INFORMATION THEORY WITH LARGE ALPHABETS AND SOURCE CODING ERROR EXPONENTS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Benjamin Gregory Kelly

August 2011

INFORMATION THEORY WITH LARGE ALPHABETS AND SOURCE CODING ERROR EXPONENTS

Benjamin Gregory Kelly, Ph.D.

Cornell University 2011

The early twenty-first century has been referred to as the 'information age'; this appears to be an apt name given the massive amounts of data that are created daily. However, the utility of the data is limited by the tools we possess to extract and manipulate the information contained within. Towards this end, in this thesis we examine some problems concerning classification and communication.

The first problem examined is that of classification in a 'large-alphabet' regime. In this large-alphabet regime, which is motivated by natural language, standard statistical approaches to classification based on chi-squared tests or maximum likelihood are inconsistent. We derive the limit (in terms of alphabet growth rate) beyond which consistent classification is impossible and propose a new consistent test that achieves this limit. We also propose a new classifier which has good empirical performance.

The second problem addressed concerns compression of sources with large alphabets. We first characterize for which alphabet growth rates is universal compression possible. We then study the permitted alphabet growth-rate in the non-universal case in which the goal is to compress a source generated by a known sequence of distributions.

We finally examine error exponents for source coding/compression problems. The error exponent characterizes the optimal exponential decay of the

error probability. For the cases of the Wyner-Ziv and source coding with side information problems we provide new upper and lower bounds on the error exponent. These bounds match for some special cases. We also make connections between source coding error exponents and graph theory and provide new upper bounds on Witsenhausen's rate and complementary graph entropy, two useful quantities from graph theory.

BIOGRAPHICAL SKETCH

Benjamin Kelly was born and raised in Nottingham, England. It was at the University of Nottingham where he earned his B.Sc.(Hons) in Computer Science (and first encountered information theory via some homework exercises on the properties of entropy in the course G51SCI). After a few years teaching high school computer science and running an IT consultancy company he returned to the University of Nottingham in 2004 and obtained his M.Phil. under Prof. David Brailsford. It was whilst working on his M.Phil. that he rediscovered information theory when working on data compression and decided to pursue further studies at Cornell. He entered Cornell in Fall 2006 as Prof. Aaron Wagner's first student. In 2010 he won the IEEE Information Theory Society's Student Best Paper award for his work on universal hypothesis testing; this work forms the basis of Chapter 2.

To Len and Grace Tomlin.

ACKNOWLEDGEMENTS

First and foremost my thanks go to my advisor, Aaron Wagner. Aaron has been a constant source of guidance, motivation, inspiration and optimism throughout my time at Cornell. Aaron exemplifies what it is to be a scholar and I have never failed to be impressed by the breadth and depth of his knowledge of information theory, probability theory and mathematics. Aaron has (attempted to :-)) instill into me the benefits of being rigorous and careful, but at the same time not always to accept the orthodoxy and not to be afraid to try out new ideas and techniques. I couldn't have hoped for a better advisor.

I also wish to thank the other members of my thesis committee. Both Lang Tong and Michael Nussbaum have been enthusiastic about my research work and always willing to make time to discuss any problems I encountered. I am also fortunate to have taken relevant courses from both, the content of which has informed the content of this thesis.

I should also thank the other Cornell professors whose classes I have taken, Camil Muscalu, Ravi Ramakrishna, Marius Ionescu, Anna Scaglione, Terry Fine, Rick Durrett and Gene Hwang. I am especially grateful to Terry for sharing with me his thoughts about life, the universe and everything. I would also like to thank David Brailsford, my master's advisor at the University of Nottingham, and Roland Backhouse, another of my Nottingham professors, for remaining interested in my work. My thanks also go to Scott Coldren, the manager of student services, who ensured I made steady progress towards my Ph.D. and didn't run afoul of the various rules and regulations.

