Theory*, 21:357–368, 1975.

[45] Robert M. Gray, David L. Neuhoff, and Donald S. Ornstein. Nonblock source coding with a fidelity criterion. *The Annals of Probability*, 3(3):478–491, 195.

[46] Martin Hellman. An extension of the shannon theory approach to cryptography. *IEEE Transactions on Information Theory*, 23(3):289–294, 1977.

[47] Michael Horstein. Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory*, 9(3):136 – 143, 1963.

[48] Arding Hsu. Data management in delayed conferencing. *Proceedings of the 10th International Conference on Data Engineering*, page 402, Feb 1994.

[49] Frederick Jelinek. Buffer overflow in variable length coding of fixed rate sources. *IEEE Transactions on Information Theory*, 14:490–501, 1968.

[50] Mark Johnson, Prakash Ishwar, Vinod Prabhakaran, Dan Schonberg, and Kannan Ramchandran. On compressing encrypted data. *IEEE Transactions on Signal Processing*, 52(10):2992–3006, 2004.

[51] Aggelos Katsaggelos, Lisimachos Kondi, Fabian Meier, Jörn Ostermann, and Guido Schuster. Mpeg-4 and ratedistortion -based shape-coding techniques. *IEEE Proc., special issue on Multimedia Signal Processing*, 86:1126–1154, 1998.

[52] Leonard Kleinrock. *Queueing Systems*. Wiley-Interscience, 1975.

[53] Amos Lapidoth and Prakash Narayan. Reliable communication under channel uncertainty. *IEEE Transactions on Signal Processing*, 44:2148–2177, 1998.

[54] Michael E. Lukacs and David G. Boyer. A universal broadband multipoint teleconferencing service for the 21st century. *IEEE Communications Magazine*, 33:36 – 43, 1995.

[55] Katalin Marton. Error exponent for source coding with a fidelity criterion. *IEEE Transactions on Information Theory*, 20(2):197–199, 1974.

[56] Ueli M. Maurer. Secret key agreement by public discussion from common information. *IEEE Transactions on Information Theory*, 39(3):733742, 1993.

[57] Neri Merhav and Ioannis Kontoyiannis. Source coding exponents for zero-delay coding with finite memory. *IEEE Transactions on Information Theory*, 49(3):609–625, 2003.

[58] David L. Neuhoff and R. Kent Gilbert. Causal source codes. *IEEE Transactions on Information Theory*, 28(5):701–713, 1982.

[59] Alon Orlitsky, Narayana P. Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. Limit results on pattern entropy. *IEEE Transactions on Information Theory*, 52(7):2954–2964, 2006.

[60] Hari Palaiyanur and Anant Sahai. Sequential decoding using side-information. *IEEE Transactions on Information Theory*, submitted, 2007.

[61] Larry L. Peterson and Bruce S. Davie. *Computer Networks: A Systems Approach*. Morgan Kaufmann, 1999.

[62] Mark Pinsker. Bounds of the probability and of the number of correctable errors for nonblock codes. *Translation from Problemy Peredachi Informatsii*, 3:44–55, 1967.

[63] Vinod Prabhakaran and Kannan Ramchandran. On secure distributed source coding. *Information Theory Workshop, Lake Tahoe, CA, USA*, September 2007.

[64] S. Sandeep Pradhan, Jim Chou, and Kannan Ramchandran. Duality between source coding and channel coding and its extension to the side information case. *IEEE Transactions on Information Theory*, 49(5):1181– 1203, May 2003.

[65] Thomas J. Richardson and Rüdiger L. Urbanke. The capacity of low-density parity-check codes under message-passing decoding. *IEEE Transactions on Information Theory*, 47(2):599–618, 2001.

[66] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[67] Anant Sahai. Why block length and delay are not the same thing. *IEEE Transactions on Information Theory*, submitted. `http://www.eecs.berkeley.edu/~sahai/Papers/FocusingBound.pdf`.

[68] Anant Sahai, Stark Draper, and Michael Gastpar. Boosting reliability over awgn networks with average power constraints and noiseless feedback. *ISIT*, 2005.

[69] Anant Sahai and Sanjoy K. Mitter. The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link: Part I: scalar systems. *IEEE Transactions on Information Theory*, 52(8):3369 – 3395, August 2006.

[70] J.P.M. Schalkwijk and Thomas Kailath. A coding scheme for additive noise channels with feedback–I: No bandwidth constraint. *IEEE Transactions on Information Theory*, 12(2):172–182, 166.

[71] Dan Schonberg. *Practical Distributed Source Coding and Its Application to the Compression of Encrypted Data, PhD thesis*. University of California, Berkeley, Berkeley, CA, 2007.

[72] Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27, 1948.

[73] Claude Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4):656–715, 1949.

[74] Claude Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record*, 7(4):142–163, 1959.

[75] Claude Shannon, Robert Gallager, and Elwyn Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels I. *Information and Control*, 10:65–103, 1967.

[76] Claude Shannon, Robert Gallager, and Elwyn Berlekamp. Lower bounds to error probability for coding on discrete memoryless channels II. *Information and Control*, 10:522–552, 1967.

[77] Paul Shields and David Neuhoff. Block and sliding-block source coding. *IEEE Transactions on Information Theory*, 23:211 – 215, 1977.

[78] Wang Shuo. *I'm your daddy*. Tianjin People's Press, 2007.

[79] David Slepian and Jack Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19, 1973.

[80] David Tse. *Variable-rate Lossy Compression and its Effects on Communication Networks, PhD thesis*. Massachusetts Institute Technology, 1994.

[81] David Tse, Robert Gallager, and John Tsitsiklis. Optimal buffer control for variable-rate lossy compression. *31st Allerton Conference*, 1993.

[82] Sergio Verdú and Steven W. McLaughlin. *Information Theory: 50 Years of Discovery*. Wiley-IEEE Press, 1999.

[83] Terry Welch. A technique for high-performance data compression. *IEEE Computer*, 17(6):8–19, 1984.

[84] Wikipedia. Biography of Wang Shuo, `http://en.wikipedia.org/wiki/Wang_Shuo`.

[85] Hirosuke Yamamoto and Kohji Itoh. Asymptotic performance of a modified schalkwijk-barron scheme for channels with noiseless feedback. *IEEE Transactions on Information Theory*, 25(6):729733, 1979.

[86] Zhusheng Zhang. *Analysis— a new perspective*. Peking University Press, 1995.

[87] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23(3):337–343, 1977.

[88] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.

Appendix A

Review of fixed-length block source coding

We review the classical fixed-length block source coding results in this Chapter.

A.1 Lossless Source Coding and Error Exponent

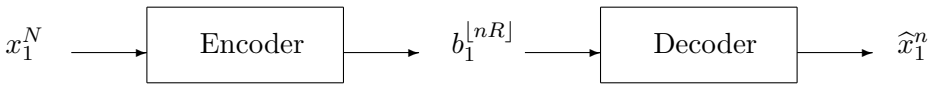

Figure A.1. Block lossless source coding

Consider a discrete memoryless iid source with distribution p_x defined on finite alphabet $\mathcal{X}$. A rate R block source coding system for n source symbols consists of an encoder-decoder pair $(\mathcal{E}_n, \mathcal{D}_n)$, as shown in Figure A.1, where

$$\begin{aligned} \mathcal{E}_n &: \mathcal{X}^n \longrightarrow \{0,1\}^{\lfloor nR \rfloor}, \qquad \mathcal{E}_n(x_1^n) = b_1^{\lfloor nR \rfloor} \\ \mathcal{D}_n &: \{0,1\}^{\lfloor nR \rfloor} \longrightarrow \mathcal{X}^n, \qquad \mathcal{D}_n(b_1^{\lfloor nR \rfloor}) = \widehat{x}_1^n \end{aligned}$$

The probability of block decoding error is $\Pr[x_1^n \neq \widehat{x}_1^n] = \Pr[x_1^n \neq \mathcal{D}_n(\mathcal{E}_n(x_1^n))]$.

In his seminal paper [72], Shannon proved that arbitrarily small error probabilities are achievable by letting n get big as long as the encoder rate is larger than the entropy of the source, $R > H(p_x)$, where $H(p_x) = \sum_{x \in \mathcal{X}} -p_x(x) \log p_x(x)$. Furthermore, it turns out that the error probability goes to zero exponentially in n.

Theorem 9 *(From [29]) For a discrete memoryless source $x \sim p_x$ and encoder rate $R < \log |\mathcal{X}|$,*

$\forall \epsilon > 0$, $\exists K < \infty$, s.t. $\forall n \geq 0$, $\exists$ a block encoder-decoder pair $\mathcal{E}_n, \mathcal{D}_n$ such that

$$\Pr[x_1^n \neq \widehat{x}_1^n] \leq K 2^{-n(E_{s,b}(R) - \epsilon)} \tag{A.1}$$

This result is asymptotically tight, in the sense that for any sequence of encoder-decoder pairs $\mathcal{E}_n, \mathcal{D}_n$,

$$\limsup_{n \to \infty} -\frac{1}{n} \log \Pr[x_1^n \neq \widehat{x}_1^n] = \limsup_{n \to \infty} -\frac{1}{n} \log \Pr[x_1^n \neq \mathcal{D}_n(\mathcal{E}_n(x_1^n))] \leq E_{s,b}(R) \tag{A.2}$$

where $E_{s,b}(R)$ is defined as the block source coding error exponent with the form:

$$E_{s,b}(R) = \min_{q: H(q) \geq R} D(q \| p_x) \tag{A.3}$$

Paralleling the definition of the Gallager function for channel coding [41], as mentioned as an exercise in [29]:

$$\begin{aligned} E_{s,b}(R) &= \sup_{\rho \geq 0} \{\rho R - E_0(\rho)\} \\ \text{where} \quad E_0(\rho) &= (1+\rho) \log[\sum_x p_x(x)^{\frac{1}{1+\rho}}] \end{aligned} \tag{A.4}$$

In [29], it is shown that if the encoder randomly assigns a *bin* number in $\{1, 2, ..., 2^{\lfloor nR \rfloor}\}$ with equal probability to the source sequence and the decoder perform a maximum likelihood or minimum empirical entropy decoding rule for the source sequences in the same *bin*, the random coding error exponent for block source coding is:

$$\begin{aligned} E_r(R) &= \min_q \{D(q \| p_x) + |R - H(q)|^+\} \\ &= \sup_{\rho \in [0,1]} \{\rho R - (1+\rho) \log[\sum_x p_x(x)^{\frac{1}{1+\rho}}]\} \end{aligned} \tag{A.5}$$

where $|t|^+ = \max\{0, t\}$. This error exponent is the same as the block coding error exponent (A.3) in the low rate regime and strictly lower than the block coding error exponent in the

high rate regime [29]. In Section 2.3.3, we show that this random coding error exponent is also achievable in the delay constrained source coding setup for streaming data. However, in Section 2.5.1 we will show that the random coding error exponent and the block coding error exponent are suboptimal everywhere for $R \in (H(p_x), \log|\mathcal{X}|)$ in the delay constrained setup.

A.2 Lossy source coding

Now consider a discrete memoryless iid source with distribution p_x defined on $\mathcal{X}$. A rate R block lossy source coding system for N source symbols consists of an encoder-decoder pair $(\mathcal{E}_N, \mathcal{D}_N)$, as shown in Figure A.2, where

$$\mathcal{E}_n : \mathcal{X}^n \longrightarrow \{0,1\}^{\lfloor nR \rfloor}, \qquad \mathcal{E}_n(x_1^n) = b_1^{\lfloor nR \rfloor}$$
$$\mathcal{D}_n : \{0,1\}^{\lfloor nR \rfloor} \longrightarrow \mathcal{Y}^n, \qquad \mathcal{D}_n(b_1^{\lfloor nR \rfloor}) = y_1^n$$

Instead of attempting to estimate the source symbols x_1^N exactly as in Chapter 2, in lossy source coding, the design goal of the system is to reconstruct the source symbols x_1^N within an average distortion $D > 0$.

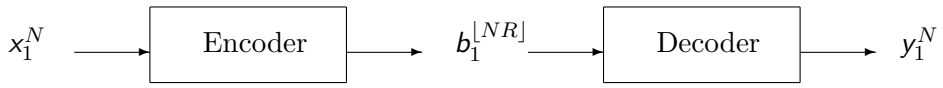

Figure A.2. Block lossy source coding

A.2.1 Rate distortion function and error exponent for block coding under average distortion

The key issue of block lossy source coding is to determine the minimum rate R such that the distortion measure $d(x_1^N, y_1^N) \leq D$ is satisfied with probability close to 1. For the average distortion, we denote by the distortion between two sequence as the average of the distortions of each individual symbols:

$$d(x_1^N, y_1^N) = \frac{1}{N} \sum_{i=1}^{N} d(x_i, y_i)$$

In [74], Shannon first proved the following *rate distortion* theorem:

Theorem 10 *The rate-distortion function $R(D)$ for average distortion measures:*

$$R(p_x, D) \triangleq \min_{W \in \mathcal{W}_D} I(p_x, W) \tag{A.6}$$

where $\mathcal{W}_D$ is the set of all transition matrices that satisfy the average distortion constraint, i.e.

$$\mathcal{W}_D = \{W : \sum_{x,y} p_x(x) W(y|x) d(x,y) \leq D\}.$$

Operationally, this lemma says, for any $\epsilon > 0$ and $\delta > 0$, for block length N big enough, there exists a code of rate $R(p_x, D) + \epsilon$, such that the average distortion between the source string x_1^N and its reconstruction y_1^N is no bigger than D with probability at least $1 - \delta$.

This problem is only interesting if the target average distortion D is higher than $\underline{D}$ and lower than $\overline{D}$, where

$$\underline{D} \triangleq \sum_{x \in \mathcal{X}} p_x(x) \min_{y \in \mathcal{Y}} d(x,y)$$

$$\overline{D} \triangleq \min_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_x(x) d(x,y).$$

The reasoning for the above statement should be trivial. Interested readers may read [6].

For distortion constraint $D > \underline{D}$, we have the following fixed-to-variable length coding result for average distortion measure. To have $\Pr[d(x_1^N, y_1^N) > D] = 0$, we can implement a universal variable length prefix-free code with code length $l_D(x_1^N)$ where

$$l_D(x_1^N) = n(R(p_{x_1^N}, D) + \delta_N) \tag{A.7}$$

where $p_{x_1^N}$ is the empirical distribution of x_1^N, and δ_N goes to 0 as N goes to infinity.

This is a simple corollary of the type covering lemma [29, 28] which is derived from the Johnson− Stein−Lovász theorem [23].

It is widely know that $R(p_x, D)$ is convex $\cup$ on D for fixed p_x [5]. However, the rate distortion function $R(p_x, D)$ for average distortion measure is in general non-concave, non-$\cap$, in the source distribution p_x for fixed distortion constraint D as pointed out in [55].

Now we present the large deviation properties of lossy source coding under average distortion measures from [55],

Theorem 11 *Block coding error exponent under average distortion:*

$$\liminf_{n\to\infty} -\frac{1}{N}\log_2 \Pr[d(x_1^N, y_1^N) > D] = E_D^{b,average}(R)$$
$$\text{where } E_D^{b,average}(R) = \min_{q_x: R(q_x,D)>R} D(q_x\|p_x) \tag{A.8}$$

where y_1^N is the reconstruction of x_1^N using an optimal rate R code.

A.3 Block distributed source coding and error exponents

In the classic block-coding Slepian-Wolf paradigm [79, 29, 39] (illustrated in Figure 4.1), full length-N vectors[1] x^N and y^N are observed by their respective encoders before communication starts. In the block coding setup, a rate-(R_x, R_y) length-N block source code consists of an encoder-decoder triplet $(\mathcal{E}_N^x, \mathcal{E}_N^y, \mathcal{D}_N)$, as we will define shortly in Definition 13.

Definition 13 *A randomized length-N rate-(R_x, R_y) block encoder-decoder triplet $(\mathcal{E}_N^x, \mathcal{E}_N^y, \mathcal{D}_N)$ is a set of maps* [2]

$$\begin{aligned}
\mathcal{E}_N^x &: \mathcal{X}^N \to \{0,1\}^{NR_x}, && \text{e.g.,}\quad \mathcal{E}_N^x(x^N) = a^{NR_x}\\
\mathcal{E}_N^y &: \mathcal{Y}^N \to \{0,1\}^{NR_y}, && \text{e.g.,}\quad \mathcal{E}_N^y(y^N) = b^{NR_y}\\
\mathcal{D}_N &: \{0,1\}^{NR_x}\times\{0,1\}^{NR_y} \to \mathcal{X}^N\times\mathcal{Y}^N, && \text{e.g.,}\quad \mathcal{D}_N(a^{NR_x}, b^{NR_y}) = (\widehat{x}^N, \widehat{y}^N)
\end{aligned}$$

where common randomness, similar to the point-to-point source coding case in Section 2.3.3, shared between the encoders and the decoder is assumed. This allows us to randomize the mappings independently of the source sequences.

The error probability typically considered in Slepian-Wolf coding is the joint error probability, $\Pr[(x^N, y^N) \neq (\widehat{x}^N, \widehat{y}^N)] = \Pr[(x^N, y^N) \neq \mathcal{D}_N(\mathcal{E}_N^x(x^N), \mathcal{E}_N^y(y^N))]$. This probability is taken over the random source vectors as well as the randomized mappings. An error exponent E is said to be achievable if there exists a family of rate-(R_x, R_y) encoders and decoders $\{(\mathcal{E}_N^x, \mathcal{E}_N^y, \mathcal{D}_N)\}$, indexed by N, such that

$$\lim_{N\to\infty} -\frac{1}{N}\log \Pr[(x^N, y^N) \neq (\widehat{x}^N, \widehat{y}^N)] \geq E. \tag{A.9}$$

[1] In the block coding part, we simply use x^N instead of x_1^N to denote the sequence of random variables $x_1, ..., x_N$.

[2] For the sake of simplicity, we assume NR_x and NR_y are integers. It should be clear that this assumption is insignificant in the asymptotic regime where N is big and the integer effect can be ignored.