Thanks are due also to my friends and fellow students at Cornell: Yücel Altuğ, Ebad Ahmed, Saif Rahman, Amine Laourine, Aaron Lei, Jason Li, Sarah Iams, Matt Ezovski, Josh Gabet, Ipek Ozil, (Dr.) Frank Ciaramello, Alireza Vahid

and Ilan Shomorony. I am particular grateful to Yücel for the many interesting conversations we had about our research over the past few years. Also to the 'senior' students who graduated shortly after I arrived, Drs. Parv Venkitasubramaniam, Anima Anandkumar, Saswat Misra and Oliver Kosut, thanks for setting such a good example.

I wish to thank my mum Anita Tomlin and grandparents Len and Grace Tomlin for encouraging me and helping me accomplish everything I set out to do. Thanks to Lala Stone and Angela Natividad for helping me learn about life. Thanks to my friends Joe Wildish, Mark Hills and Sam Bloor for staying in touch even as I disappeared to a different continent. And finally I thank Emily Christensen for putting up with me for the past two years and her parents Maureen and Andy for providing me with PG Tips and scones.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

> There were 5 exabytes of information created between the dawn of civilization through 2003 ... but that much information is now created every two days, and the pace is increasing.
>
> Eric Schmidt, CEO Google, *Techonomy conference, August 2010*.

1.1 Motivation and Overview

The sheer volume of information produced by mankind today presents many challenges to communication engineers, computer scientists, and statisticians, such as how to efficiently store and transmit the data, and how to use the data to make accurate predictions and decisions. The fields of information theory and statistics are especially suited to providing answers to these problems since in both fields, it is often assumed that a very large (infinite) amount of data is available.

There is, of course, a sharp distinction between *very large* and *infinite* amounts of data. It is not necessarily the case that studying an abstract model, where the number of observations goes to infinity, says anything about the problem faced by the engineer, who may be asked to design a solution for a problem in which only one hundred observations are available. Fortunately, it turns out that in many cases the solutions and insights provided by studying these abstract models in fact work well when applied to real-world problems. This can be seen, for example, in the area of channel coding, where practical error correct-

ing codes such as LDPC (which approach the limit established by Shannon) now exist; or in the area of data compression where algorithms such as Lempel-Ziv can compress a source at rates close to the entropy fundamental limit. Loosely speaking, the asymptotic analysis establishes the limits of what can be accomplished, e.g. how much data can be sent over a particular noisy channel or by how much we can compress a file, and then with the limits established, begins the search for practical schemes.

For a typical information processing problem any practical scheme can be split into three phrases. First the raw data is acquired, say by some sensing mechanism; the data is then stored, or perhaps transmitted and stored at a remote site; finally the data is processed into some usable form. In this thesis we focus on the final two phases, and in particular we specialize and study the problems of compression (storage and transmission) and classification (processing).

We first study compression and classification of large-alphabet sources. As explained in the next section, the large-alphabet model captures some asymptotic properties present in natural language data that are not captured by the conventional model often used. Natural language is an important class of raw data and covers a wide spectrum of sources, including blogs, webpages and books.

The final two chapters of the thesis examine compression/transmission of data in the presence of correlated 'side information'. The goal is to characterize the error exponent, the speed of the exponential decay of the error probability, which allows the performance of various schemes to be compared to a fundamental limit.

1.2 Classification with Large Alphabets

A fundamental problem when dealing with natural language data is that of classification. In its simplest form the problem is as follows: a classifier is given a document and has to decide whether the document is about topic one or topic two. As an example, the classifier could be an email client, and topic one is "spam" and topic two is "not spam"; or the classifier could be a search engine bot deciding whether a webpage is written in "French" or "English"; or the classifier could be trying to decide which of two authors wrote a particular text.