In this thesis, we study random source vectors $(\mathsf{x}^N, \mathsf{y}^N)$ that are iid across time but may have dependencies at any given time:

$$p_{\mathsf{xy}}(x^N, y^N) = \prod_{i=1}^{N} p_{\mathsf{xy}}(x_i, y_i).$$

Without loss of generality, we assume the marginal distributions are non zero, i.e. $p_{\mathsf{x}}(x) > 0$ for all $x \in \mathcal{X}$, and $p_{\mathsf{y}}(y) > 0$ for all $y \in \mathcal{Y}$,

For such iid sources, upper and lower bounds on the achievable error exponents are derived in [39, 29]. These results are summarized by the following Lemma.

Theorem 12 *(Lower bound) Given a rate pair (R_x, R_y) such that $R_x > H(\mathsf{x}|\mathsf{y})$, $R_y > H(\mathsf{y}|\mathsf{x})$, $R_x + R_y > H(\mathsf{x},\mathsf{y})$. Then, for all*

$$E < \min_{\bar{\mathsf{x}},\bar{\mathsf{y}}} D(p_{\bar{\mathsf{x}}\bar{\mathsf{y}}} \| p_{\mathsf{xy}}) + \left| \min[R_x + R_y - H(\bar{\mathsf{x}},\bar{\mathsf{y}}), R_x - H(\bar{\mathsf{x}}|\bar{\mathsf{y}}), R_y - H(\bar{\mathsf{y}}|\bar{\mathsf{x}})] \right|^+ \quad \text{(A.10)}$$

there exists a family of randomized encoder-decoder mappings as defined in Definition 13 such that (A.9) is satisfied. In (A.10) the function $|z|^+ = z$ if $z \geq 0$ and $|z|^+ = 0$ if $z < 0$.

(Upper bound) Given a rate pair (R_x, R_y) such that $R_x > H(\mathsf{x}|\mathsf{y})$, $R_y > H(\mathsf{y}|\mathsf{x})$, $R_x + R_y > H(\mathsf{x},\mathsf{y})$. Then, for all

$$E > \min \left\{ \min_{\bar{\mathsf{x}},\bar{\mathsf{y}}: R_x < H(\bar{\mathsf{x}}|\bar{\mathsf{y}})} D(p_{\bar{\mathsf{x}}\bar{\mathsf{y}}} \| p_{\mathsf{xy}}), \min_{\bar{\mathsf{x}},\bar{\mathsf{y}}: R_y < H(\bar{\mathsf{y}}|\bar{\mathsf{x}})} D(p_{\bar{\mathsf{x}}\bar{\mathsf{y}}} \| p_{\mathsf{xy}}), \min_{\bar{\mathsf{x}},\bar{\mathsf{y}}: R_x + R_y < H(\bar{\mathsf{x}},\bar{\mathsf{y}})} D(p_{\bar{\mathsf{x}}\bar{\mathsf{y}}} \| p_{\mathsf{xy}}) \right\} \quad \text{(A.11)}$$

there does not exists a randomized encoder-decoder mapping as defined in Definition 13 such that (A.9) is satisfied.

In both bounds $(\bar{\mathsf{x}}, \bar{\mathsf{y}})$ are arbitrary random variables with joint distribution $p_{\bar{\mathsf{x}}\bar{\mathsf{y}}}$.

Remark: As long as (R_x, R_y) is in the interior of the achievable region, i.e., $R_x > H(\mathsf{x}|\mathsf{y})$, $R_y > H(\mathsf{y}|\mathsf{x})$ and $R_x + R_y > H(\mathsf{x},\mathsf{y})$ then the lower-bound (A.10) is positive. The achievable region is illustrated in Fig A.3. As shown in [29], the upper and lower bounds (A.11) and (A.10) match when the rate pair (R_x, R_y) is achievable and close to the boundary of the region. This is analogous to the high rate regime in channel coding, or the low rate regime in source coding, where the random coding bound (analogous to (A.10)) and the sphere packing bound (analogous to (A.11)) agree.

Theorem 12 can also be used to generate bounds on the exponent for source coding with decoder side-information (i.e., y observed at the decoder), and for source coding without side information (i.e., y is independent of x and the decoder only needs to decode x).

Corollary 3 *(Source coding with decoder side-information) Consider a Slepian-Wolf problem where y is known by the decoder. Given a rate R_x such that $R_x > H(\mathsf{x}|\mathsf{y})$, then for all*

$$E < \min_{\bar{x},\bar{y}} D(p_{\bar{x}\bar{y}} \| p_{\mathsf{x}\mathsf{y}}) + |R_x - H(\bar{x}|\bar{y})|^+, \tag{A.12}$$

there exists a family of randomized encoder-decoder mappings as defined in Definition 13 such that (A.9) is satisfied.

The proof of Corollary 3 follows from Theorem 12 by letting R_y be sufficiently large ($> \log|\mathcal{X}|$). Similarly, by letting y be independent of x so that $H(\mathsf{x}|\mathsf{y}) = H(\mathsf{x})$, we get the following random-coding bound for the point-to-point case of a single source x which is the random coding part of Theorem 9 in Section A.1.

Corollary 4 *(point-to-point) Consider a Slepian-Wolf problem where y is independent of x, Given a rate R_x such that $R_x > H(\mathsf{x})$, for all*

$$E < \min_{\bar{x}} D(p_{\bar{x}} \| p_{\mathsf{x}}) + |R_x - H(\bar{x})|^+ = E_r(R_x) \tag{A.13}$$

there exists a family of randomized encoder-decoder triplet as defined in Definition 13 such that (A.9) is satisfied.

A.4 Review of block source coding with side-information

As shown in Figure 5.1, the sources are iid random variables $\mathsf{x}_1^n, \mathsf{y}_1^n$ from a finite alphabet $\mathcal{X} \times \mathcal{Y}$ with distribution $p_{\mathsf{x}\mathsf{y}}$. Without loss of generality, we assume that $p_{\mathsf{x}}(x) > 0$, $\forall x \in \mathcal{X}$ and $p_{\mathsf{y}}(y) > 0$, $\forall y \in \mathcal{Y}$. x_1^n is the source known to the encoder and y_1^n is the side-information known only to the decoder. A rate R block source coding system for n source symbols consists of an encoder-decoder pair $(\mathcal{E}_n, \mathcal{D}_n)$. Where

$$\begin{aligned} &\mathcal{E}_n : \mathcal{X}^n \to \{0,1\}^{\lfloor nR \rfloor}, && \mathcal{E}_n(x_1^n) = b_1^{\lfloor nR \rfloor} \\ &\mathcal{D}_n : \{0,1\}^{\lfloor nR \rfloor} \times \mathcal{Y}^n \to \mathcal{X}^n, && \mathcal{D}_n(b_1^{\lfloor nR \rfloor}, y_1^n) = \hat{x}_1^n \end{aligned}$$

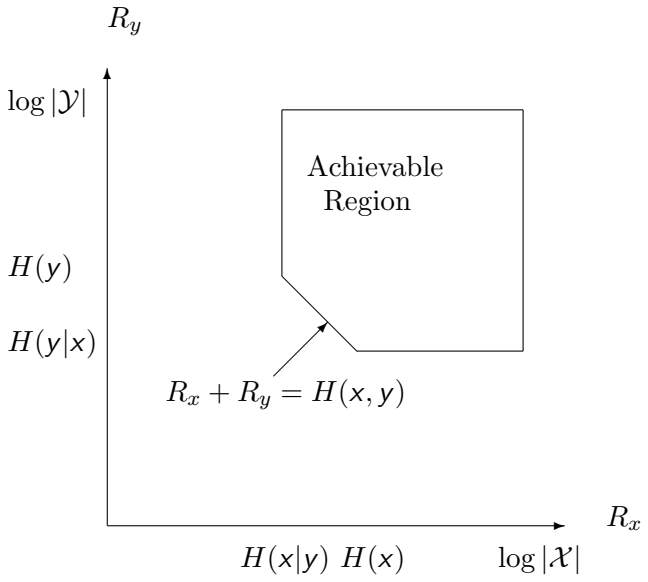

Figure A.3. Achievable region for Slepian-Wolf source coding

The error probability is $\Pr[x_1^n \neq \hat{x}_1^n] = \Pr[x_1^n \neq \mathcal{D}_n(\mathcal{E}_n(x_1^n), y_1^n)]$. The exponent $E_{si,b}(R)$ is achievable if $\exists$ a family of $\{(\mathcal{E}_n, \mathcal{D}_n)\}$, s.t.

$$\lim_{n\to\infty} -\frac{1}{n}\log_2 \Pr[x_1^n \neq \hat{x}_1^n] = E_{si,b}(R) \tag{A.14}$$

The relevant results of [29, 39] are summarized into the following theorem.

Theorem 13 *$E_{si,b}^{lower}(R) \leq E_{si,b}(R) \leq E_{si,b}^{upper}(R)$ where*

$$\begin{aligned} E_{si,b}^{lower}(R) &= \min_{q_{xy}}\{D(q_{xy}\|p_{xy}) + |0, R - H(q_{x|y})|^+\} \\ E_{si,b}^{upper}(R) &= \min_{q_{xy}:H(q_{x|y})\geq R}\{D(q_{xy}\|p_{xy})\} \end{aligned}$$

Where these two error exponents can also be expressed in a parameterized way:

$$\begin{aligned} E_{si,b}^{lower}(R) &= \max_{\rho\in[0,1]} \rho R - \log[\sum_y [\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}]^{1+\rho}] \\ E_{si,b}^{upper}(R) &= \max_{\rho\in[0,\infty]} \rho R - \log[\sum_y [\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}]^{1+\rho}] \end{aligned}$$

It should be clear that both of these bounds are only positive if $R > H(p_{x|y})$. As shown in [29], the two bounds are the same in the low rate regime. Furthermore, with *encoder* side-information, similar to the case in Theorem 9, the block coding error exponent is $E_{si,b}^{upper}(R)$. Thus in the low rate regime, the error exponents are the same with or without encoder side-information. Due to the duality between channel coding and source coding with decoder side-information, in the high rate regime one can derive an expurgated bound, as done by Gallager [41] for channel coding, to tighten the random coding error exponent. Interested readers may read Ahlswede's very classical paper [2] and some recent papers [22, 64].

Appendix B

Tilted Entropy Coding and its delay constrained performance

In this appendix, we introduce another way to achieve the delay constrained source coding error exponent in Theorem 1. This coding scheme is identical to our universal coding scheme introduced in Section 2.4.1, except that we use a non-universal fixed-to-variable length prefix-free source code based on *tilted* entropy code. The rest of the system is identical to that in Figure 2.11. This derivation first appeared in our paper [20] which is a minor variation of the scheme analyzed in [49]. Just as in the universal coding scheme in Section 2.4.1 and [49], the queuing behavior is what determines the delay constrained error exponent. Rather than modeling the buffer as a random walk with and analyze the stationary distribution of the process as in [49], here we give an alternate derivation by applying Cramér's theorem.

B.1 Tilted entropy coding

We replace the universal optimal source code with the following tilted entropy code first introduced by Jelinek [49]. This code is a Shannon-code[1] built for a particular tilted distribution for p_x.

Definition 14 *The λ-tilted entropy code is an instantaneous code $\mathcal{C}_{N_\lambda}$ for $\lambda > -1$. It is a*

[1]Code length proportional to the logarithm of the inverse of probability.

mapping from $\mathcal{X}^N$ to a variable number of binary bits.

$$\mathcal{C}_{N_\lambda}(x_1^N) = b_1^{l(x_1^N)}$$

where $l(x_1^N)$ is the codeword length for source sequence x_1^N. The first bit is always 1 and the rest of the codewords are the Shannon codewords based on the λ tilted distribution of p_x

$$\begin{aligned} l(x_1^N) &= 1 + \lceil -\log_2 \frac{p_x(x_1^N)^{\frac{1}{1+\lambda}}}{\sum_{s_1^N \in \mathcal{X}^N} p_x(s_1^N)^{\frac{1}{1+\lambda}}} \rceil \\ &\leq 2 - \sum_{i=1}^{N} \log_2 \frac{p_x(x_i)^{\frac{1}{1+\lambda}}}{\sum_{s\in\mathcal{X}} p_x(s)^{\frac{1}{1+\lambda}}} \end{aligned} \tag{B.1}$$

From the definition of $l(\vec{x})$, we have the following fact:

$$\begin{aligned} \log_2(\sum_{\vec{x}\in\mathcal{X}^N} p_x(\vec{x}) 2^{\lambda l(\vec{x})})\} &\leq 2\lambda + N\log_2[(\sum_x p_x(x)^{1-\frac{\lambda}{1+\lambda}})(\sum_x p_x(x)^{\frac{1}{1+\lambda}})^\lambda] \\ &= 2\lambda + N(1+\lambda)\log_2[\sum_x p_x(x)^{\frac{1}{1+\lambda}}] \\ &= 2\lambda + NE_0(\lambda) \end{aligned} \tag{B.2}$$

This definition is valid for any $\lambda > -1$. As will be seen later in the proof, for a rate R delay constrained source coding system, the optimal λ is $\lambda = \rho^*$, where ρ^* is defined in (2.57):

$$\rho^* R = E_0(\rho^*)$$

From (B.1), the longest code length l_N^λ is

$$l_N^\lambda \leq 2 - N\log_2 \frac{p_{x_{\min}}^{\frac{1}{1+\lambda}}}{\sum_{s\in\mathcal{X}} p_x(s)^{\frac{1}{1+\lambda}}} \tag{B.3}$$

Where $p_{x_{\min}} = \min_{x\in\mathcal{X}} p_x(x)$. The constant 2 is insignificant compared to N.

This variable-length code is turned into a fixed rate R code as follows that in Section 2.4.1. The delay constrained source coding scheme is illustrated in Figure 2.11. At time kN, $k = 1, 2, ...$ the encoder $\mathcal{E}$ uses the variable length code $\mathcal{C}_{N_\lambda}$ to encode the k^{th} source block $\vec{x}_k = x_{(k-1)N+1}^{kN}$ into a binary sequence $b(k)_1, b(k)_2, ...b(k)_{l(\vec{x}_k)}$. This codeword is pushed into a FIFO queue with infinite buffer-size. The encoder drains a bit from the queue every $\frac{1}{R}$ seconds. If the queue is empty, the encoder simply sends 0's to the decoder until there are new bits pushed in the queue.

The decoder knows the variable length code book and the prefix-free nature[2] of the Shannon code guarantees that everything can be decoded correctly.

B.2 Error Events

The number of the bits Ω_k in the encoder buffer at time kN, is a random walk process with negative drift and a reflecting barrier. At time $(k+d)N$, where $dN = \Delta$, the decoder can make an error in estimating $\vec{x}_k$ if and only if part of the variable length code for source block $\vec{x}_k$ is still in the encoder buffer.

Since the FIFO queue drains deterministically, it means that when the k-th block's codeword entered the queue, it was already doomed to miss its deadline of $dN = \Delta$. Formally, for an error to occur, the number of bits in the buffer $\Omega_k \geq \lfloor dNR = \Delta R \rfloor$. Thus, meeting a specific end-to-end latency constraint over a fixed-rate noiseless link is like the buffer overflow events analyzed in [49]. Define the random time $t_k N$ to be the last time before time kN when the queue was empty. A missed-deadline occurs only if $\sum_{i=t_k+1}^{k} l(\vec{x}_i) > (d+k-t_k)NR = (kN - t_k N + \Delta)R$.

For arbitrary $1 \leq t \leq k-1$, define the error event

$$P_N^{k,d}(t) = P\left(\sum_{i=t}^{k} l(\vec{x}_i) > (d+k-t)NR\right) = P\left(\sum_{i=t}^{k} l(\vec{x}_i) > (kN - t_k N + \Delta)R\right).$$

Using Cramér theorem[31], we derive a tight, in the large deviation sense, upper bound on $P_N^{k,d}(t)$.

As a simple corollary of the Cramér theorem, we have an upper bound on $P_N^{k,d}(t)$.

$$\begin{aligned} P_N^{k,d}(t) &= P\left(\sum_{i=t+1}^{k} l(\vec{x}_i) \geq (d+k-t)NR\right) \\ &= P\left(\frac{1}{k-t}\sum_{i=t+1}^{k} l(\vec{x}_i) \geq \frac{(d+k-t)NR}{k-t}\right) \\ &\leq (k-t)^{|\mathcal{X}|^N} 2^{-E_N(\mathcal{C}_{N_\lambda}, R, k-t, d)} \end{aligned}$$

Where by using the fact in (B.2) and noticing that the 2λ is insignificant in (B.2), we have

[2]The initial 1 is not really required since the decoder knows the rate at which source-symbols are arriving at the encoder. Thus, it knows when the queue is empty and does not need to even interpret the 0s it receives.

the large deviation exponent:

$$\begin{aligned} E_N(\mathcal{C}_{N_\lambda}, R, k-t, d) &\geq (k-t) \sup_{\rho \in \mathcal{R}^+} \{\rho \frac{(d+k-t)NR}{k-t} - \log_2(\sum_{\vec{x} \in \mathcal{X}^N} p_x(\vec{x}) 2^{\rho l(\vec{x})})\} \\ &\geq (k-t)[\lambda \frac{(d+k-t)NR}{k-t} - \log_2(\sum_{\vec{x} \in \mathcal{X}^N} p_x(\vec{x}) 2^{\lambda l(\vec{x})})] \\ &\geq (k-t)N(\lambda \frac{(d+k-t)R}{k-t} - \frac{2\lambda}{N} - E_0(\lambda)) \\ &= dN\lambda R + (k-t)N[\lambda R - E_0(\lambda) - \frac{2\lambda}{N}] \end{aligned} \tag{B.4}$$

B.3 Achievability of delay constrained error exponent $E_s(R)$

We only need to show that for any $\epsilon > 0$, by appropriate choice of N, λ, it is possible to achieve an error exponent with delay of $E_s(R) - \epsilon$, i.e. for all i, Δ,

$$\Pr[x_j \neq \widehat{x}_j(j+\Delta)] \leq K 2^{-\Delta(E_s(R)-\epsilon)}$$

From (2.57), we know that $E_s(R) = E_0(\rho^*)$, where $\rho^* R = E_0(\rho^*)$.