A statistical formalisation of the problem is to suppose that the document is a sequence of words $\mathbf{Z} = Z_1, \ldots, Z_n$, and is the output of a memoryless source with distribution p or q, and that we (the classifier) know p and q. The optimum solution to this problem was given by Neyman and Pearson [1]. In the case where the document length goes to infinity, it can be shown that the Neyman-Pearson solution is *consistent*, i.e. has classification error tending to zero [2]. However, a more practically relevant scenario is when the underlying distributions are not available to us, but instead we have access to training data $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{X}$ is known to be generated according to the distribution P (topic one) and $\mathbf{Y}$ generated according Q (topic two). We are then given the third sequence $\mathbf{Z}$ and we perform a binary classification (i.e. a hypothesis test), to decide whether $\mathbf{Z}$ is generated by according to topic one or topic two.

One model for this problem is to suppose that $\mathbf{X} = X_1^n$ is a realization of a discrete memoryless source (DMS) emitting symbols with some fixed, but unknown, distribution p on a finite alphabet $\mathcal{A}$ (and similarly $\mathbf{Y} = Y_1^n$ is generated by a DMS with a different unknown distribution q). The problem is then to

decide whether $\mathbf{Z} = Z_1^n$ was generated by distribution p or distribution q, using only $\mathbf{X}$ and $\mathbf{Y}$. The classical information-theoretic approach is to let the blocklength, n, increase so that we see longer realizations, and be satisfied by a classifier that performs well in the limit as n goes to infinity.

For certain scenarios this classical asymptotic is inappropriate. For example in natural language, with words as our base symbols, $\mathbf{X}$ and $\mathbf{Y}$ are strings containing n words each generated according to p and q. Studies of English text [3] however, suggest that 1) as the blocklength grows, so does the number of words we encounter, *without bound*; and 2) English text tends to comprise a large number of words that occur $\Theta(1)$ times. Yet in the traditional asymptotic with a fixed and finite alphabet, the law of large numbers (LLN) applies, implying that all words will eventually appear and the count of any word will increase without bound. Notice that this observation precludes the use of the Zipf-Mandelbrot distribution [4, 5], often used to model (ranked) word frequencies, because as the blocklength tends to infinity, a string generated according to this distribution would still be dominated by $\Theta(1)$ words appearing $\Theta(n)$ times. The presence of a LLN is roughly equivalent to being able to "learn" the underlying distributions from the data via the convergence of empirical distributions, and can itself be another reason to reject the asymptotic if such an assumption is unrealistic for the application. Note that if we model language with some fixed-order Markov chain, similar issues arise.

1.2.1 Contributions

In Chapter 2 we investigate the classification problem in an alternative asymptotic, where the (discrete) alphabet and underlying distributions generating the data can vary with n. To tackle the problem we formulate it as a sequence of composite[1] binary hypothesis testing problems and ask under what conditions on the distributions p_n, q_n and alphabet $\mathcal{A}_n$ is it possible to have *universally consistent* tests, i.e. a sequence of tests (one for each n) that asymptotically makes no error for any sequence of pairs of distributions on $\mathcal{A}_n$. Note that this problem is non-trivial because here, unlike in the classical asymptotic, the empirical distributions of the test and training data need not converge to the underlying distributions.

Our primary focus is the case in which the underlying distributions belong to the class of α-large-alphabet distributions, i.e. distributions whose underlying symbol probabilities are all order $n^{-\alpha}$ and alphabet size order is order n^{α} (see Def. 2.1, Sect 2.1 for a precise definition). For these sources we provide a simple test and prove that it is universally consistent when $0 \leq \alpha < 2$. We also show that universally consistent classification for these sources is impossible when $\alpha \geq 2$. We also prove that two commonly used tests from classical statistics, the chi-squared test and generalized likelihood ratio test (GLRT), are universally consistent for $0 \leq \alpha < 1$, but both tests fail when $\alpha = 1$.

Our study of α-large-alphabet sources offers insights into the hypothesis testing problem for inhomogeneous sources (i.e. non α-large-alphabet sources whose symbol probabilities are arbitrary) with growing alphabets. Firstly,

[1]Using the nomenclature from statistics, a hypothesis is *simple* if the distribution is fully known and otherwise we say the hypothesis is *composite*.

our results show that universally consistent tests for up-to sub-linear alphabet growth exist. Secondly, our converse result implies that testing for arbitrary sources is not possible when the underlying alphabet grows quadratically or faster. Finally, we illustrate that a key problem in classifying inhomogeneous data concerns how to handle symbols whose probabilities are of different orders. The chi-squared test and GLRT employ a kind of normalization, which attempts to put the differences between the symbol counts in the data on the same scale. Yet, for α-large alphabet sources these differences are naturally on the same scale and we show that this normalization can cause a systematic inconsistency. Our new test relies solely on the unnormalized counts, and we show that for inhomogeneous data our test is inconsistent precisely due to its lack of normalization.

We show by proving that when given an infinite amount of training data (i.e. the classifier exactly knows the underlying distributions p_n and q_n) consistent testing is possible for any rate of alphabet growth; we also provide an achievable error exponent.

We conclude with some observations on classification of general sources (i.e. beyond the α-large-alphabet model) and propose a practical classification algorithm for this problem.

1.2.2 Related Work

The case of hypothesis testing between fixed distributions on a finite alphabet has been well studied. For this simple-versus-simple case, a fundamental result on the existence of optimum tests is due to Neyman and Pearson, [1]; Cher-

noff [6, 7] also provides exponential error guarantees. For the simple-versus-composite case, a key result concerning the problem of asymptotically optimum tests (in an error exponent sense) is Hoeffding [8].

The composite-versus-composite case with fixed distributions on finite alphabets has also received some attention. The problem of determining a test with a prescribed exponential error decay under one hypothesis and that is uniformly most powerful under the other is considered by Gutman [9] (see also Ziv [10]). Feder and Merhav [11] propose a "competitive minimax" approach, in which one minimizes the worst case ratio between the probability of error of a universal test and the minimum probability of error attainable when the distributions are known.

For the case of growing alphabets, the existence of consistent tests for the simple-versus-composite problem is studied by Barron [12], Paninski [13] and Ermakov [14]. The works [12, 13] also address the converse problem of determining the the smallest growth rate beyond which (respectively) uniformly exponentially consistent and consistent tests do not exist.

An alternate line of investigation into the simple-versus-composite case with growing alphabets studied the Pitman and Bahadur efficiencies of the likelihood and chi-square tests [15, 16]. Moderate and large deviation results for these statistics in the same regime are also available [17]. In [18, Ch.4 §3] Read and Cressie study the power divergence family with growing alphabets, which includes the chi-square and likelihood tests as members; the Bahadur efficiency of this family with growing alphabets is investigated in [19].

The composite-versus-composite case with growing alphabets is addressed

in limited form by Wagner et al. [20], who develop a probability estimator for the "rare-events" regime where underlying probabilities are all order $\Theta(n^{-1})$ and therefore alphabet size is order $\Theta(n)$. Other practical approaches may also be taken, see for example Orlitsky-Santhanam-Zhang (OSZ) [21, 22], support vector machines [23], and techniques from pattern recognition and machine learning [24].

1.3 Compression of Large Alphabet Sources

Compression of a sequence of independent and identically distributed (i.i.d.) random variables is arguably one of most basic problems in information theory. If we suppose that p is a probability mass function on some finite alphabet $\mathcal{A}$, then the entropy of the source, $H(p)$, specifies the fewest number of symbols required to represent a source $X^n \sim p^n$. In the fixed-rate setting, this is accomplished via the specification of a block encoder that maps source sequences of length n to some fixed message set, i.e. $f_n : \mathcal{A}^{\times n} \to \mathcal{M}_n$, along with a decoder $g_n : \mathcal{M}_n \to \mathcal{A}^{\times n}$ that inverts this mapping. Shannon [25] showed that the error probability, $\Pr(g_n(f_n(X^n)) \neq X^n)$, can be made arbitrarily small provided that n is sufficiently large and $n^{-1} \log |\mathcal{M}_n| > H(p)$. A converse result states that if $n^{-1} \log |\mathcal{M}_n| < H(p)$ then the probability of error must remain bounded away from zero.