Proof: For the tilted entropy source coding scheme of $\mathcal{C}_{N,\lambda}$, the decoding error for the k^{th} source block at time $(k+d)N$ is[3] $P_N^{k,d}$. $t_k N$ is the last time before k when the buffer is empty.

$$\begin{aligned} \Pr[x_j \neq \widehat{x}_j(j+\Delta)] &\leq \sum_{t=0}^{k} P_N^{k,d}(t) \\ &= \sum_{t=0}^{k-1} P\left(t_k = t, \sum_{i=t+1}^{k} l(\vec{x}_i) \geq (d+k-t)NR\right) \\ &\leq \sum_{t=0}^{k-1} P\left(\sum_{i=t+1}^{k} l(\vec{x}_i) \geq (d+k-t)NR\right) \\ &\leq \sum_{t=0}^{k-1} (k-t+1)^{|\mathcal{X}|^N} 2^{-E_N(\mathcal{C}_{N_\lambda}, R, k-t, d)} \\ &= 2^{-dN\lambda R} \sum_{t=0}^{k-1} (k-t+1)^{|\mathcal{X}|^N} 2^{-(k-t)N[\lambda R - E_0(\lambda) - \frac{\lambda}{N}]} \end{aligned}$$

The above equality is true for all N, λ. Pick $\lambda = \rho^* - \frac{\epsilon}{R}$. From Figure 2.12, we know that $\lambda R - E_0(\lambda) > 0$. Choose $N > \frac{\lambda}{\lambda R - E_0(\lambda))}$. Then define

$$K(\epsilon, R, N) = \sum_{i=0}^{\infty} (i+1)^{|\mathcal{X}|^N} 2^{-iN[\lambda R - E_0(\lambda) - \frac{\lambda}{N}]} < \infty$$

[3]Here we denote j by kN, and Δ by dN.

which is guaranteed to be finite since dying exponentials dominate polynomials in sums. Thus:

$$\Pr[x_j \neq \widehat{x}_j(j+\Delta)] \leq K(\epsilon)2^{-dN\lambda R} = K(\epsilon)2^{-\Delta\lambda R} = K(\epsilon, R, N)2^{-\Delta(\rho^* R - \epsilon)}$$

where the constant $K(\epsilon, R)$ does not depend on the delay in question. Since $E_0(\rho^*) = \rho^* R$ and the encoder also is not targeted to the delay Δ, this scheme achieves the desired delay constrained exponent as promised.

Appendix C

Proof of the concavity of the rate distortion function under the peak distortion measure

In this appendix, we prove Lemma 5 which states that the rate distortion function $R(p, D)$ is concave $\cap$ in p.

Proof: To show that $R(p, D)$ is concave $\cap$ in p, it is enough to show that for any two distributions p_0 and p_1 and for any $\lambda \in [0, 1]$,

$$R(p_\lambda, D) \geq \lambda R(p_0, D) + (1 - \lambda) R(p_1, D)$$

where $p_\lambda = \lambda p_0 + (1 - \lambda) p_1$. Define:

$$W^* = \arg\min_{W \in \mathcal{W}_D} I(p_\lambda, W)$$

From the definition of $R(p, D)$ we know that

$$\begin{aligned} R(p_\lambda, D) &= I(p_\lambda, W^*) \\ &\geq \lambda I(p_0, W^*) + (1 - \lambda) I(p_1, W^*) \qquad \text{(C.1)} \\ &\geq \lambda \min_{W \in \mathcal{W}_D} I(p_0, W) + (1 - \lambda) \min_{W \in \mathcal{W}_D} I(p_1, W) \\ &= \lambda R(p_0, D) + (1 - \lambda) R(p_1, D) \end{aligned}$$

(C.1) is true because $I(p, W)$ is concave $\cap$ in p for fixed W and $p_\lambda = \lambda p_0 + (1 - \lambda) p_1$. The rest are by the definition. $\square$

Appendix D

Derivation of the upper bound on the delay constrained lossy source coding error exponent

In this appendix, we prove the converse of Theorem 2. The proof is similar to that of the converse of the delay constrained lossless source coding error exponent in Theorem 1.

To bound the best possible error exponent with fixed delay, we consider a block coding encoder/decoder pair that is constructed by the delay constrained encoder/decoder pair and translate the block-coding error exponent for peak distortion in Lemma 6 to the delay constrained error exponent. The arguments are analogous to the "focusing bound" derivation in [67] for channel coding with feedback and extremely similar to that of the lossless source coding case in Theorem 1. We summarize the converse of Theorem 2 in the following proposition.

Proposition 12 *For fixed-rate encodings of discrete memoryless sources, it is not possible to achieve an lossy source coding error exponent with fixed-delay higher than*

$$\inf_{\alpha>0} \frac{1}{\alpha} E_D^b((\alpha+1)R) \tag{D.1}$$

from the definition of delay constrained lossy source coding error exponent in Definition 4, the statement of this proposition is equivalent to the following statement:

For any $E > \inf_{\alpha>0} \frac{1}{\alpha} E_D^b((\alpha+1)R)$*, there exists an positive* ϵ*, such that for any* $K < \infty$*, there exists* $i > 0$*,* $\Delta > 0$ *and*

$$\Pr[d(x_i, y_i(i+\Delta)) \geq D] > K 2^{-\Delta(E_D(R)-\epsilon)}$$

Proof: We show the proposition by contradiction. Suppose that the delay-constrained error exponent can be higher than $\inf_{\alpha>0} \frac{1}{\alpha} E_D^b((\alpha+1)R)$. Then according to Definition 4, there exists a delay-constrained source coding system, such that for some $E > \inf_{\alpha>0} \frac{1}{\alpha} E_D^b((\alpha+1)R)$, for any positive real value ϵ, there exists $K < \infty$, such that for all $i > 0$, $\Delta > 0$

$$\Pr[d(x_i, y_i(i+\Delta)) > D] \leq K 2^{-\Delta(E-\epsilon)}$$

so we choose $\epsilon > 0$, such that

$$E - \epsilon > \inf_{\alpha>0} \frac{1}{\alpha} E_D^b((\alpha+1)R) \tag{D.2}$$

Then consider a block coding scheme $(\mathcal{E}, \mathcal{D})$ that is built on the delay constrained lossy source coding system. The encoder of the block coding system is the *same* as the delay-constrained lossy source encoder, and the block decoder $\mathcal{D}$ works as follows:

$$y_1^i = (y_1(1+\Delta), y_2(2+\Delta), ..., y_i(i+\Delta))$$

Now the block decoding distortion of this coding system can be upper bounded as follows, for any $i > 0$ and $\Delta > 0$:

$$\begin{aligned}
\Pr[d(x_1^i, y_1^i) > D] &= \sum_{t=1}^{i} \Pr[\max_i d(x_i, y_i) > D] \\
&\leq \sum_{t=1}^{i} \Pr[d(x_i, y_i(i+\Delta)) > D] \\
&\leq \sum_{t=1}^{i} K 2^{-\Delta(E-\epsilon)} \\
&= iK 2^{-\Delta(E-\epsilon)}
\end{aligned} \tag{D.3}$$

The block coding scheme $(\mathcal{E}, \mathcal{D})$ is a block source coding system for i source symbols by using $\lfloor R(i+\Delta) \rfloor$ bits hence has a rate $\frac{\lfloor R(i+\Delta) \rfloor}{i} \leq \frac{i+\Delta}{i} R$. From the block coding result in Lemma 6, we know that the lossy source coding error exponent $E_D^b(R)$ is monotonically increasing in R, so $E_D^b(\frac{\lfloor R(i+\Delta) \rfloor}{i}) \leq E_D^b(\frac{R(i+\Delta)}{i})$. Again from Lemma 6, we know that the block coding error probability can be bounded in the following way:

$$\Pr[d(x_1^i, y_1^i) > D] > 2^{-i(E_D^b(\frac{R(i+\Delta)}{i})+\epsilon_i)} \tag{D.4}$$

where $\lim_{i\to\infty} \epsilon_i = 0$.

Combining (D.3) and (D.4), we have:

$$2^{-i(E_D^b(\frac{R(i+\Delta)}{i})+\epsilon_i)} < iK2^{-\Delta(E-\epsilon)}$$

Now let $\alpha = \frac{\Delta}{i}$, $\alpha > 0$, then the above inequality becomes:

$$2^{-i(E_D^b(R(1+\alpha))+\epsilon_i)} < 2^{-i\alpha(E-\epsilon-\theta_i)} \text{ and hence:}$$

$$E - \epsilon < \frac{1}{\alpha}(E_D^b(R(1+\alpha)+\epsilon_i) + \theta_i \tag{D.5}$$

where $\theta_i = \frac{\log K}{i}$, so $\lim_{i\to\infty} \theta_i = 0$. The above inequality is true for all i, $\alpha > 0$, and $\lim_{i\to\infty} \theta_i = 0$, $\lim_{i\to\infty} \epsilon_i = 0$. Taking all these into account, we have:

$$E - \epsilon \leq \inf_{\alpha>0} \frac{1}{\alpha} E_D^b(R(1+\alpha)) \tag{D.6}$$

Now (D.6) contradicts with the assumption in (D.2), thus the proposition is proved. ■

Appendix E

Bounding source atypicality under a distortion measure

In this appendix, we prove Lemma 7 in Section 3.4.2

Proof: We only need to show the case for $r > R(p_x, D)$. By Cramér's theorem [31], for all $\epsilon_1 > 0$, there exists K_1, such that

$$\begin{aligned}\Pr[\sum_{i=1}^{n} l_D(\vec{x}_i) > nNr] &= \Pr[\frac{1}{n}\sum_{i=1}^{n} l_D(\vec{x}_i) > Nr] \\ &\leq K_1 2^{-n(\inf\limits_{z>Nr} I(z)-\epsilon_1)}\end{aligned}$$

where the rate function $I(z)$ is [31]:

$$I(z) = \sup_{\rho\geq 0}\{\rho z - \log_2(\sum_{(\vec{x}\in\mathcal{X}^N} p_x(\vec{x})2^{\rho l_D(\vec{x})})\} \tag{E.1}$$

It is clear that $I(z)$ is monotonically increasing with z and $I(z)$ is continuous. Thus

$$\inf_{z>Nr} I(z) = I(Nr) \tag{E.2}$$

Using the upper bound on $l_D(\vec{x})$ in (3.6):

$$\begin{aligned}\log_2(\sum_{\vec{x}\in\mathcal{X}^N} p_x(\vec{x})2^{\rho l_D(\vec{x})}) &\leq \log_2(\sum_{q_x\in\mathcal{T}^N} 2^{-ND(q_x\|p_x)}2^{\rho(\delta_N+NR(q_x,D))}) \\ &\leq \log_2(2^{N\epsilon_N}2^{-N\min_{q_x}\{D(q_x\|p_x)-\rho R(q_x,D)-\rho\delta_N\}}) \\ &= N\big(-\min_{q_x}\{D(q_x\|p_x) - \rho R(q_x,D) - \rho\delta_N\} + \epsilon_N\big)\end{aligned}$$

where $\mathcal{T}^N$ is the set of types of $\mathcal{X}^N$, and $2^{N\epsilon_N}$ is the number of types in $\mathcal{X}^N$, $0 < \epsilon_N \leq \frac{|\mathcal{X}|\log_2(N+1)}{N}$, so ϵ_N goes to 0 as N goes to infinity.

Substitute the above inequalities into (E.1):

$$I(Nr) \geq N\big(\sup_{\rho\geq 0}\{\min_{q_x} \rho(r - R(q_x, D) - \delta_N) + D(q_x\|p_x)\} - \epsilon_N\big) \tag{E.3}$$

Next we show that $I(Nr) \geq N(E_D^b(r) + \epsilon'_N)$ where ϵ'_N goes to 0 as N goes to infinity. We show the existence of a saddle point of the function

$$f(q_x, \rho) = \rho(r - R(q_x, D) - \delta_N) + D(q_x\|p_x)$$

Obviously, for fixed q_x, $f(q_x, \rho)$ is a linear function of ρ, thus concave $\cap$. Also for fixed $\rho \geq 0$, $f(q_x, \rho)$ is a convex $\cup$ function of q_x, because both $-R(q_x, D)$ and $D(q_x\|p_x)$ are convex $\cup$ in q_x. Write

$$g(u) = \min_{q_x}\sup_{\rho\geq 0}(f(q_x, \rho) + \rho u)$$

Showing that $g(u)$ is finite around $u = 0$ establishes the existence of the saddle point as shown in Exercise 5.25 [10].

$$\begin{aligned}
\min_{q_x}\sup_{\rho\geq 0} f(q, \rho) + \rho u &=_{(a)} \min_{q_x}\sup_{\rho\geq 0} \rho(r - R(q_x, D) - \delta_N + u) + D(q_x\|p_x) \\
&\leq_{(b)} \min_{q_x: R(q_x,D)\geq r-\delta_N+u} \sup_{\rho\geq 0} \rho(r - R(q_x, D) - \delta_N + u) + D(q_x\|p_x) \\
&\leq_{(c)} \min_{q_x: R(q_x,D)\geq r-\delta_N+u} D(q_x\|p_x) \\
&<_{(d)} \infty
\end{aligned}$$

(a) is by definition. (b) is true because $R(p_x, D) < r < \overline{R}_D$, thus for very small δ_N and u, $R(p_x, D) < r - \delta_N + u < \overline{R}_D$. Thus there exists a distribution q_x, s.t. $R(q_x, D) \geq r - \delta_N + u$. (c) is because $R(q_x, D) \geq r - \delta_N + u$ and $\rho \geq 0$. (d) is true because we might as well assume that $p_x(x) > 0$ for all $x \in \mathcal{X}$, and $r - \delta_N + u < \overline{R}_D$. Thus we proved the existence of the saddle point of $f(q, \rho)$.

$$\sup_{\rho\geq 0}\{\min_{q} f(q, \rho)\} = \min_{q}\{\sup_{\rho\geq 0} f(q, \rho)\} \tag{E.4}$$

Note that if $R(q_x, D) < r - \delta_N$, ρ can be chosen to be arbitrarily large to make $\rho(r - R(q_x, D) - \delta_N) + D(q_x\|p_x)$ arbitrarily large. Thus the q_x to minimize $\sup_{\rho\geq 0} \rho(r - R(q_x, D) - \delta_N) + D(q_x\|p_x)$ satisfies $r - R(q_x, D) - \delta_N \geq 0$. So

$$\begin{aligned}
\min_{q_X}\{\sup_{\rho\geq 0}\rho(r-R(q_X,D)-\delta_N)+D(q_{XY}\|p_{XY})\} &=_{(a)} \min_{q_X:R(q_X,D)\geq r-\delta_N}\sup_{\rho\geq 0} \\
&\qquad \{\rho(r-R(q_X,D)-\delta_N)+D(q_X\|p_X)\} \\
&=_{(b)} \min_{q_X:R(q_X,D)\geq r-\delta_N}\{D(q_X\|p_X)\} \\
&=_{(c)} E_D^b(r-\delta_N) \qquad\qquad (E.5)
\end{aligned}$$

(a) follows from the argument above. (b) is because $r-R(q_X,D)-\delta_N\leq 0$ and $\rho\geq 0$, and hence $\rho=0$ maximizes $\rho(r-R(q_X,D)-\delta_N)$. (c) is by definition in (3.5). By combining (E.3), (E.4) and (E.5), letting N be sufficiently big so that δ_N is sufficiently small, and noticing that $E_D^b(r)$ is continuous in r, we get the desired bound in (3.7). □

Appendix F

Bounding individual error events for distributed source coding

In this appendix, we prove Lemma 9 and Lemma 11. The proofs have some similarities to those for the point-to-point lossless source coding in Propositions 3 and 4.

F.1 ML decoding: Proof of Lemma 9

In this section we give the proof of Lemma 9, this part has the same flavor as the proof of Proposition 4 in Section 2.3.3. The technical tool used here is the standard Chernoff bound argument, or "the ρ thing" in Gallager's book [41]. This technique is perfected by Gallager in the derivation of the error exponents for a series of problems, cf. multiple-access channel in [37], degraded broadcast channel in [40] and for distributed source coding in [39].