Yet many practical compression problems do not satisfy the hypotheses of this result. There are two reasons for this. First, the result makes the unrealistic assumption that the underlying distribution is known *a priori*; in practice we are often provided with X^n and asked to compress it as well as we can. It is there-

fore natural to seek instead compression schemes that can compress a source generated by *any* underlying distribution in some reasonably-large class. The second reason is that, as mentioned in the previous section, the statistical model on which this result is based is often a poor fit for real data. Many existing algorithms choose to bypass working on the source's natural alphabet by mapping each source symbol onto multiple symbols in a smaller alphabet, such as bits. The disadvantage of this approach, however, is that the dependence of the source may become very long range.

In Chapter 3 we consider fixed-rate universal compression of general large-alphabet sources. We suppose that we are given a sequence of alphabets, $\{\mathcal{A}_n\}$, and distributions on those alphabets, $\{p_n\}$, and observe a source X^n generated i.i.d. according to the nth distribution on the nth alphabet. We determine when there exist codes that can compress any i.i.d. source asymptotically as well as the best code for that source.

1.3.1 Related Work

Ziv [26] appears to be the first to examine fixed-rate universal compression of sources over fixed alphabets. He shows that universal codes with exponential decay of the error probability exist for sources whose non-universal minimal achievable rates are smaller than the coding rate. Nowadays it is well known that block codes that are universal with respect to the class i.i.d. distributions exist, and in fact these codes can be made to be error-exponent optimal for each source. [27, Th. 2.15].

Universal compression of large alphabet sources is also examined by Orlit-

sky and Santhanam [28] (see also [29] for a modified result), however their focus is on compression redundancy of *variable length* codes for i.i.d. sources. They show that the compression redundancy goes to zero when the alphabet grows sublinearly, but is bounded away from zero when the alphabet grows linearly. Other work investigating variable length compression redundancy of sources on fixed alphabets includes [30, 31, 32]. Although practical lossless compression algorithms are typically variable rate, the fixed-rate framework in this work provides a more natural starting point for studying large alphabets in other coding problems, such as the Slepian-Wolf problem and channel coding.

Our results rely on an information spectrum [33] characterization of lossless compression. This characterization turns the question of the existence of codes of a prescribed rate into a question about the probability that the information random variable, $-n^{-1}\log p_n^n(X^n)$, exceeds the given threshold.

1.3.2 Contributions

In Section 3.2 we show that universal compression of large-alphabet sources is possible when the alphabet grows sublinearly, i.e. $|\mathcal{A}_n| = o(n)$. The scheme we use to achieve this growth rate is not new. Our main contribution is the converse: we show that there are families of alphabets that grow linearly for which universal compression is impossible, even with randomized codes. The converse hinges on the fact that with linear alphabets it is possible to find collections of i.i.d. sources, each having the same entropy, such that a mixture of the sources has an entropy that is strictly larger by an amount that is linear in the blocklength.

In Section 3.3 we introduce and study the problem of source coding with distributional side information in which the decoder is given the distribution of the source, but the encoder knows only that the distribution is i.i.d. over a particular alphabet. If randomized encoders and decoders are permitted then we show that universal coding is possible for *any* alphabet growth rate. This result is reminiscent of the result of Slepian and Wolf [34], who show that decoder knowledge of a correlated random variable Y reduces the required rate from the entropy of the source to the conditional entropy given the side information, and this performance is not achievable if the side information is absent at the decoder. Likewise, here if the decoder alone knows the distribution then universal coding is possible for any alphabet growth, but without the side information sub-linear growth is the best rate that can be handled.

In Section 3.4 we conclude by showing that non-universal compression (i.e. compression of a source with known distribution sequence $\{p_n\}$) is possible at the entropy rate $\{H(p_n)\}$ if and only if $n^{-1/2} \log |\mathcal{A}_n| \to 0$. This is in stark contrast to the variable length case, where it is possible to design a code with normalized expected codeword length arbitrarily close to the entropy, $H(p_n)$ for any alphabet growth rate [35, Eqn. 5.37].

1.4 Reliability in Source Coding with Side Information

In a typical lossy data compression problem a source is to be compressed by an encoder at a prescribed rate so that a decoder may reproduce the source to within some desired fidelity (distortion). Sometimes present, in addition to the data to be compressed, is some correlated information that can be utilized by

a second encoder, that is able to send a separate message to the decoder. We refer to this kind of problem as source coding with side information (SCSI). The set-up is depicted in Fig. 1.1, where a source X is compressed by encoder one to a rate R_1 with the decoder having access to encoded side information Y, compressed at rate R_2 by encoder two, as well as the compressed version of X from the first encoder.

The SCSI scenario arises in a variety of applications. For example, in video applications [36] X can represent a current frame, and Y a separate correlated frame sent from a second encoder; by taking the second rate to be large, Y can even represent the frame(s) preceding the current frame X in the stream. While the previous frames are certainly available to the encoder, the encoder's coding scheme can be simplified by not making use of this information and leaving the decoder to exploit the interframe dependence. A second example can be found in communication in networks with relays [37]. A source sends a message X to a sink in a network containing a relay. One mode of operation for the relay is "compress and forward", i.e. for the relay to send a compressed version of its observation, Y, of the source-sink message to the sink. This compressed message can be used by the sink to further aid its decoding. SCSI appears in applications even beyond communication, for example (with minor changes) it has been proposed as a model for rate-constrained pattern recognition [38].

For the lossless problem with partial side information (SCPSI)[2], and the lossy problem with full side information (Wyner-Ziv), the "rate region" problem, i.e. determining the rates required to meet a given average distortion constraint, is solved. In this chapter, we study these two problems from an error-exponent

[2]Also known as the "One Helper" problem, Wyner's problem [39] or the Ahlswede-Körner problem [40].

standpoint. Our motivation for doing so is three-fold:

- In the applications mentioned above the average distortion of a compression scheme is not the only important metric. Indeed, a video compression system with good average performance but that frequently yields poor images, or a communication system that suffers from frequent outages is usually deemed unacceptable. In addition to minimizing the average distortion, one would like to minimize the fraction of time in which the images are poor or the relay is unable to help.
- In some important cases, there is no *rate loss*, meaning that there is no difference in the rate-distortion performance between the SCSI problem and the problem in which the side information Y is available to the encoder as well as the decoder. In particular, it is well known that this is true of both the binary erasure and quadratic Gaussian forms of the problem [41]. This raises the question of whether these two systems are equivalent when performance is measured via error exponents instead of the average distortion.
- Recently a connection has been established between error exponents in channel coding and the stabilization of linear systems over noisy channels [42], and there is a known interdependence between source- and channel-coding error exponents. Thus new techniques in source-coding error exponents could aid our understanding of problems at the intersection of communication and control [43].