The bound depends on whether $l \leq k$ or $l \geq k$. Consider the case for $l \leq k$,

$$\begin{aligned} p_n(l,k) &= \sum_{x_1^n, y_1^n} p_{xy}(x_1^n, y_1^n) \\ &\quad \Pr[\exists\ (\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{B}_x(x_1^n) \times \mathcal{B}_y(y_1^n) \cap \mathcal{F}_n(l,k,x_1^n,y_1^n) \text{ s.t. } p_{xy}(x_1^n,y_1^n) < p_{xy}(\tilde{x}_1^n,\tilde{y}_1^n)] \\ &\leq \sum_{x_1^n, y_1^n} \min\Big[1, \sum_{\substack{(\tilde{x}_1^n, \tilde{y}_1^n) \in \mathcal{F}_n(l,k,x_1^n,y_1^n) \\ p_{xy}(x_1^n,y_1^n) < p_{xy}(\tilde{x}_1^n,\tilde{y}_1^n)}} \Pr[\tilde{x}_1^n \in \mathcal{B}_x(x_1^n), \tilde{y}_1^n \in \mathcal{B}_y(y_1^n)]\Big] p_{xy}(x_1^n, y_1^n) \end{aligned} \tag{F.1}$$

$$\leq \sum_{x_l^n, y_l^n} \min\Big[1, \sum_{\substack{(\tilde{x}_l^n, \tilde{y}_l^n) \text{ s.t. } \tilde{y}^{k-1} = y^{k-1} \\ p_{xy}(x_l^n, y_l^n) < p_{xy}(\tilde{x}_l^n, \tilde{y}_l^n)}} 4 \times 2^{-(n-l+1)R_x - (n-k+1)R_y}\Big] p_{xy}(x_l^n, y_l^n) \tag{F.2}$$

$$\leq 4 \times \sum_{x_l^n, y_l^n} \min\Big[1, \sum_{\tilde{x}_l^n, \tilde{y}_k^n} 2^{-(n-l+1)R_x - (n-k+1)R_y} \mathit{1}[p_{xy}(\tilde{x}_l^{k-1}, y_l^{k-1}) p_{xy}(\tilde{x}_k^n, \tilde{y}_k^n) > p_{xy}(x_l^n, y_l^n)]\Big] p_{xy}(x_l^n, y_l^n)$$

$$\leq 4 \times \sum_{x_l^n, y_l^n} \min\left[1, \sum_{\tilde{x}_l^n, \tilde{y}_k^n} 2^{-(n-l+1)R_x - (n-k+1)R_y} \min\left[1, \frac{p_{xy}(\tilde{x}_l^{k-1}, y_l^{k-1}) p_{xy}(\tilde{x}_k^n, \tilde{y}_k^n)}{p_{xy}(x_l^n, y_l^n)}\right]\right] p_{xy}(x_l^n, y_l^n)$$

$$\leq 4\times \sum_{x_l^n, y_l^n} \left[\sum_{\tilde{x}_l^n, \tilde{y}_k^n} e^{-(n-l+1)R_x - (n-k+1)R_y} \left[\frac{p_{xy}(\tilde{x}_l^{k-1}, y_l^{k-1}) p_{xy}(\tilde{x}_k^n, \tilde{y}_k^n)}{p_{xy}(x_l^n, y_l^n)}\right]^{\frac{1}{1+\rho}}\right]^{\rho} p_{xy}(x_l^n, y_l^n) \tag{F.3}$$

$$= 4 \times 2^{-(n-l+1)\rho R_x - (n-k+1)\rho R_y} \sum_{x_l^n, y_l^n} \left[\sum_{\tilde{x}_l^n, \tilde{y}_k^n} [p_{xy}(\tilde{x}_l^{k-1}, y_l^{k-1}) p_{xy}(\tilde{x}_k^n, \tilde{y}_k^n)]^{\frac{1}{1+\rho}}\right]^{\rho} p_{xy}(x_l^n, y_l^n)^{\frac{1}{1+\rho}}$$

$$= 4 \times 2^{-(n-l+1)\rho R_x - (n-k+1)\rho R_y} \sum_{y_l^{k-1}} \Big[\sum_{x_l^{k-1}} p_{xy}(x_l^{k-1}, y_l^{k-1})^{\frac{1}{1+\rho}}\Big] \Big[\sum_{\tilde{x}_l^{k-1}} p_{xy}(\tilde{x}_l^{k-1}, y_l^{k-1})^{\frac{1}{1+\rho}}\Big]^{\rho} \Big[\sum_{\tilde{x}_k^n, \tilde{y}_k^n} p_{xy}(\tilde{x}_k^n, \tilde{y}_k^n)^{\frac{1}{1+\rho}}\Big]^{\rho} \sum_{x_k^n, y_k^n} p_{xy}(x_k^n, y_k^n)^{\frac{1}{1+\rho}}$$

$$= 4 \times 2^{-(n-l+1)\rho R_x - (n-k+1)\rho R_y} \left[\sum_{y_l^{k-1}} \Big[\sum_{x_l^{k-1}} p_{xy}(x_l^{k-1}, y_l^{k-1})^{\frac{1}{1+\rho}}\Big]^{1+\rho}\right] \Big[\sum_{x_k^n, y_k^n} p_{xy}(x_k^n, y_k^n)^{\frac{1}{1+\rho}}\Big]^{1+\rho}$$

$$= 4 \times 2^{-(n-l+1)\rho R_x - (n-k+1)\rho R_y} \left[\sum_{y} \Big[\sum_{x} p_{x,y}(x,y)^{\frac{1}{1+\rho}}\Big]^{1+\rho}\right]^{k-l} \Big[\sum_{x,y} p_{x,y}(x,y)^{\frac{1}{1+\rho}}\Big]^{(1+\rho)(n-k+1)} \tag{F.4}$$

$$= 4 \times 2^{-(k-l)\left[\rho R_x - \log\left[\sum_y \left[\sum_x p_{x,y}(x,y)^{\frac{1}{1+\rho}}\right]^{1+\rho}\right]\right]} 2^{-(n-k+1)\left[\rho(R_x+R_y) - (1+\rho)\log\left[\sum_{x,y} p_{x,y}(x,y)^{\frac{1}{1+\rho}}\right]\right]}$$

$$= 4 \times 2^{-(k-l)E_{x|y}(R_x,\rho)-(n-k+1)E_{xy}(R_x,R_y,\rho)} \tag{F.5}$$

$$= 4 \times 2^{-(n-l+1)\left[\frac{k-l}{n-l+1}E_{x|y}(R_x,\rho)+\frac{n-k+1}{n-l+1}E_{xy}(R_x,R_y,\rho)\right]} \tag{F.6}$$

$$\leq 4 \times 2^{-(n-l+1)\sup_{\rho\in[0,1]}\left[\frac{k-l}{n-l+1}E_{x|y}(R_x,\rho)+\frac{n-k+1}{n-l+1}E_{xy}(R_x,R_y,\rho)\right]} \tag{F.7}$$

$$= 4 \times 2^{-(n-l+1)E_x^{ML}\left(R_x,R_y,\frac{k-l}{n-l+1}\right)}$$

$$= 4 \times 2^{-(n-l+1)E_x(R_x,R_y,\frac{k-l}{n-l+1})}. \tag{F.8}$$

In (F.1) we explicitly indicate the three conditions that a suffix pair $(\tilde{x}_l^n, \tilde{y}_k^n)$ must satisfy to result in a decoding error. In (F.2) we sum out over the common prefixes (x_1^{l-1}, y_1^{l-1}), and use the fact that the random binning is done independently at each encoder, thus we can use the inequalities in (4.6) and (4.7). We get (F.3) by limiting ρ to the interval $0 \leq \rho \leq 1$, as in (2.25). Getting (F.4) from (F.3) follows by a number of basic manipulations. In (F.4) we get the single letter expression by again using the memorylessness of the sources. In (F.5) we use the definitions of $E_{x|y}$ and E_{xy} from (4.9) in Theorem 3. Noting that the bound holds for all $\rho \in [0, 1]$, optimizing over ρ results in (F.7). Finally, using the definition in (4.8) and the remark following Theorem 5 that the maximum-likelihood and universal exponents are equal gives (F.8). The bound on $p_n(l, k)$ when $l > k$, is developed in the same fashion. □

F.2 Universal decoding: Proof of Lemma 11

In this section, we prove Lemma 11. We use the techniques called *method of types* developed by Imre Csiszár. This method is elegantly explained in [29] and [28]. The proof here is similar to that for the point-to-point case in the proof of Proposition 4. However, we need the concept of V-shells defined in [29]. Essentially, a V-shell is the conditional type set of y sequence given x sequence where x and y are from the alphabet $\mathcal{X} \times \mathcal{Y}$. The technique of V-shell is originally used in the proof of channel coding theorems in [29]. Due to the duality of the channel coding and source coding with decoding information, it is no surprise that it proves a very powerful tool in the distributed source coding problems. That being said, here is the proof.

The error probability $p_n(l,k)$ can be thought as starting from (F.2) with the condition $(k-l)H(\tilde{x}_l^{k-1}|\tilde{y}_l^{k-1}) + (n-k+1)H(\tilde{x}_k^n, \tilde{y}_k^n) < (k-l)H(x_l^{k-1}|y_l^{k-1}) + (n-k+1)H(x_k^n, y_k^n)$ substituted for $p_{xy}(\tilde{x}_l^n, \tilde{y}_l^n) > p_{xy}(x_l^n, y_l^n)$, we get

$$p_n(l,k) \leq \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \sum_{\substack{y_l^{k-1} \in \mathcal{T}_{P^{k-l}}, \\ y_k^n \in \mathcal{T}_{P^{n-k}}}} \sum_{\substack{x_l^{k-1} \in \mathcal{T}_{V^{k-l}}(y_l^{k-1}), \\ x_k^n \in \mathcal{T}_{V^{n-k}(y_k^n)}}} \min\Bigg[1, \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k} \text{ s.t.} \\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l}) < \\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} \sum_{\tilde{y}_k^n \in \mathcal{T}_{\tilde{P}^{n-k}}} \sum_{\tilde{x}_l^{k-1} \in \mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})} \sum_{\tilde{x}_k^n \in \mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)} 4 \times 2^{-(n-l+1)R_x - (n-k+1)R_y} \Bigg] p_{xy}(x_1^n, y_1^n) \tag{F.9}$$

In (F.9) we enumerate all the source sequences in a way that allows us to focus on the types of the important subsequences. We enumerate the possibly misleading candidate sequences in terms of their suffixes types. We restrict the sum to those pairs $(\tilde{x}_1^n, \tilde{y}_1^n)$ that could lead to mistaken decoding, defining the compact notation $S(P^{n-k}, P^{k-l}, V^{n-k}, V^{k-l}) \triangleq (k-l)H(V^{k-l}|P^{k-l}) + (n-k+1)H(P^{n-k} \times V^{n-k})$, which is the weighted suffix entropy condition rewritten in terms of types.

Note that the summations within the minimization in (F.9) do not depend on the arguments within these sums. Thus, we can bound this sum separately to get a bound on

the number of possibly misleading source pairs $(\tilde{x}_1^n, \tilde{y}_1^n)$.

$$\sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}\ \text{s.t.}\\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l})<\\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} \sum_{\tilde{y}_k^n\in\mathcal{T}_{\tilde{P}^{n-k}}} \sum_{\tilde{x}_l^{k-1}\in\mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})} \sum_{\tilde{x}_k^n\in\mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)}$$

$$\leq \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}\ \text{s.t.}\\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l})<\\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} \sum_{\tilde{y}_k^n\in\mathcal{T}_{\tilde{P}^{n-k}}} |\mathcal{T}_{\tilde{V}^{k-l}(y_l^{k-1})}||\mathcal{T}_{\tilde{V}^{n-k}(\tilde{y}_k^n)}| \tag{F.10}$$

$$\leq \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}\ \text{s.t.}\\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l})<\\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} |\mathcal{T}_{\tilde{P}^{n-k}}|2^{(k-l)H(\tilde{V}^{k-l}|P^{k-l})}2^{(n-k+1)H(\tilde{V}^{n-k}|\tilde{P}^{n-k})} \tag{F.11}$$

$$\leq \sum_{\substack{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}\ \text{s.t.}\\ S(\tilde{P}^{n-k},P^{k-l},\tilde{V}^{n-k},\tilde{V}^{k-l})<\\ S(P^{n-k},P^{k-l},V^{n-k},V^{k-l})}} 2^{(k-l)H(\tilde{V}^{k-l}|P^{k-l})+(n-k+1)H(\tilde{P}^{n-k}\times\tilde{V}^{n-k})} \tag{F.12}$$

$$\leq \sum_{\tilde{V}^{n-k},\tilde{V}^{k-l},\tilde{P}^{n-k}} 2^{(k-l)H(V^{k-l}|P^{k-l})+(n-k+1)H(P^{n-k}\times V^{n-k})} \tag{F.13}$$

$$\leq (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|}2^{(k-l)H(V^{k-l}|P^{k-l})+(n-k+1)H(P^{n-k}\times V^{n-k})} \tag{F.14}$$

In (F.10) we sum over all $\tilde{x}_l^{k-1} \in \mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})$. In (F.11) we use standard bounds, e.g., $|\mathcal{T}_{\tilde{V}^{k-l}}(y_l^{k-1})| \leq 2^{(k-l)H(\tilde{V}^{k-l}|P^{k-l})}$ for $y_l^{k-1} \in \mathcal{T}_{P^{k-l}}$. We also sum over all $\tilde{x}_k^n \in \mathcal{T}_{\tilde{V}^{n-k}}(\tilde{y}_k^n)$ and over all $\tilde{y}_k^n \in \mathcal{T}_{\tilde{P}^{n-k}}$ in (F.11). By definition of the decoding rule $(\tilde{x}_1^n, \tilde{y}_1^n)$ can only lead to a decoding error if $(k-l)H(\tilde{V}^{k-l}|P^{k-l})] + (n-k+1)H(\tilde{P}^{n-k}\times\tilde{V}^{n-k}) < (k-l)H(V^{k-l}|P^{k-l}) + (n-k+1)H(P^{n-k}\times V^{n-k})$. In (F.14) we apply the polynomial bound on the number of types.

We substitute (F.14) into (F.9) and pull out the polynomial term, giving

$$\begin{aligned}
p_n(l,k) \leq{}& (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \sum_{\substack{y_l^{k-1}\in\mathcal{T}_{P^{k-l}},\\ y_k^n\in\mathcal{T}_{P^{n-k}}}} \sum_{\substack{x_l^{k-1}\in\mathcal{T}_{V^{k-l}}(y_l^{k-1}),\\ x_k^n\in\mathcal{T}_{V^{n-k}(y_k^n)}}} \\
&\min\left[1, 4\times 2^{-(k-l)[R_x-H(V^{k-l}|P^{k-l})]-(n-k+1)[R_x+R_y-H(V^{n-k}\times P^{n-k})]}\right] p_{x_l^n,y_l^n}(x_l^n,y_l^n) \\
\leq{}& 4\times(n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}} \\
&\times 2^{\max\left[0,-(k-l)[R_x-H(V^{k-l}|P^{k-l})]-(n-k+1)[R_x+R_y-H(V^{n-k}\times P^{n-k})]\right]} \\
&\times 2^{-(k-l)D(V^{k-l}\times P^{k-l}\|p_{xy})-(n-k+1)D(V^{n-k}\times P^{n-k}\|p_{xy})}
\end{aligned} \tag{F.15}$$

$$\leq 4 \times (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}}$$
$$2^{-(n-l+1)\left[\lambda D(V^{k-l}\times P^{k-l}\|p_{xy})+\bar{\lambda}D(V^{n-k}\times P^{n-k}\|p_{xy})+\left|\lambda[R_x-H(V^{k-l}|P^{k-l})]+\bar{\lambda}[R_x+R_y-H(V^{n-k}\times P^{n-k})]\right|^+\right]} \tag{F.16}$$

$$\leq 4 \times (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-k},P^{k-l}} \sum_{V^{n-k},V^{k-l}}$$
$$2^{-(n-l+1)\inf_{\tilde{x},\tilde{y},\bar{x},\bar{y}}\left[\lambda D(p_{\tilde{x},\tilde{y}}\|p_{xy})+\bar{\lambda}D(p_{\bar{x},\bar{y}}\|p_{xy})+\left|\lambda[R_x-H(\tilde{x}|\tilde{y})]+\bar{\lambda}[R_x+R_y-H(\bar{x},\bar{y})]\right|^+\right]} \tag{F.17}$$

$$\leq 4 \times (n-l+2)^{4|\mathcal{X}||\mathcal{Y}|} 2^{-(n-l+1)E_x(R_x,R_y,\lambda)}$$
$$\leq K_1 2^{-(n-l+1)[E_x(R_x,R_y,\lambda)-\epsilon]} \tag{F.18}$$

In (F.15) we use the memorylessness of the source, and exponential bounds on the probability of observing (x_l^{k-1}, y_l^{k-1}) and (x_k^n, y_k^n). In (F.16) we pull out $(n-l+1)$ from all terms, by noticing that $\lambda = (k-l)/(n-l+1) \in [0,1]$ and $\bar{\lambda} \triangleq 1-\lambda = (n-k+1)/(n-l+1)$. In (F.17) we minimize the exponent over all choices of distributions $p_{\tilde{x},\tilde{y}}$ and $p_{\bar{x},\bar{y}}$. In (F.18) we define the universal random coding exponent $E_x(R_x, R_y, \lambda) \triangleq \inf_{\tilde{x},\tilde{y},\bar{x},\bar{y}}\{\lambda D(p_{\tilde{x},\tilde{y}}\|p_{xy}) + \bar{\lambda}D(p_{\bar{x},\bar{y}}\|p_{xy}) + \left|\lambda[R_x - H(\tilde{x}|\tilde{y})] + \bar{\lambda}[R_x + R_y - H(\bar{x},\bar{y})]\right|^+\}$ where $0 \leq \lambda \leq 1$ and $\bar{\lambda} = 1-\lambda$. We also incorporate the number of conditional and marginal types into the polynomial bound, as well as the sum over k, and then push the polynomial into the exponent since for any polynomial F, $\forall E, \epsilon > 0$, there exists $C \in (0,\infty)$, s.t. $F(\Delta)e^{-\Delta E} \leq Ce^{-\Delta(E-\epsilon)}$. $\square$

Appendix G

Equivalence of ML and universal error exponents and tilted distributions

In this appendix, we give a detail proof of Lemma 8. This section also contains some very useful analysis on the source coding error exponents and their geometry, tilted distributions and their properties. The key mathematical tool is convex optimization. The fundamental lemmas which cannot be easily found in the literature in Section G.3 are used throughout this thesis. For notation simplicity, we change the logarithm from base 2 in the main body of the thesis to base e in this section.

Our goal in Theorem 5 is to show that the maximum likelihood (ML) error exponent equals the universal error exponent. It is sufficient to show that for all γ,

$$E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma)$$

Where the ML error exponent:

$$
\begin{aligned}
E_x^{ML}(R_x, R_y, \gamma) &= \sup_{\rho \in [0,1]} \{\gamma E_{x|y}(R_x, \rho) + (1-\gamma) E_{xy}(R_x, R_y, \rho)\} \\
&= \sup_{\rho \in [0,1]} \{\rho R^{(\gamma)} - \gamma \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho}) \\
&\quad -(1-\gamma)(1+\rho)\log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})\} \\
&= \sup_{\rho \in [0,1]} \{E_x^{ML}(R_x, R_y, \gamma, \rho)\}
\end{aligned}
$$

Write the function inside the sup argument as $E_x^{ML}(R_x, R_y, \gamma, \rho)$. The universal error exponent:

$$
\begin{aligned}
& E_x^{UN}(R_x, R_y, \gamma) \\
&= \inf_{q_{xy}, o_{xy}} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) \\
&\quad +|\gamma(R_x - H(q_{x|y})) + (1-\gamma)(R_x + R_y - H(o_{xy}))|^+\} \\
&= \inf_{q_{xy}, o_{xy}} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) + |R^{(\gamma)} - \gamma H(q_{x|y}) - (1-\gamma) H(o_{xy})|^+\}
\end{aligned}
$$

Here we define $R^{(\gamma)} = \gamma R_x + (1-\gamma)(R_x + R_y) > \gamma H(p_{x|y}) + (1-\gamma) H(p_{xy})$. For notational simplicity, we write q_{xy} and o_{xy} as two arbitrary joint distributions on $\mathcal{X} \times \mathcal{Y}$ instead of $p_{\bar{x}\bar{y}}$ and $p_{\bar{\bar{x}}\bar{\bar{y}}}$. We still write p_{xy} as the distribution of the source.

Before the proof, we define a pair of distributions that we need.

Definition 15 *Tilted distribution of p_{xy}: p_{xy}^{ρ}, for all $\rho \in [-1, \infty)$*

$$
p_{xy}^{\rho}(x,y) = \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}}
$$

The entropy of the tilted distribution is written as $H(p_{xy}^{\rho})$. Obviously $p_{xy}^{0} = p_{xy}$.