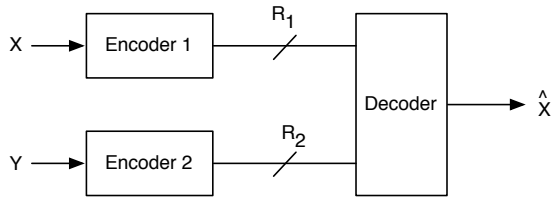

Figure 1.1: The source coding with side information (SCSI) problem

1.4.1 Related Work

Error exponents for both SCPSI and Wyner-Ziv were studied by Arutyunyan and Marutyan [44]. However, their results were not proven rigorously and appear to be unduly strong; they have recently been retracted [45]. Eswaran and Gastpar [46] have established an achievable exponent for the general multiterminal source-coding problem, which yields an achievable exponent for these two problems in particular. Their approach is based on determining the rate of convergence of the Markov lemma, and is fundamentally different from the approach used in this chapter. The approach used here arguably reveals greater insight into both the design of coding schemes for these problems and theoretical questions such as the exponent loss for the Binary Erasure and Gaussian Wyner-Ziv problems.

For the SCPSI problem in particular, Csiszár and Körner [27, pg. 268] provide an upper bound on the reliability function. This bound is formally improved in the present work by using a more refined change-of-measure argument. For the Wyner-Ziv problem, Jayaraman and Berger [47] studied the exponent associated with the binning error probability. One of the goals of this work is to show that a binning error is only one of two competing error events. In this sense, at the error exponent level the Wyner-Ziv problem resembles the

problem of distributed hypothesis testing [48].

The Wyner-Ziv problem is in a sense "dual" to the problem of channel coding with side information (CCSI) [49]. Comparing the results in this work to error exponent studies of the CCSI problem [50, 51], however, show that this duality breaks down at the level of error exponents. In particular, in the CCSI problem, the encoder can force the realization of the auxiliary random variable to have a specified joint distribution with the side information. In the Wyner-Ziv problem, however, the encoder must rely on the law of large numbers to ensure this. At the rate level, atypical realizations can be ignored and this difference is immaterial. At the level of error exponents, on the other hand, the two are quite different, and the Wyner-Ziv setup is more challenging.

There is a substantial literature on error exponents for simpler source coding problems such as lossless compression with side information available at encoder and decoder (full side information) [52, 53, 54], the Slepian-Wolf problem [55, 56, 57], and lossy compression without side information [58, 59]. None of the these problems involve optimization over an auxiliary random variable, however, and we shall see that the presence of auxiliary random variables makes the error exponent problem more interesting.

1.4.2 Contributions

Our key contributions are achievable exponents and converse bounds for the SCPSI and Wyner-Ziv problems. The conventional approach to proving coding theorems for the these problems [35] relies on typicality-based arguments and yields error exponents that are essentially zero. By using more sophisticated

covering and decoding techniques, we obtain lower bounds that are strictly positive for all achievable rates and distortions. Both achievable exponents have a natural interpretation as a two-player game between nature and the code designer, with nature's goal to minimize the exponent and the code designer's goal to maximize it.

In Section 4.2 we give our results for the SCPSI problem. Our upper bound uses a change-of-measure argument that is more refined than the conventional approach [27, pg. 268] and yields a formally better bound. This bound more accurately captures the structure of the problem and might be applicable to other network information setups. The proof also uses the Karush-Kuhn-Tucker (KKT) conditions in a novel way to obtain cardinality bounds on the auxiliary random variable.

In Section 4.3 we give our results for the Wyner-Ziv problem. We supply results for both the discrete-memoryless and Gaussian versions of the problem. Our analysis indicates that the optimization of the coding scheme is a richer problem than it is when the goal is to minimize the average distortion. In particular, there is a tension in the choice of the test channel. If the test channel is "clean" then the codebook is large, which results in a high binning error probability and a low error exponent. On the other hand, if the test channel is "noisy" then the binning error probability is low, but the decoder must rely heavily on the side information Y^n to reconstruct X^n. A small deviation in the empirical distribution of Y^n from its true distribution will then cause an error, which again leads to a poor error exponent. The optimum choice of the test channel balances these two competing error events.

Section 4.4 applies the Wyner-Ziv results to the Binary Erasure and Gaussian

problem, where we illustrate the aforementioned tension numerically for the Binary Erasure version of the problem.