Definition 16 *$x-y$ tilted distribution of p_{xy}: $\bar{p}_{xy}^{\rho}$, for all $\rho \in [-1, +\infty)$*

$$
\begin{aligned}
\bar{p}_{xy}^{\rho}(x,y) &= \frac{[\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}]^{1+\rho}}{\sum_t [\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}]^{1+\rho}} \times \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}} \\
&= \frac{A(y,\rho)}{B(\rho)} \times \frac{C(x,y,\rho)}{D(y,\rho)}
\end{aligned}
$$

Where

$$\begin{aligned} A(y,\rho) &= [\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}]^{1+\rho} = D(y,\rho)^{1+\rho} \\ B(\rho) &= \sum_s [\sum_t p_{xy}(s,t)^{\frac{1}{1+\rho}}]^{1+\rho} = \sum_y A(y,\rho) \\ C(x,y,\rho) &= p_{xy}(x,y)^{\frac{1}{1+\rho}} \\ D(y,\rho) &= \sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}} = \sum_x C(x,y,\rho) \end{aligned}$$

The marginal distribution for y is $\frac{A(y,\rho)}{B(\rho)}$. Obviously $\bar{p}^0_{xy} = p_{xy}$. Write the conditional distribution of x given y under distribution $\bar{p}^\rho_{xy}$ as $\bar{p}^\rho_{x|y}$, where $\bar{p}^\rho_{x|y}(x,y) = \frac{C(x,y,\rho)}{D(y,\rho)}$, and the conditional entropy of x given y under distribution $\bar{p}^\rho_{xy}$ as $H(\bar{p}^\rho_{x|y})$. Obviously $H(\bar{p}^0_{x|y}) = H(p_{x|y})$.

The conditional entropy of x given y for the $x-y$ tilted distribution is

$$H(\bar{p}^\rho_{x|y=y}) = -\sum_x \frac{C(x,y,\rho)}{D(y,\rho)} \log(\frac{C(x,y,\rho)}{D(y,\rho)})$$

We introduce $A(y,\rho)$, $B(\rho)$, $C(x,y,\rho)$, $D(y,\rho)$ to simplify the notations. Some of their properties are shown in Lemma 25.

While tilted distributions are common optimal distributions in large deviation theory, it is useful to contemplate why we need to introduce these *two* tilted distributions. In the proof of Lemma 8, through a Lagrange multiplier argument, we will show that $\{p^\rho_{xy} : \rho \in [-1,+\infty)\}$ is the family of distributions that minimize the Kullback−Leibler distance to p_{xy} with fixed *entropy* and $\{\bar{p}^\rho_{xy} : \rho \in [-1,+\infty)\}$ is the family of distributions that minimize the Kullback−Leibler distance to p_{xy} with fixed *conditional entropy*. Using a Lagrange multiplier argument, we parametrize the universal error exponent $E_x^{UN}(R_x,R_y,\gamma)$ in terms of ρ and show the equivalence of the universal and maximum likelihood error exponents.

Now we are ready to prove Lemma 5: $E_x^{ML}(R_x,R_y,\gamma) = E_x^{UN}(R_x,R_y,\gamma)$.

Proof:

G.1 case 1: $\gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}) < R^{(\gamma)} < \gamma H(\bar{p}^1_{x|y}) + (1-\gamma)H(p^1_{xy})$.

First, from Lemma 31 and Lemma 32:

$$\frac{\partial E_x^{ML}(R_x, R_y, \gamma, \rho)}{\partial \rho} = R^{(\gamma)} - \gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p^{\rho}_{xy})$$

Then, using Lemma 22 and Lemma 26, we have:

$$\frac{\partial^2 E_x^{ML}(R_x, R_y, \gamma, \rho)}{\partial \rho} \leq 0$$

.

So ρ maximize $E_x^{ML}(R_x, R_y, \gamma, \rho)$, if and only if:

$$0 = \frac{\partial E_x^{ML}(R_x, R_y, \gamma, \rho)}{\partial \rho} = R^{(\gamma)} - \gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p^{\rho}_{xy}) \tag{G.1}$$

Because $R^{(\gamma)}$ is in the interval $[\gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}), \gamma H(\bar{p}^{1}_{x|y}) + (1-\gamma)H(p^{1}_{xy})]$ and the entropy functions monotonically-increase over ρ, we can find $\rho^* \in (0,1)$, s.t.

$$\gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy}) = R^{(\gamma)}$$

Using Lemma 29 and Lemma 30 we get:

$$E_x^{ML}(R_x, R_y, \gamma) \quad = \quad \gamma D(\bar{p}^{\rho^*}_{xy} \| p_{xy}) + (1-\gamma) D(p^{\rho^*}_{xy} \| p_{xy}) \tag{G.2}$$

Where $\gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy}) = R^{(\gamma)}$, ρ^* is generally unique because both $H(\bar{p}^{\rho}_{x|y})$ and $H(p^{\rho}_{xy})$ are strictly increasing with ρ.

Secondly

$$\begin{aligned}
& E_x^{UN}(R_x, R_y, \gamma) \\
= & \inf_{q_{xy}, o_{xy}} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) + |R^{(\gamma)} - \gamma H(q_{x|y}) - (1-\gamma)H(o_{xy})|^+\} \\
= & \inf_b \{ \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) + |R^{(\gamma)} - b)|^+\} \\
= & \inf_{b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})} \{ \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) \\
& + |R^{(\gamma)} - b)|^+\}
\end{aligned} \tag{G.3}$$

The last equality is true because, for $b < \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}) < R^{(\gamma)}$,

$$
\begin{aligned}
& \inf_{q_{xy},o_{xy}:\gamma H(q_{x|y})+(1-\gamma)H(o_{xy})=b} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy}) + |R^{(\gamma)} - b|^+\} \\
\geq \; & 0 + R^{(\gamma)} - b \\
= \; & \inf_{q_{xy},o_{xy}:H(q_{x|y})=H(p_{x|y}),H(o_{xy})=H(p_{xy})} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy}) + |R^{(\gamma)} - b|^+\} \\
\geq \; & \inf_{q_{xy},o_{xy}:H(q_{x|y})=H(p_{x|y}),H(o_{xy})=H(p_{xy})} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy}) \\
& + |R^{(\gamma)} - \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})|^+\} \\
\geq \; & \inf_{q_{xy},o_{xy}:\gamma H(q_{x|y})+(1-\gamma)H(o_{xy})=\gamma H(p_{x|y})+(1-\gamma)H(p_{xy})} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy}) \\
& + |R^{(\gamma)} - \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})|^+\}
\end{aligned}
$$

Fixing $b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})$, the inner infimum in (G.3) is an optimization problem on q_{xy}, o_{xy} with equality constraints $\sum_x \sum_y q_{xy}(x,y) = 1$, $\sum_x \sum_y o_{xy}(x,y) = 1$ and $\gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b$ and the obvious inequality constraints $0 \leq q_{xy}(x,y) \leq 1, 0 \leq o_{xy}(x,y) \leq 1, \forall x, y$. In the following formulation of the optimization problem, we relax one equality constraint to an inequality constraint $\gamma H(q_{x|y})+(1-\gamma)H(o_{xy}) \geq b$ to make the optimization problem *convex*. It turns out later that the optimal solution to the relaxed problem is also the optimal solution to the original problem because $b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})$. The resulting optimization problem is:

$$
\begin{aligned}
& \inf_{q_{xy},o_{xy}} \{\gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy})\} \\
& \text{s.t.} \sum_x \sum_y q_{xy}(x,y) = 1 \\
& \sum_x \sum_y o_{xy}(x,y) = 1 \\
& b - \gamma H(q_{x|y}) - (1-\gamma)H(o_{xy}) \leq 0 \\
& 0 \leq q_{xy}(x,y) \leq 1, \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y} \\
& 0 \leq o_{xy}(x,y) \leq 1, \quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y}
\end{aligned} \tag{G.4}
$$

The above optimization problem is *convex* because the objective function and the inequality constraint functions are convex and the equality constraint functions are affine[10]. The Lagrange multiplier function for this convex optimization problem is:

$$
\begin{aligned}
& L(q_{xy}, o_{xy}, \rho, \mu_1, \mu_2, \nu_1, \nu_2, \nu_3, \nu_4) \\
= \; & \gamma D(q_{xy}||p_{xy}) + (1-\gamma) D(o_{xy}||p_{xy}) + \\
& \mu_1(\sum_x \sum_y q_{xy}(x,y) - 1) + \mu_2(\sum_x \sum_y o_{xy}(x,y) - 1) + \\
& \rho(b - \gamma H(q_{x|y}) - (1-\gamma) H(o_{xy})) + \\
& \sum_x \sum_y \big\{ \nu_1(x,y)(-q_{xy}(x,y)) + \nu_2(x,y)(1 - q_{xy}(x,y)) \\
& \qquad + \nu_3(x,y)(-o_{xy}(x,y)) + \nu_4(x,y)(1 - o_{xy}(x,y)) \big\}
\end{aligned}
$$

Where ρ, μ_1, μ_2 are real numbers and $\nu_i \in R^{|\mathcal{X}||\mathcal{Y}|}$, $i = 1, 2, 3, 4$.

According to the KKT conditions for convex optimization[10], q_{xy}, o_{xy} minimize the convex optimization problem in (G.4) if and only if the following conditions are simultaneously satisfied for some q_{xy}, o_{xy}, μ_1, μ_2, ν_1, ν_2, ν_3, ν_4 and ρ:

$$
\begin{aligned}
0 \; = \; & \frac{\partial L(q_{xy}, o_{xy}, \rho, \mu_1, \mu_2, \nu_1, \nu_2, \nu_3, \nu_4)}{\partial q_{xy}(x,y)} \\
= \; & \gamma[-\log(p_{xy}(x,y)) + (1+\rho)(1 + \log(q_{xy}(x,y))) + \rho \log(\sum_s q_{xy}(s,y))] \\
& + \mu_1 - \nu_1(x,y) - \nu_2(x,y)
\end{aligned} \tag{G.5}
$$

$$
\begin{aligned}
0 \; = \; & \frac{\partial L(q_{xy}, o_{xy}, \rho, \mu_1, \mu_2, \nu_1, \nu_2, \nu_3, \nu_4)}{\partial o_{xy}(x,y)} \\
= \; & (1-\gamma)[-\log(p_{xy}(x,y)) + (1+\rho)(1 + \log(o_{xy}(x,y)))] + \mu_2 - \nu_3(x,y) - \nu_4(x,y)
\end{aligned} \tag{G.6}
$$

For all x, y and

$$
\begin{aligned}
& \sum_x \sum_y q_{xy}(x,y) = 1 \\
& \sum_x \sum_y o_{xy}(x,y) = 1 \\
& \rho(\gamma H(q_{x|y}) + (1-\gamma) H(o_{xy}) - b) = 0 \\
& \rho \geq 0 \\
& \nu_1(x,y)(-q_{xy}(x,y)) = 0, \quad \nu_2(x,y)(1 - q_{xy}(x,y)) = 0 \quad \forall x, y \\
& \nu_3(x,y)(-o_{xy}(x,y)) = 0, \quad \nu_4(x,y)(1 - o_{xy}(x,y)) = 0 \quad \forall x, y \\
& \nu_i(x,y) \geq 0, \quad \forall x, y, i = 1, 2, 3, 4
\end{aligned} \tag{G.7}
$$

Solving the above standard Lagrange multiplier equations (G.5), (G.6) and (G.7), we have:

$$\begin{aligned}
q_{xy}(x,y) &= \frac{[\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho_b}}]^{1+\rho_b}}{\sum_t[\sum_s p_{xy}(s,t)^{\frac{1}{1+\rho_b}}]^{1+\rho_b}} \frac{p_{xy}(x,y)^{\frac{1}{1+\rho_b}}}{\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho_b}}} \\
&= \bar{p}_{xy}^{\rho_b}(x,y) \\
o_{xy}(x,y) &= \frac{p_{xy}(x,y)^{\frac{1}{1+\rho_b}}}{\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho_b}}} \\
&= p_{xy}^{\rho_b}(x,y) \\
\nu_i(x,y) &= 0 \quad \forall x,y,i=1,2,3,4 \\
\rho &= \rho_b
\end{aligned} \tag{G.8}$$

Where ρ_b satisfies the following condition

$$\gamma H(\bar{p}_{x|y}^{\rho_b}) + (1-\gamma)H(p_{xy}^{\rho_b}) = b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})$$

and thus $\rho_b \geq 0$ because both $H(\bar{p}_{x|y}^{\rho})$ and $H(p_{xy}^{\rho})$ are monotonically increasing with ρ as shown in Lemma 22 and Lemma 26.

Notice that all the KKT conditions are simultaneously satisfied with the inequality constraint $\gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) \geq b$ being met with equality. Thus, the relaxed optimization problem has the same optimal solution as the original problem as promised. The optimal q_{xy} and o_{xy} are the $x-y$ tilted distribution $\bar{p}_{xy}^{\rho_b}$ and standard tilted distribution $p_{xy}^{\rho_b}$ of p_{xy} with the same parameter $\rho_b \geq 0$. chosen s.t.

$$\gamma H(\bar{p}_{x|y}^{\rho_b}) + (1-\gamma)H(p_{xy}^{\rho_b}) = b$$

Now we have :

$$\begin{aligned}
& E_x^{UN}(R_x, R_y, \gamma) \\
= & \inf_{b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})} \{ \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b} \{ \\
& \qquad \gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy}) + |R^{(\gamma)} - b|^+ \} \} \\
= & \inf_{b \geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})} \{ \gamma D(\bar{p}_{xy}^{\rho_b}||p_{xy}) + (1-\gamma)D(p_{xy}^{\rho_b}||p_{xy}) + |R^{(\gamma)} - b|^+ \} \\
= & \min[\inf_{\rho \geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}_{x|y}^{\rho}) + (1-\gamma)H(p_{xy}^{\rho})} \{ \\
& \qquad \gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) + (1-\gamma)D(p_{xy_\rho}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}_{x|y}^{\rho}) - (1-\gamma)H(p_{xy}^{\rho}) \}, \\
& \inf_{\rho \geq 0: R^{(\gamma)} \leq \gamma H(\bar{p}_{x|y}^{\rho}) + (1-\gamma)H(p_{xy}^{\rho})} \{ \gamma D(\bar{p}_{xy}^{\rho}||p_{xy}) + (1-\gamma)D(p_{xy_\rho}||p_{xy}) \}]
\end{aligned} \tag{G.9}$$

Notice that $H(p^{\rho}_{xy})$, $H(\bar{p}^{\rho}_{x|y})$, $D(\bar{p}^{\rho}_{xy}||p_{xy})$ and $D(p^{\rho}_{xy}||p_{xy})$ are all strictly increasing with $\rho > 0$ as shown in Lemma 22, Lemma 23, Lemma 26 and Lemma 27 later in this appendix. We have:

$$\begin{aligned}\inf_{\rho\geq 0: R^{(\gamma)} \leq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})} & \{\gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy})\} \\ & = \gamma D(\bar{p}^{\rho^*}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}||p_{xy}) \end{aligned} \tag{G.10}$$

where $R^{(\gamma)} = \gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy})$. Applying the results in Lemma 28 and Lemma 24, we get:

$$\begin{aligned}&\inf_{\rho\geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})} \{\gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy}) + R^{(\gamma)} \\ &\qquad -\gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p^{\rho}_{xy})\} \\ &= \gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p^{\rho}_{xy})|_{\rho=\rho^*} \\ &= \gamma D(\bar{p}^{\rho^*}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}||p_{xy}) \end{aligned} \tag{G.11}$$

This is true because for $\rho : R^{(\gamma)} \geq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})$, we know $\rho \leq 1$ because of the range of $R^{(\gamma)}$: $R^{(\gamma)} < \gamma H(\bar{p}^{1}_{x|y}) + (1-\gamma)H(p^{1}_{xy})$. Substituting (G.10) and (G.11) into (G.9), we get

$$\begin{aligned} E_x^{UN}(R_x, R_y, \gamma) &= \gamma D(\bar{p}^{\rho^*}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}||p_{xy}) \\ &\quad \text{where} \quad R^{(\gamma)} = \gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy}) \end{aligned} \tag{G.12}$$

So for $\gamma H(p_{x|y}) + (1-\gamma)H(p_{xy}) \leq R^{(\gamma)} \leq \gamma H(\bar{p}^{1}_{x|y}) + (1-\gamma)H(p^{1}_{xy})$, from (G.2) we have the desired property:

$$E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma)$$

G.2 case 2: $R^{(\gamma)} \geq \gamma H(\bar{p}^{1}_{x|y}) + (1-\gamma)H(p^{1}_{xy})$.

In this case, for all $0 \leq \rho \leq 1$

$$\frac{\partial E_x^{ML}(R_x, R_y, \gamma, \rho)}{\partial \rho} = R^{(\gamma)} - \gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p^{\rho}_{xy}) \geq R^{(\gamma)} - \gamma H(\bar{p}^{1}_{x|y}) - (1-\gamma)H(p^{1}_{xy}) \geq 0$$

So ρ takes value 1 to maximize the error exponent $E_x^{ML}(R_x, R_y, \gamma, \rho)$, thus

$$E_x^{ML}(R_x, R_y, \gamma) = R^{(\gamma)} - \gamma \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{2}})^2) - 2(1-\gamma)\log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{2}}) \tag{G.13}$$

Using the same convex optimization techniques as case G.1, we notice the fact that $\rho^* \geq 1$ for $R^{(\gamma)} = \gamma H(\bar{p}^{\rho^*}_{x|y}) + (1-\gamma)H(p^{\rho^*}_{xy})$. Then applying Lemma 28 and Lemma 24, we have:

$$\begin{aligned}
&\inf_{\rho\geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})} \{\gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy}) \\
&\qquad\qquad +R^{(\gamma)} - \gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p_{xy_\rho})\} \\
&= \gamma D(\bar{p}^{1}_{xy}||p_{xy}) + (1-\gamma)D(p^{1}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{1}_{x|y}) - (1-\gamma)H(p^{1}_{xy})
\end{aligned}$$

And

$$\begin{aligned}
&\inf_{\rho\geq 0: R^{(\gamma)} \leq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})} \{\gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy})\}] \\
= \;& \gamma D(\bar{p}^{\rho^*}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}||p_{xy}) \\
= \;& \gamma D(\bar{p}^{\rho^*}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho^*}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{\rho^*}_{x|y}) - (1-\gamma)H(p^{\rho^*}_{xy}) \\
\leq \;& \gamma D(\bar{p}^{1}_{xy}||p_{xy}) + (1-\gamma)D(p^{1}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{1}_{x|y}) - (1-\gamma)H(p^{1}_{xy})
\end{aligned}$$

Finally:

$$\begin{aligned}
& E_x^{UN}(R_x, R_y, \gamma) \\
= \;& \inf_{b\geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})} \{ \inf_{q_{xy}, o_{xy}: \gamma H(q_{x|y}) + (1-\gamma)H(o_{xy}) = b} \{ \\
& \gamma D(q_{xy}||p_{xy}) + (1-\gamma)D(o_{xy}||p_{xy}) + |R^{(\gamma)} - b|^+\}\} \\
= \;& \inf_{b\geq \gamma H(p_{x|y}) + (1-\gamma)H(p_{xy})} \{\gamma D(\bar{p}^{\rho_b}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho_b}_{xy}||p_{xy}) + |R^{(\gamma)} - b|^+\} \\
= \;& \min[\inf_{\rho\geq 0: R^{(\gamma)} \geq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})} \{ \\
& \gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{\rho}_{x|y}) - (1-\gamma)H(p^{\rho}_{xy})\}, \\
& \inf_{\rho\geq 0: R^{(\gamma)} \leq \gamma H(\bar{p}^{\rho}_{x|y}) + (1-\gamma)H(p^{\rho}_{xy})} \{\gamma D(\bar{p}^{\rho}_{xy}||p_{xy}) + (1-\gamma)D(p^{\rho}_{xy}||p_{xy})\}] \\
= \;& \gamma D(\bar{p}^{1}_{xy}||p_{xy}) + (1-\gamma)D(p^{1}_{xy}||p_{xy}) + R^{(\gamma)} - \gamma H(\bar{p}^{1}_{x|y}) - (1-\gamma)H(p^{1}_{xy}) \\
= \;& R^{(\gamma)} - \gamma \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{2}})^2) - 2(1-\gamma)\log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{2}}) \qquad \text{(G.14)}
\end{aligned}$$

The last equality is true by setting $\rho = 1$ in Lemma 29 and Lemma 30.

Again, $E_x^{ML}(R_x, R_y, \gamma) = E_x^{UN}(R_x, R_y, \gamma)$, thus we finish the proof. □

G.3 Technical Lemmas on tilted distributions

Some technical lemmas we used in the above proof of Lemma 8 are now discussed:

Lemma 22 $\frac{\partial H(p^{\rho}_{xy})}{\partial \rho} \geq 0$

Proof: From the definition of the tilted distribution we have the following observation:

$$\log(p^{\rho}_{xy}(x_1, y_1)) - \log(p^{\rho}_{xy}(x_2, y_2)) = \log(p_{xy}(x_1, y_1)^{\frac{1}{1+\rho}}) - \log(p_{xy}(x_2, y_2)^{\frac{1}{1+\rho}})$$

Using the above equality, we first derive the derivative of the tilted distribution, for all x, y

$$\begin{aligned}
\frac{\partial p^{\rho}_{xy}(x,y)}{\partial \rho} &= \frac{-1}{(1+\rho)^2} \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}} \log(p_{xy}(x,y))(\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}})}{(\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}})^2} \\
&\quad - \frac{-1}{(1+\rho)^2} \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}} (\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}} \log(p_{xy}(s,t)))}{(\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}})^2} \\
&= \frac{-1}{1+\rho} p^{\rho}_{xy}(x,y)[\log(p_{xy}(x,y)^{\frac{1}{1+\rho}}) - \sum_t \sum_s p^{\rho}_{xy}(s,t) \log(p_{xy}(s,t)^{\frac{1}{1+\rho}})] \\
&= \frac{-1}{1+\rho} p^{\rho}_{xy}(x,y)[\log(p^{\rho}_{xy}(x,y)) - \sum_t \sum_s p^{\rho}_{xy}(s,t) \log(p^{\rho}_{xy}(s,t))] \\
&= -\frac{p^{\rho}_{xy}(x,y)}{1+\rho}[\log(p^{\rho}_{xy}(x,y)) + H(p^{\rho}_{xy})] \qquad \text{(G.15)}
\end{aligned}$$

Then:

$$
\begin{aligned}
\frac{\partial H(p^{\rho}_{xy})}{\partial \rho} &= -\frac{\partial \sum_{x,y} p^{\rho}_{xy}(x,y)\log(p^{\rho}_{xy}(x,y))}{\partial \rho} \\
&= -\sum_{x,y}(1+\log(p^{\rho}_{xy}(x,y)))\frac{\partial p^{\rho}_{xy}(x,y)}{\partial \rho} \\
&= \sum_{x,y}(1+\log(p^{\rho}_{xy}(x,y)))\frac{p^{\rho}_{xy}(x,y)}{1+\rho}(\log(p^{\rho}_{xy}(x,y))+H(p^{\rho}_{xy})) \\
&= \frac{1}{1+\rho}\sum_{x,y} p^{\rho}_{xy}(x,y)\log(p^{\rho}_{xy}(x,y))(\log(p^{\rho}_{xy}(x,y))+H(p^{\rho}_{xy})) \\
&= \frac{1}{1+\rho}[\sum_{x,y} p^{\rho}_{xy}(x,y)(\log(p^{\rho}_{xy}(x,y)))^2 - H(p^{\rho}_{xy})^2] \\
&= \frac{1}{1+\rho}[\sum_{x,y} p^{\rho}_{xy}(x,y)(\log(p^{\rho}_{xy}(x,y)))^2 \sum_{x,y} p^{\rho}_{xy}(x,y) - H(p^{\rho}_{xy})^2] \\
&\geq_{(a)} \frac{1}{1+\rho}[(\sum_{x,y} p^{\rho}_{xy}(x,y)\log(p^{\rho}_{xy}(x,y)))^2 - H(p^{\rho}_{xy})^2] \\
&= 0
\end{aligned} \tag{G.16}
$$

where (a) is true by the Cauchy-Schwartz inequality. □

Lemma 23 $\frac{\partial D(p^{\rho}_{xy}\|p_{xy})}{\partial \rho} = \rho \frac{\partial H(p^{\rho}_{xy})}{\partial \rho}$

Proof: As shown in Lemma 29 and Lemma 31 respectively:

$$D(p^{\rho}_{xy}\|p_{xy}) = \rho H(p^{\rho}_{xy}) - (1+\rho)\log(\sum_{x,y} p_{xy}(x,y)^{\frac{1}{1+\rho}})$$

$$H(p^{\rho}_{xy}) = \frac{\partial (1+\rho)\log(\sum_{y}\sum_{x} p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial \rho}$$

We have:

$$
\begin{aligned}
\frac{\partial D(p^{\rho}_{xy}\|p_{xy})}{\partial \rho} &= H(p^{\rho}_{xy}) + \rho\frac{\partial H(p^{\rho}_{xy})}{\partial \rho} - \frac{\partial (1+\rho)\log(\sum_{y}\sum_{x} p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial \rho} \\
&= H(p^{\rho}_{xy}) + \rho\frac{\partial H(p^{\rho}_{xy})}{\partial \rho} - H(p^{\rho}_{xy}) \\
&= \rho\frac{\partial H(p^{\rho}_{xy})}{\partial \rho}
\end{aligned} \tag{G.17}
$$

□

Lemma 24 $sign\frac{\partial[D(p^\rho_{xy}\|p_{xy})-H(p^\rho_{xy})]}{\partial\rho} = sign(\rho-1)$.

Proof: Combining the results of the previous two lemmas, we have:

$$\frac{\partial D(p^\rho_{xy}\|p_{xy}) - H(p^\rho_{xy})}{\partial\rho} = (\rho-1)\frac{\partial H(p^\rho_{xy})}{\partial\rho} = sign(\rho-1)$$

□

Lemma 25 *Properties of* $\frac{\partial A(y,\rho)}{\partial\rho}$, $\frac{\partial B(\rho)}{\partial\rho}$, $\frac{\partial C(x,y,\rho)}{\partial\rho}$, $\frac{\partial D(y,\rho)}{\partial\rho}$ *and* $\frac{\partial H(\bar{p}^\rho_{x|y=y})}{\partial\rho}$

First,

$$\begin{aligned}
\frac{\partial C(x,y,\rho)}{\partial\rho} &= \frac{\partial p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\partial\rho} = -\frac{1}{1+\rho}p_{xy}(x,y)^{\frac{1}{1+\rho}}\log(p_{xy}(x,y)^{\frac{1}{1+\rho}}) \\
&= -\frac{C(x,y,\rho)}{1+\rho}\log(C(x,y,\rho)) \\
\frac{\partial D(y,\rho)}{\partial\rho} &= \frac{\partial\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}}{\partial\rho} = -\frac{1}{1+\rho}\sum_s p_{xy}(s,y)^{\frac{1}{1+\rho}}\log(p_{xy}(s,y)^{\frac{1}{1+\rho}}) \\
&= -\frac{\sum_x C(x,y,\rho)\log(C(x,y,\rho))}{1+\rho}
\end{aligned} \tag{G.18}$$

For a differentiable function $f(\rho)$,

$$\frac{\partial f(\rho)^{1+\rho}}{\partial\rho} = f(\rho)^{1+\rho}\log(f(\rho)) + (1+\rho)f(\rho)^\rho\frac{\partial f(\rho)}{\partial\rho}$$

So

$$\begin{aligned}
\frac{\partial A(y,\rho)}{\partial\rho} &= \frac{\partial D(y,\rho)^{1+\rho}}{\partial\rho} = D(y,\rho)^{1+\rho}\log(D(y,\rho)) + (1+\rho)D(y,\rho)^\rho\frac{\partial D(y,\rho)}{\partial\rho} \\
&= D(y,\rho)^{1+\rho}(\log(D(y,\rho)) - \sum_x\frac{C(x,y,\rho)}{D(y,\rho)}\log(C(x,y,\rho))) \\
&= D(y,\rho)^{1+\rho}(-\sum_x\frac{C(x,y,\rho)}{D(y,\rho)}\log(\frac{C(x,y,\rho)}{D(y,\rho)})) \\
&= A(y,\rho)H(\bar{p}^\rho_{x|y=y}) \\
\frac{\partial B(\rho)}{\partial\rho} &= \sum_y\frac{\partial A(y,\rho)}{\partial\rho} = \sum_y A(y,\rho)H(\bar{p}^\rho_{x|y=y}) = B(\rho)\sum_y\frac{A(y,\rho)}{B(\rho)}H(\bar{p}^\rho_{x|y=y}) \\
&= B(\rho)H(\bar{p}^\rho_{x|y})
\end{aligned}$$

And last:

$$
\begin{aligned}
&\frac{\partial H(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}=y})}{\partial \rho} \\
= \;& -\sum_x \left[\frac{\frac{\partial C(x,y,\rho)}{\partial \rho}}{D(y,\rho)} - \frac{C(x,y,\rho)\frac{\partial D(y,\rho)}{\partial \rho}}{D(y,\rho)^2}\right]\left[1+\log\left(\frac{C(x,y,\rho)}{D(y,\rho)}\right)\right] \\
= \;& -\sum_x \left[\frac{-\frac{C(x,y,\rho)}{1+\rho}\log(C(x,y,\rho))}{D(y,\rho)} + \frac{C(x,y,\rho)\frac{\sum_s C(s,y,\rho)\log(C(s,y,\rho))}{1+\rho}}{D(y,\rho)^2}\right]\left[1+\log\left(\frac{C(x,y,\rho)}{D(y,\rho)}\right)\right] \\
= \;& \frac{1}{1+\rho}\sum_x [\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y)\log(C(x,y,\rho)) - \bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y)\sum_s \bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(s,y)\log(C(s,y,\rho))] \\
& \times[1+\log(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y))] \\
= \;& \frac{1}{1+\rho}\sum_x \bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y)[\log(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y)) - \sum_s \bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(s,y)\log(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(s,y))][1+\log(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y))] \\
= \;& \frac{1}{1+\rho}\sum_x \bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y)\log(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y))[\log(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y)) - \sum_s \bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(s,y)\log(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(s,y))] \\
= \;& \frac{1}{1+\rho}\sum_x \bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y)\log(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y))\log(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y)) - \frac{1}{1+\rho}[\sum_x \bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y)\log(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y))]^2 \\
\geq \;& 0 \qquad \text{(G.19)}
\end{aligned}
$$

The inequality is true by the Cauchy-Schwartz inequality and by noticing that $\sum_x \bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}(x,y) = 1$. □

These properties will again be used in the proofs in the following lemmas.

Lemma 26 $\frac{\partial H(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}})}{\partial \rho} \geq 0$

Proof:

$$
\begin{aligned}
\frac{\partial \frac{A(y,\rho)}{B(\rho)}}{\partial \rho} &= \frac{1}{B(\rho)^2}\left(\frac{\partial A(y,\rho)}{\partial \rho}B(\rho) - \frac{\partial B(\rho)}{\partial \rho}A(y,\rho)\right) \\
&= \frac{1}{B(\rho)^2}(A(y,\rho)H(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}=y})B(\rho) - H(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}})B(\rho)A(y,\rho)) \\
&= \frac{A(y,\rho)}{B(\rho)}(H(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}=y}) - H(\bar{p}^{\rho}_{\mathsf{x}|\mathsf{y}}))
\end{aligned}
$$

Now,

$$
\begin{aligned}
\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho} &= \frac{\partial}{\partial \rho}\sum_y \frac{A(y,\rho)}{B(\rho)}\sum_x \frac{C(x,y,\rho)}{D(y,\rho)}[-\log(\frac{C(x,y,\rho)}{D(y,\rho)})] \\
&= \frac{\partial}{\partial \rho}\sum_y \frac{A(y,\rho)}{B(\rho)} H(\bar{p}^{\rho}_{x|y=y}) \\
&= \sum_y \frac{A(y,\rho)}{B(\rho)}\frac{\partial H(\bar{p}^{\rho}_{x|y=y})}{\partial \rho} + \sum_y \frac{\partial \frac{A(y,\rho)}{B(\rho)}}{\partial \rho} H(\bar{p}^{\rho}_{x|y=y}) \\
&\geq \sum_y \frac{\partial \frac{A(y,\rho)}{B(\rho)}}{\partial \rho} H(\bar{p}^{\rho}_{x|y=y}) \\
&= \sum_y \frac{A(y,\rho)}{B(\rho)}(H(\bar{p}^{\rho}_{x|y=y}) - H(\bar{p}^{\rho}_{x|y}))H(\bar{p}^{\rho}_{x|y=y}) \\
&= \sum_y \frac{A(y,\rho)}{B(\rho)} H(\bar{p}^{\rho}_{x|y=y})^2 - H(\bar{p}^{\rho}_{x|y})^2 \\
&= (\sum_y \frac{A(y,\rho)}{B(\rho)} H(\bar{p}^{\rho}_{x|y=y})^2)(\sum_y \frac{A(y,\rho)}{B(\rho)}) - H(\bar{p}^{\rho}_{x|y})^2 \\
&\geq_{(a)} (\sum_y \frac{A(y,\rho)}{B(\rho)} H(\bar{p}^{\rho}_{x|y=y}))^2 - H(\bar{p}^{\rho}_{x|y})^2 \\
&= 0
\end{aligned} \tag{G.20}
$$

where (a) is again true by the Cauchy-Schwartz inequality. □

Lemma 27 $\frac{\partial D(\bar{p}^{\rho}_{xy}\|p_{xy})}{\partial \rho} = \rho \frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho}$

Proof: As shown in Lemma 30 and Lemma 32 respectively:

$$D(\bar{p}^{\rho}_{xy}\|p_{xy}) = \rho H(\bar{p}^{\rho}_{x|y}) - \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})$$

$$H(\bar{p}^{\rho}_{x|y}) = \frac{\partial \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})}{\partial \rho}$$

We have:

$$
\begin{aligned}
\frac{\partial D(\bar{p}^{\rho}_{xy}\|p_{xy})}{\partial \rho} &= H(\bar{p}^{\rho}_{x|y}) + \rho \frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho} - \frac{\partial \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})}{\partial \rho} \\
&= H(\bar{p}^{\rho}_{x|y}) + \rho \frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho} - H(\bar{p}^{\rho}_{x|y}) \\
&= \rho \frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial \rho}
\end{aligned} \tag{G.21}
$$

□

Lemma 28 $sign\frac{\partial[D(\bar{p}^{\rho}_{xy}\|p_{xy})-H(\bar{p}^{\rho}_{x|y})]}{\partial\rho} = sign(\rho-1)$.

Proof: Using the previous lemma, we get:

$$\frac{\partial D(\bar{p}^{\rho}_{xy}\|p_{xy}) - H(\bar{p}^{\rho}_{x|y})}{\partial\rho} = (\rho-1)\frac{\partial H(\bar{p}^{\rho}_{x|y})}{\partial\rho}$$

Then by Lemma 26, we get the conclusion. □

Lemma 29

$$\rho H(p^{\rho}_{xy}) - (1+\rho)\log(\sum_y\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}) = D(p^{\rho}_{xy}\|p_{xy})$$

Proof: By noticing that $\log(p_{xy}(x,y)) = (1+\rho)[\log(p^{\rho}_{xy}(x,y)) + \log(\sum_{s,t} p_{xy}(s,t)^{\frac{1}{1+\rho}})]$. We have:

$$\begin{aligned}
D(p^{\rho}_{xy}\|p_{xy}) &= -H(p^{\rho}_{xy}) - \sum_{x,y} p^{\rho}_{xy}(x,y)\log(p_{xy}(x,y)) \\
&= -H(p^{\rho}_{xy}) - \sum_{x,y} p^{\rho}_{xy}(x,y)(1+\rho)[\log(p^{\rho}_{xy}(x,y)) + \log(\sum_{s,t} p_{xy}(s,t)^{\frac{1}{1+\rho}})] \\
&= -H(p^{\rho}_{xy}) + (1+\rho)H(p^{\rho}_{xy}) - (1+\rho)\sum_{x,y} p^{\rho}_{xy}(x,y)\log(\sum_{s,t} p_{xy}(s,t)^{\frac{1}{1+\rho}}) \\
&= \rho H(p^{\rho}_{xy}) - (1+\rho)\log(\sum_{s,t} p_{xy}(s,t)^{\frac{1}{1+\rho}}) \qquad \text{(G.22)}
\end{aligned}$$

□

Lemma 30

$$\rho H(\bar{p}^{\rho}_{x|y}) - \log(\sum_y(\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho}) = D(\bar{p}^{\rho}_{xy}\|p_{xy})$$

Proof:

$$\begin{aligned}
D(\bar{p}^{\rho}_{xy}\|p_{xy}) &= \sum_y \sum_x \frac{A(y,\rho)}{B(\rho)} \frac{C(x,y,\rho)}{D(y,\rho)} \log(\frac{\frac{A(y,\rho)}{B(\rho)}\frac{C(x,y,\rho)}{D(y,\rho)}}{p_{xy}(x,y)}) \\
&= \sum_y \sum_x \frac{A(y,\rho)}{B(\rho)} \frac{C(x,y,\rho)}{D(y,\rho)} [\log(\frac{A(y,\rho)}{B(\rho)}) + \log(\frac{C(x,y,\rho)}{D(y,\rho)}) - \log(p_{xy}(x,y))] \\
&= -\log(B(\rho)) - H(\bar{p}^{\rho}_{x|y}) \\
&\quad + \sum_y \sum_x \frac{A(y,\rho)}{B(\rho)} \frac{C(x,y,\rho)}{D(y,\rho)} [\log(D(y,\rho)^{1+\rho}) - \log(C(x,y,\rho)^{1+\rho})] \\
&= -\log(B(\rho)) - H(\bar{p}^{\rho}_{x|y}) + (1+\rho) H(\bar{p}^{\rho}_{x|y}) \\
&= -\log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho}) + \rho H(\bar{p}^{\rho}_{x|y})
\end{aligned}$$

□

Lemma 31

$$H(p^{\rho}_{xy}) = \frac{\partial (1+\rho) \log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial \rho}$$

Proof:

$$\begin{aligned}
&\frac{\partial (1+\rho) \log(\sum_y \sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})}{\partial \rho} \\
&= \log(\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}) - \sum_y \sum_x \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}} \log(p_{xy}(x,y)^{\frac{1}{1+\rho}}) \\
&= -\sum_y \sum_x \frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}} \log(\frac{p_{xy}(x,y)^{\frac{1}{1+\rho}}}{\sum_t \sum_s p_{xy}(s,t)^{\frac{1}{1+\rho}}}) \\
&= H(p^{\rho}_{xy}) \qquad \text{(G.23)}
\end{aligned}$$

□

Lemma 32

$$H(\bar{p}^{\rho}_{x|y}) = \frac{\partial \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})}{\partial \rho}$$

Proof: Notice that $B(\rho) = \sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho}$, and $\frac{\partial B(\rho)}{\partial \rho} = B(\rho) H(\bar{p}^{\rho}_{x|y})$ as shown in Lemma 25. It is clear that:

$$\begin{aligned} \frac{\partial \log(\sum_y (\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}})^{1+\rho})}{\partial \rho} &= \frac{\partial \log(B(\rho))}{\partial \rho} \\ &= \frac{1}{B(\rho)} \frac{\partial B(\rho)}{\partial \rho} \\ &= H(\bar{p}^{\rho}_{x|y}) \end{aligned} \tag{G.24}$$

□

Appendix H

Bounding individual error events for source coding with side-information

In this appendix, we give the proofs for Lemma 13 and Lemma 14. The proofs here are very similar to those for point-to-point source coding problem in Propositions 3 and 4.

H.1 ML decoding: Proof of Lemma 13

In this section, we prove Lemma 13. The argument resembles the proof of Lemma 9 in Appendix F.1.

$$\begin{aligned} p_n(l) &= \sum_{x_1^n, y_1^n} \Pr\left[\exists\, \tilde{x}_1^n \in \mathcal{B}_x(x_1^n) \cap \mathcal{F}_n(l, x_1^n) \text{ s.t. } p_{xy}(\tilde{x}_1^n, y_1^n) \geq p_{xy}(x_1^n, y_1^n)\right] p_{xy}(x_1^n, y_1^n) \\ &\leq \sum_{x_1^n, y_1^n} \min\Big[1, \sum_{\substack{\tilde{x}_1^n \in \mathcal{F}_n(l, x_1^n) \text{s.t.} \\ p_{xy}(x_1^n, y_1^n) \leq p_{xy}(\tilde{x}_1^n, y_1^n)}} \Pr[\tilde{x}_1^n \in \mathcal{B}_s(x_1^n)]\Big] p_{xy}(x_1^n, y_1^n) \end{aligned} \tag{H.1}$$

$$
\begin{aligned}
\leq & \sum_{y_1^n} \sum_{x_1^{l-1}, x_l^n} \min\Big[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ p_{xy}(x_l^n, y_l^n) < p_{xy}(\tilde{x}_l^n, y_l^n)}} 2 \times 2^{-(n-l+1)R}\Big] p_{xy}(x_1^{l-1}, y_1^{l-1}) p_{xy}(x_l^n, y_l^n) \\
\leq & 2 \times \sum_{y_l^n} \sum_{x_l^n} \min\Big[1, \sum_{\substack{\tilde{x}_l^n \text{ s.t.} \\ p_{xy}(x_l^n, y_l^n) < p_{xy}(\tilde{x}_l^n, y_l^n)}} 2^{-(n-l+1)R}\Big] p_{xy}(x_l^n, y_l^n) \\
= & 2 \times \sum_{y_l^n} \sum_{x_l^n} \min\Big[1, \sum_{\tilde{x}_l^n} \mathit{1}[p_{xy}(\tilde{x}_l^n, y_l^n) > p_{xy}(x_l^n, y_l^n)] 2^{-(n-l+1)R}\Big] p_{xy}(x_l^n, y_l^n) \\
\leq & 2 \times \sum_{y_l^n} \sum_{x_l^n} \min\left[1, \sum_{\tilde{x}_l^n} \min\left[1, \frac{p_{xy}(\tilde{x}_l^n, y_l^n)}{p_{xy}(x_l^n, y_l^n)}\right] 2^{-(n-l+1)R}\right] p_{xy}(x_l^n, y_l^n) \\
\leq & 2 \times \sum_{y_l^n} \sum_{x_l^n} \left[\sum_{\tilde{x}_l^n} \left[\frac{p_{xy}(\tilde{x}_l^n, y_l^n)}{p_{xy}(x_l^n, y_l^n)}\right]^{\frac{1}{1+\rho}} 2^{-(n-l+1)R}\right]^{\rho} p_{xy}(x_l^n, y_l^n) \\
= & 2 \times \sum_{y_l^n} \sum_{x_l^n} p_{xy}(x_l^n, y_l^n)^{\frac{1}{1+\rho}} \left[\sum_{\tilde{x}_l^n} [p_{xy}(\tilde{x}_l^n, y_l^n)]^{\frac{1}{1+\rho}}\right]^{\rho} 2^{-(n-l+1)\rho R} \\
= & 2 \times \sum_{y} \left[\sum_{x} p_{xy}(x, y)^{\frac{1}{1+\rho}}\right]^{(n-l+1)} \left[\sum_{x} p_{xy}(x, y)^{\frac{1}{1+\rho}}\right]^{(n-l+1)\rho} 2^{-(n-l+1)\rho R} \\
= & 2 \times \sum_{y} \left[\sum_{x} p_{xy}(x, y)^{\frac{1}{1+\rho}}\right]^{(n-l+1)(1+\rho)} 2^{-(n-l+1)\rho R} \\
= & 2 \times 2^{-(n-l+1)\left(\rho R - \log[\sum_y [\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}]^{1+\rho}]\right)}
\end{aligned} \tag{H.2}
$$

for all $\rho \in [0, 1]$. Every step of the proof follows the same argument in the proof of Lemma 9. Finally, we optimize the exponent over $\rho \in [0, 1]$ in (H.2) to prove the lemma. □

H.2 Universal decoding: Proof of Lemma 14

In this section, we give the details of the proof of Lemma 14. The argument resembles that in the proof of Lemma 11 in Appendix F.2.

The error probability $p_n(l)$ can be thought as starting from (H.1) with the condition $H(\tilde{x}_l^n, y_l^n) < H(x_l^n, y_l^n)$ substituted for $p_{xy}(\tilde{x}_l^n, y_l^n) > p_{xy}(x_l^n, y_l^n)$, we get

$$p_n(l) \leq \sum_{P^{n-l}} \sum_{V^{n-l}} \sum_{y_l^n \in \mathcal{T}_{P^{n-l}}} \sum_{x_l^n \in \mathcal{T}_{V^{n-l}(y_l^n)}} \min\Bigg[1, \sum_{\substack{\tilde{V}^{n-l} \text{ s.t.} \\ H(P^{n-l} \times \tilde{V}^{n-l}) < \\ H(P^{n-l} \times V^{n-l})}} \sum_{\tilde{x}_l^n \in \mathcal{T}_{\tilde{V}^{n-l}}(\tilde{y}_l^n)} 2 \times 2^{-(n-l+1)R}\Bigg] p_{xy}(x_l^n, y_l^n) \tag{H.3}$$

In (H.3) we enumerate all the source sequences in a way that allows us to focus on the types of the important subsequences. We enumerate the possibly misleading candidate sequences in terms of their suffixes types. We restrict the sum to those sequences $\tilde{x}_l^n$ that could lead to mistaken decoding, defining the notation $H(P^{n-l} \times V^{n-l})$, which is the joint entropy of the type P^{n-l} with V-shell V^{n-l}.

Note that the summations within the minimization in (H.3) do not depend on the arguments within these sums. Thus, we can bound this sum separately to get a bound on the number of possibly misleading source sequences $\tilde{x}_l^n$.

$$\begin{aligned}
\sum_{\substack{\tilde{V}^{n-l} \text{ s.t.} \\ H(P^{n-l} \times \tilde{V}^{n-l}) < \\ H(P^{n-l} \times V^{n-l})}} \sum_{\tilde{x}_l^n \in \mathcal{T}_{\tilde{V}^{n-l}}(\tilde{y}_l^n)} &\leq \sum_{\substack{\tilde{V}^{n-l} \text{ s.t.} \\ H(P^{n-l} \times \tilde{V}^{n-l}) < \\ H(P^{n-l} \times V^{n-l})}} |\mathcal{T}_{\tilde{V}^{n-l}}(\tilde{y}_l^n)| \\
&\leq \sum_{\substack{\tilde{V}^{n-l} \text{ s.t.} \\ H(P^{n-l} \times \tilde{V}^{n-l}) < \\ H(P^{n-l} \times V^{n-l})}} 2^{(n-l+1)H(\tilde{V}^{n-l}|P^{n-l})} \\
&\leq \sum_{\tilde{V}^{n-l}} 2^{(n-l+1)H(V^{n-l}|P^{n-l})} \qquad \text{(H.4)} \\
&\leq (n-l+2)^{|\mathcal{X}||\mathcal{Y}|} 2^{(n-l+1)H(V^{n-l}|P^{n-l})} \qquad \text{(H.5)}
\end{aligned}$$

Every inequality is obvious by following the same argument in the proof of Lemma 11. (H.4) is true because $H(P^{n-l} \times \tilde{V}^{n-l}) < H(P^{n-l} \times V^{n-l})$ is equivalent to $H(\tilde{V}^{n-l}|P^{n-l}) < H(V^{n-l}|P^{n-l})$.

We substitute (H.5) into (H.3) and pull out the polynomial term, giving

$$\begin{aligned}
p_n(l) \leq & (n-l+2)^{|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-l}} \sum_{V^{n-l}} \sum_{y_l^n \in \mathcal{T}_{P^{n-l}}} \sum_{x_l^n \in \mathcal{T}_{V^{n-l}(y_l^n)}} \min \Big[1, \\
& 2 \times 2^{-(n-l+1)(R-H(V^{n-l}|P^{n-l}))} \Big] p_{xy}(x_l^n, y_l^n) \\
\leq & (n-l+2)^{|\mathcal{X}||\mathcal{Y}|} \sum_{P^{n-l}} \sum_{V^{n-l}} 2^{-(n-l+1)\max\{0, R-H(V^{n-l}|P^{n-l})\}} \\
& \times 2^{-(n-l+1)D(P^{n-l} \times V^{n-l} \| p_{xy})} \\
\leq & 2 \times (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \times 2^{-(n-l+1)\inf_{\tilde{x},\tilde{y}}\{D(p_{\tilde{x},\tilde{y}} \| p_{xy}) + |R - H(V^{n-l}|P^{n-l})|^+\}} \\
= & 2 \times (n-l+2)^{2|\mathcal{X}||\mathcal{Y}|} \times 2^{-(n-l+1)E_{si}^{lower}(R)}
\end{aligned}$$

All steps are obvious by following the same argument in the proof of Lemma 11.

□

Appendix I

Proof of Corollary 1

I.1 Proof of the lower bound

From Theorem 6, we know that

$$\begin{aligned} E_{si}(R) &\geq E_{si}^{lower}(R) \\ &= \min_{q_{\mathsf{xy}}}\{D(q_{\mathsf{xy}}\|p_{\mathsf{xy}}) + |R - H(q_{\mathsf{x}|\mathsf{y}})|^+\} \quad \text{(I.1)} \\ &= \max_{\rho\in[0,1]} \rho R - \log[\sum_y [\sum_x p_{\mathsf{xy}}(x,y)^{\frac{1}{1+\rho}}]^{1+\rho}] \quad \text{(I.2)} \end{aligned}$$

where the equivalence between (I.1) and (I.2) is shown by a Lagrange multiplier argument in [39] and detailed in Appendix G. Next we show that (I.2) is equal to the random source coding error exponent $E_r(R, p_{\mathsf{s}})$ defined in (A.5) by replacing p_{x} with p_{s}.

$$\begin{aligned} E_{si}^{lower}(R) &=_{(a)} \max_{\rho\in[0,1]} \rho R - \log[\sum_y [\sum_x p_{\mathsf{xy}}(x,y)^{\frac{1}{1+\rho}}]^{1+\rho}] \\ &=_{(b)} \max_{\rho\in[0,1]} \rho R - \log[\sum_y [\sum_x (\frac{p_{\mathsf{xy}}(x|y)}{|\mathcal{S}|})^{\frac{1}{1+\rho}}]^{1+\rho}] \\ &=_{(c)} \max_{\rho\in[0,1]} \rho R - \log[\frac{1}{|\mathcal{S}|}\sum_y [\sum_s p_{\mathsf{s}}(s)^{\frac{1}{1+\rho}}]^{1+\rho}] \\ &=_{(d)} \max_{\rho\in[0,1]} \rho R - \log[[\sum_s p_{\mathsf{s}}(s)^{\frac{1}{1+\rho}}]^{1+\rho}] \\ &=_{(e)} E_r(R, p_{\mathsf{s}}) \end{aligned}$$

(a) is by definition. (b) is because the marginal distribution of y is uniform on $\mathcal{S}$, thus $p_{\mathsf{xy}}(x,y) = p_{\mathsf{xy}}(x|y)p_{\mathsf{y}}(y) = p_{\mathsf{xy}}(x|y)\frac{1}{|\mathcal{S}|}$. (c) is true because $\mathsf{x} = \mathsf{y} \oplus \mathsf{s}$. (d) is obvious. (e) is by the definition of the random coding error exponent in Equation (A.5). □

I.2 Proof of the upper bound

From Theorem 7, we know that

$$\begin{aligned}
E_{si}(R) &\leq_{(a)} E_{si}^{upper}(R) \\
&=_{(b)} \min\{\inf_{q_{xy},\alpha\geq 1: H(q_{x|y})>(1+\alpha)R}\{\frac{1}{\alpha}D(q_{xy}\|p_{xy})\}, \\
&\qquad \inf_{q_{xy},1\geq\alpha\geq 0: H(q_{x|y})>(1+\alpha)R}\{\frac{1-\alpha}{\alpha}D(q_x\|p_x)+D(q_{xy}\|p_{xy})\}\} \\
&\leq_{(c)} \inf_{q_{xy},1\geq\alpha\geq 0: H(q_{x|y})>(1+\alpha)R}\{\frac{1-\alpha}{\alpha}D(q_x\|p_x)+D(q_{xy}\|p_{xy})\} \\
&\leq_{(d)} \inf_{q_{xy},1\geq\alpha\geq 0: H(q_{x|y})>(1+\alpha)R}\{D(q_{xy}\|p_{xy})\} \\
&\leq_{(e)} \inf_{q_{xy}: H(q_{x|y})>R}\{D(q_{xy}\|p_{xy})\} \\
&=_{(f)} E_{si,b}^{upper}(R)
\end{aligned} \tag{I.3}$$

where (a) and (b) are by Theorem 7, (c) is obvious. (d) is true because $D(q_x\|p_x) \geq 0$. (e) is true because $\{(q_{xy},0)|H(q_{x|y}) > R\} \subseteq \{(q_{xy},\alpha)|H(q_{x|y}) > (1+\alpha)R\}$. (f) is by the definition of the upper bound on block coding error exponent in Theorem 13. By using the other definition of $E_{si,b}^{upper}(R)$ in Theorem 13, we have:

$$\begin{aligned}
E_{si}^{upper}(R) &=_{(a)} \max_{\rho\in[0,\infty]} \rho R - \log[\sum_y[\sum_x p_{xy}(x,y)^{\frac{1}{1+\rho}}]^{1+\rho}] \\
&=_{(b)} \max_{\rho\in[0,\infty]} \rho R - \log[\sum_y[\sum_x (\frac{p_{xy}(x|y)}{|\mathcal{S}|})^{\frac{1}{1+\rho}}]^{1+\rho}] \\
&=_{(c)} \max_{\rho\in[0,\infty]} \rho R - \log[\frac{1}{|\mathcal{S}|}\sum_y[\sum_s p_{\mathsf{s}}(s)^{\frac{1}{1+\rho}}]^{1+\rho}] \\
&=_{(d)} \max_{\rho\in[0,\infty]} \rho R - \log[[\sum_s p_{\mathsf{s}}(s)^{\frac{1}{1+\rho}}]^{1+\rho}] \\
&=_{(e)} E_{s,b}(R, p_{\mathsf{s}})
\end{aligned}$$

where (a)-(d) follows exactly the same arguments as those in the previous section I.1. (e) is by the definition of block coding error exponent in (A.4). □

Appendix J

Proof of Theorem 8

The proof here is almost identical to that for the delay constrained lossless source coding error exponent $E_s(R)$ in Chapter 2. In the converse part, we use a different argument which is *genie* based [67]. Readers might find that proof interesting.

J.1 Achievability of $E_{ei}(R)$

In this section, we introduce a universal coding scheme which achieves the delay constrained error exponent in Theorem 8. The coding scheme only depends on the size of the alphabet $|\mathcal{X}|, |\mathcal{Y}|$, not the distribution of the source. We first describe our universal coding scheme.

A block-length N is chosen that is much smaller than the target end-to-end delays, while still being large enough. Again we are interested in the performance with asymptotically large delays Δ, For a discrete memoryless source X, side information Y and large block-length N, the best possible variable-length code is given in [29] and consists of two stages: first describing the type of the block $\vec{x}_i, \vec{y}_i$ using at most $O(|\mathcal{X}||\mathcal{Y}| \log_2 N)$ bits and then describing which particular realization has occurred by using a variable $NH(\vec{x}_i|\vec{y}_i)$ bits where $\vec{x}_i$ is the i^{th} block of length N and $H(\vec{x}_i|\vec{y}_i)$ is the empirical conditional entropy of sequence $\vec{x}_i$ given $\vec{y}_i$. The overhead $O(|\mathcal{X}||\mathcal{Y}| \log_2 N)$ is asymptotically negligible and the code is also universal in nature. It is easy to verify that the average code length: $\lim_{N\to\infty} \frac{E_{p_{xy}}(l(\vec{x},\vec{y}))}{N} = H(p_{x|y})$ This code is obviously a prefix-free code. Write $l(\vec{x}_i, \vec{y}_i)$ as

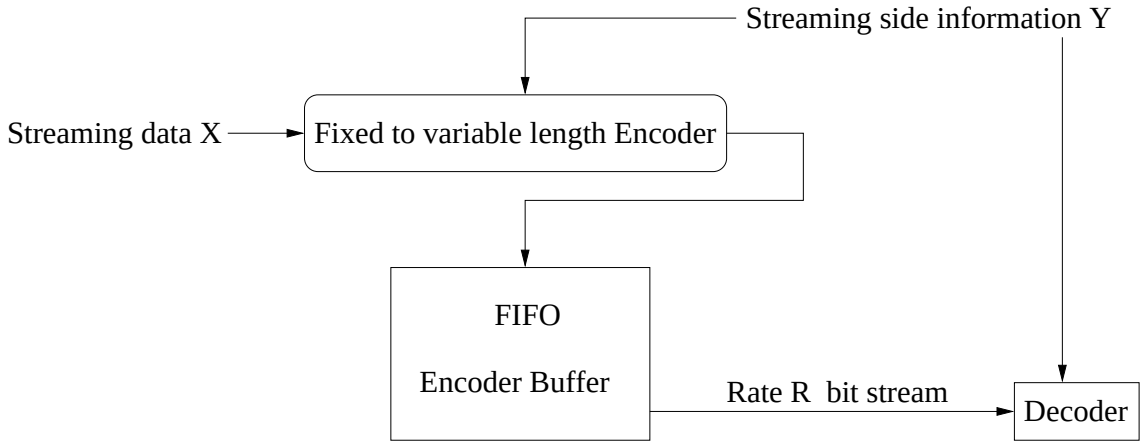

Figure J.1. A universal delay constrained source coding system with both encoder and decoder side-information

the length of the codeword for $\vec{x}_i, \vec{y}_i$, write $2^{N\epsilon_N}$ as the number of types in $\mathcal{X}^N \times \mathcal{Y}^N$ then:

$$NH(\vec{x}_i|\vec{y}_i) \leq l(\vec{x}_i, \vec{y}_i) = N(H(\vec{x}_i|\vec{y}_i) + \epsilon_N) \tag{J.1}$$

where $0 < \epsilon_N \leq \frac{|\mathcal{X}||\mathcal{Y}|\log_2(N+1)}{N}$ goes to 0 as N goes to infinity. The binary sequence describing the source is fed to the FIFO buffer described in Figure J.1. Notice that if the buffer is empty, the output of the buffer can be gibberish binary bits. The decoder simply discards these meaningless bits because it is aware that the buffer is empty.

Proposition 13 *For the iid source $\sim p_{xy}$ using the universal causal code described above, for all ϵ, there exists $K < \infty$, s.t. for all t, Δ:*

$$\Pr(\vec{x}_t \neq \hat{\vec{x}}_t((t+\Delta)N) \leq K2^{-\Delta N(E_{ei}(R)-\epsilon)}$$

where $\hat{\vec{x}}_t((t+\Delta)N)$ is the estimate of $\vec{x}_t$ at time $(t+\Delta)N$. Before the proof, we have the following lemma to bound the probabilities of atypical source behavior.

Lemma 33 *(Source atypicality) for all $\epsilon > 0$, block length N large enough, there exists $K < \infty$, s.t. for all n, if $r < \log_2|\mathcal{X}|$:*

$$\Pr(\sum_{i=1}^{n} l(\vec{x}_i, \vec{y}_i) > nNr) \leq K2^{-nN(E_{ei,b}(r)-\epsilon)} \tag{J.2}$$

Proof: Only need to show the case for $r > H(p_{x|y})$. By the Cramér's theorem[31], for all $\epsilon_1 > 0$, there exists K_1, such that

$$\Pr(\sum_{i=1}^{n} l(\vec{x}_i, \vec{y}_i) > nNr) = \Pr(\frac{1}{n}\sum_{i=1}^{n} l(\vec{x}_i, \vec{y}_i) > Nr) \leq K_1 2^{-n(\inf_{z>Nr} I(z)-\epsilon_1)} \tag{J.3}$$

where the rate function $I(z)$ is [31]:

$$I(z) = \sup_{\rho\in\mathcal{R}}\{\rho z - \log_2(\sum_{(\vec{x},\vec{y})\in\mathcal{X}^N\times\mathcal{Y}^N} p_{xy}(\vec{x},\vec{y})2^{\rho l(\vec{x},\vec{y})})\} \tag{J.4}$$

$$Write \quad I(z,\rho) = \rho z - \log_2(\sum_{(\vec{x},\vec{y})\in\mathcal{X}^N\times\mathcal{Y}^N} p_{xy}(\vec{x},\vec{y})2^{\rho l(\vec{x},\vec{y})})$$

Obviously $I(z,0) = 0$, $z > Nr > NH(p_{x|y})$ thus for large N: .

$$\frac{\partial I(z,\rho)}{\partial\rho}|_{\rho=0} = z - \sum_{(\vec{x},\vec{y})\in\mathcal{X}^N\times\mathcal{Y}^N} p_{xy}(\vec{x},\vec{y})l(\vec{x},\vec{y}) \geq 0$$

By the Hölder inequality, for all ρ_1, ρ_2, and for all $\theta \in (0,1)$:

$$\begin{aligned}(\sum_i p_i 2^{\rho_1 l_i})^\theta (\sum_i p_i 2^{\rho_2 l_i})^{(1-\theta)} &\geq \sum_i (p_i^\theta 2^{\theta\rho_1 l_i})(p_i^{(1-\theta)} 2^{(1-\theta)\rho_2 l_i}) \\ &= \sum_i p_i 2^{(\theta\rho_1+(1-\theta)\rho_2)l_i}\end{aligned}$$

This shows that $\log_2(\sum_{(\vec{x},\vec{y})\in\mathcal{X}^N\times\mathcal{Y}^N} p_{xy}(\vec{x},\vec{y})2^{\rho l(\vec{x},\vec{y})})$ is a convex function on ρ, thus $I(z,\rho)$ is a concave $\cap$ function on ρ for fixed z. Then $\forall z > 0$, $\forall \rho < 0$, $I(z,\rho) < 0$, which means that the ρ to maximize $I(z,\rho)$ is positive. This implies that $I(z)$ is monotonically increasing with z and obviously $I(z)$ is continuous. Thus

$$\inf_{z>Nr} I(z) = I(Nr) \tag{J.5}$$

Using the upper bound on $l(\vec{x},\vec{y})$ in (J.1):

$$\begin{aligned}\log_2(\sum_{(\vec{x},\vec{y})\in\mathcal{X}^N\times\mathcal{Y}^N} p_{xy}(\vec{x},\vec{y})2^{\rho l(\vec{x},\vec{y})}) &\leq \log_2(\sum_{q_{xy}\in\mathcal{T}^N} 2^{-ND(q_{xy}\|p_{xy})} 2^{\rho(\epsilon_N + NH(q_{x|y}))}) \\ &\leq 2^{N\epsilon_N} 2^{-N\min_q\{D(q_{xy}\|p_{xy})-\rho H(q_{x|y})-\rho\epsilon_N\}}) \\ &= N\big(-\min_q\{D(q_{xy}\|p_{xy}) - \rho H(q_{x|y}) - \rho\epsilon_N\} + \epsilon_N\big)\end{aligned}$$

where $0 < \epsilon_N \leq \frac{|\mathcal{X}||\mathcal{Y}|\log_2(N+1)}{N}$ goes to 0 as N goes to infinity.

$\mathcal{T}^N$ is the set of all types of $\mathcal{X}^N \times \mathcal{Y}^N$.

Substitute the above inequalities to $I(Nr)$ defined in (J.4):

$$I(Nr) \geq N\big(\sup_{\rho}\{\min_{q}\rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy}\|p_{xy})\} - \epsilon_N\big) \tag{J.6}$$

We next show that $I(Nr) \geq N(E_{ei,b}(r) + \epsilon)$ where ϵ goes to 0 as N goes to infinity. This can be proved by a somewhat complicated Lagrange multiplier method also used in [12]. We give a new proof based on the existence of saddle point of the min-max function. Define

$$f(q, \rho) = \rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy}\|p_{xy})$$

Obviously, for fixed q, $f(q, \rho)$ is a linear function of ρ, thus concave $\cap$. Also for fixed ρ, $f(q, \rho)$ is a convex function of q. Define $g(u) = \min_q \sup_\rho(f(q, \rho) + \rho u)$, it is enough [10] to show that $g(u)$ is finite in the neighbor of $u = 0$ to establish the existence of the saddle point.

$$\begin{aligned} g(u) &=_{(a)} \min_q \sup_\rho(f(q, \rho) + \rho u) \\ &=_{(b)} \min_q \sup_\rho(\rho(r - H(q_{x|y}) - \epsilon_N + u) + D(q_{xy}\|p_{xy})) \\ &\leq_{(c)} \min_{q:H(q_{x|y})=r-\epsilon_N+u} D(q_{xy}\|p_{xy}) \\ &\leq_{(d)} \infty \end{aligned} \tag{J.7}$$

(a) and (b) are by definition. (c) is true because $H(p_{x|y}) < r < \log_2|\mathcal{X}|$, thus for very small ϵ_N and u, $H(p_{x|y}) < r - \epsilon_N + u < \log_2|\mathcal{X}|$. Thus there exists distribution q, s.t. $H(q_{x|y}) = r - \epsilon_N + u$. (d) is true because the assumption on p_{xy} that the marginal $p_x(x) > 0$ for all $x \in \mathcal{X}$. Thus we proved the existence of the saddle point of $f(q, \rho)$.

$$\sup_\rho\{\min_q f(q, \rho)\} = \min_q\{\sup_\rho f(q, \rho)\} \tag{J.8}$$

Notice that if $H(q_{x|y}) \neq r + \epsilon_N$, then ρ can be chose to be arbitrarily large or small to make $\rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy}\|p_{xy})$ arbitrarily large. Thus the q to minimize $\sup_\rho \rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy}\|p_{xy})$ satisfies that $r - H(q_{x|y}) - \epsilon_N = 0$, so:

$$\begin{aligned} \min_q\{\sup_\rho \rho(r - H(q_{x|y}) - \epsilon_N) + D(q_{xy}\|p_{xy})\} &=_{(a)} \min_{q:H(q_{x|y})=r+\epsilon_N}\{D(q_{xy}\|p_{xy})\} \\ &=_{(b)} \min_{q:H(q_{x|y})\geq r+\epsilon_N}\{D(q_{xy}\|p_{xy})\} \\ &=_{(c)} E_{ei,b}(r + \epsilon_N) \end{aligned} \tag{J.9}$$

(a) is following the argument above, (b) is true because the distribution q to minimize the KL divergence is always on the boundary of the feasible set [26] when p is not in the

feasible set. (c) is definition. Combining (J.6), (J.8) and (J.9), and let N be sufficiently big, thus ϵ_N sufficiently small, and notice that $E_{ei}(r)$ is continuous in r, we get the the desired bound in (J.2). □

Now we are ready to prove Proposition 13.

Proof: We give an upper bound on the decoding error on $\vec{x}_t$ at time $(t+\Delta)N$. At time $(t+\Delta)N$, the decoder *cannot* decode $\vec{x}_t$ with *zero* error probability iff the binary strings describing $\vec{x}_t$ are *not* all out of the buffer. Since the encoding buffer is FIFO, this means that the number of outgoing bits from some time t_1 to $(t+\Delta)N$ is less than the number of the bits in the buffer at time t_1 plus the number of incoming bits from time t_1 to time tN. Suppose that the buffer is last empty at time $tN-nN$ where $0 \le n \le t$. Given this condition, the decoding error occurs only if $\sum_{i=0}^{n-1} l(\vec{x}_{t-i}, \vec{y}_{t-i}) > (n+\Delta)NR$. Write l_{max} as the longest code length, $l_{max} \le |\mathcal{X}||\mathcal{Y}| \log_2(N+1) + N \log_2 |\mathcal{X}|$. Then $\Pr(\sum_{i=0}^{n-1} l(\vec{x}_{t-i}, \vec{y}_{t-i}) > (n+\Delta)NR) > 0$ only if $n > \frac{(n+\Delta)NR}{l_{max}} > \frac{\Delta NR}{l_{max}} \triangleq \beta\Delta$

$$
\begin{aligned}
\Pr(\vec{x}_t \neq \vec{x}_t((t+\Delta)N) \quad &\le \quad \sum_{n=\beta\Delta}^{t} \Pr(\sum_{i=0}^{n-1} l(\vec{x}_{t-i}, \vec{y}_{t-i}) > (n+\Delta)NR) \qquad \text{(J.10)} \\
&\le_{(a)} \quad \sum_{n=\beta\Delta}^{t} K_1 2^{-nN(E_{ei,b}(\frac{(n+\Delta)NR}{nN})-\epsilon_1)} \\
&\le_{(b)} \quad \sum_{n=\gamma\Delta}^{\infty} K_2 2^{-nN(E_{ei,b}(R)-\epsilon_2)} + \sum_{n=\beta\Delta}^{\gamma\Delta} K_2 2^{-\Delta N(\min_{\alpha>1}\{\frac{E_{ei,b}(\alpha R)}{\alpha-1}\}-\epsilon_2)} \\
&\le_{(c)} \quad K_3 2^{-\gamma\Delta N(E_{ei,b}(R)-\epsilon_2)} |\gamma\Delta - \beta\Delta| K_3 2^{-\Delta N(E_{ei}(R)-\epsilon_2)} \\
&\le_{(d)} \quad K 2^{-\Delta N(E_{ei}(R)-\epsilon)}
\end{aligned}
$$

where, $K_i's$ and $\epsilon_i's$ are properly chosen real numbers. (a) is true because of Lemma 33. Define $\gamma = \frac{E_{ei}(R)}{E_{ei,b}(R)}$, in the first part of (b), we only need the fact that $E_{ei,b}(R)$ is non decreasing with R. In the second part of (b), we write $\alpha = \frac{n+\Delta}{n}$ and take the α to minimize the error exponent. The first term of (c) comes from the sum of a geometric series. The second term of (c) is by the definition of $E_{ei}(R)$. (d) is by the definition of γ. ■

J.2 Converse

To bound the best possible error exponent with fixed delay, we consider a genie-aided encoder/decoder pair and translate the block-coding error exponent to the delay constrained

error exponent. The arguments are analogous to the "focusing bound" derivation in [67] for channel coding with feedback.

Proposition 14 *For fixed-rate encodings of discrete memoryless sources, it is not possible to achieve an error exponent with fixed-delay better than*

$$\inf_{\alpha>0} \frac{1}{\alpha} E_{ei,b}((\alpha+1)R) \tag{J.11}$$

Proof: For simplicity of exposition, we ignore integer effects arising from the finite nature of Δ, R, etc. For every $\alpha > 0$ and delay Δ, consider a code running over its fixed-rate noiseless channel till time $\frac{\Delta}{\alpha} + \Delta$. By this time, the decoder have committed to estimates for the source symbols up to time $i = \frac{\Delta}{\alpha}$. The total number of bits used during this period is $(\frac{\Delta}{\alpha} + \Delta)R$.

Now consider a genie that gives the encoder access to the first i source symbols at the beginning of time, rather than forcing the encoder to get the source symbols one at a time. Simultaneously, loosen the requirements on the decoder by only demanding correct estimates for the first i source symbols by the time $\frac{\Delta}{\alpha} + \Delta$. In effect, the deadline for decoding the *past* source symbols is extended to the deadline of the i-th symbol itself.

Any lower-bound to the error probability of the new problem is clearly also a bound for the original problem. Furthermore, the new problem is just a fixed-length block-coding problem requiring the encoding of i source symbols into $(\frac{\Delta}{\alpha}+\Delta)R$ bits. The rate per symbol is

$$\begin{aligned}((\frac{\Delta}{\alpha} + \Delta)R)\frac{1}{i} &= ((\frac{\Delta}{\alpha} + \Delta)R)\frac{\alpha}{\Delta} \\ &= (\alpha+1)R\end{aligned}$$

Lemma 12 tells us that such a code has a probability of error that is at least exponential in $iE_{ei,b}((\alpha+1)R)$. Since $i = \frac{\Delta}{\alpha}$, this translates into an error exponent of at most $\frac{E_{ei,b}((\alpha+1)R)}{\alpha}$ with block length Δ.

Since this is true for all $\alpha > 0$, we have a bound on the reliability function $E_{ei}(R)$ with fixed delay Δ:

$$E_{ei}(R) \leq \inf_{\alpha>0} \frac{1}{\alpha} E_{ei,b}((\alpha+1)R)$$

The minimizing α tells how much of the past $(\frac{\Delta}{\alpha})$ is involved in the dominant error event. ■