Our results present evidence that, for both the binary erasure and Gaussian cases, there is likely a difference in the error exponents between conventional Wyner-Ziv and the version of the problem in which the side information is available at both encoder and decoder (an "exponent loss"). This is in contrast to the rate-distortion version of the problem, for which the two scenarios have identical performance. Determining whether the reliability functions are indeed different is an interesting topic for future work.

An application of our results on discrete-memoryless Wyner-Ziv allows us to determine the reliability function exactly (for a range of rates) for the lossless functional source coding problem, in which the goal is to reproduce a function $g(X)$ at the decoder (see section 4.3.1).

1.5 Improved Source Coding Exponents via Witsenhausen's Rate

In Chapter 5[3] we improve the results of Chapter 4 for the special case of full side information depicted in Figure 1.2.

[3] ©2011 IEEE. Portions, reprinted, with permission, from [Kelly and Wagner, "Improved Source Coding Exponents via Witsenhausen's Rate", to appear in IEEE Transactions on Information Theory].

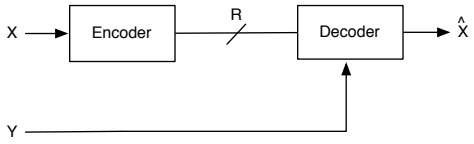

Figure 1.2: Source coding with full side information

1.5.1 Related Work

Bounds on the rate of decay of the error probability for this problem, the so-called *error exponent*, were determined by Csiszár and Körner [53] whose results include a universally attainable random coding exponent and a non-universal 'expurgated' exponent. Previously Gallager [52] derived a non-universal exponent that was later shown to be universally attainable by Csiszár, Körner and Marton [55].

Although our interest is in error exponents and therefore, necessarily, the vanishing error probability formulation of full side information problems, our improvements are derived from the study of a related zero-error problem. The zero-error formulation of source coding with full side information was studied by Witsenhausen [60], who showed that for fixed blocklength, n, the fewest number of messages required so that the decoder can reproduce the source with no error, i.e. $P_{XY}^n(X^n = \hat{X}^n) = 1$, is $\gamma(G_X^n)$, the chromatic number of the n-fold strong product of the characteristic graph of the source; see Section 5.1 for definitions and Körner and Orlitsky [61] for a comprehensive overview of the applications of graph theory in zero-error information theory.

Asymptotically, the required rate, sometimes referred to as Witsenhausen's

rate in the literature, is therefore

$$R(G) = \lim_{n\to\infty} \frac{1}{n} \log \gamma(G^n). \tag{1.1}$$

(Note that the limit in (1.1) exists by sub-additivity and appealing to Fekete's lemma.) Witsenhausen's rate may also be expressed as an optimization over input distributions of the complementary graph entropy functional [62, 63], but no single letter expression for this functional is known. Existing bounds on $R(G)$ include $\log \gamma(G)$, which follows by noting that $\gamma(G^n) \leq \gamma(G)^n$, and graph entropy [64], which bounds complementary graph entropy. The second contribution of this work is a new upper bound on $R(G)$, attained by introducing a new graph functional and showing that it is an upper bound on complementary graph entropy. Our method combines graph- and information-theoretic techniques.

We use the Witsenhausen coding idea and our new functional to give improved error exponents for the full side information problems. The key observation is that all sequences in some typeclasses can be communicated without error using the Witsenhausen scheme, and doing so can strictly improve the error exponent by eliminating certain error events. Unlike existing schemes this requires that the encoder be nonuniversal, although the only knowledge of the source distribution required is the position of the zeroes in the channel matrix that connects the source and the side information.

1.5.2 Contributions

Section 5.1 contains definitions and preliminary facts, including the definition of the new graph functional.

Section 5.2 gives some useful properties of our new functional.

In Section 5.3 we motivate the functional and give our first result, a single letter, computable bound on Witsenhausen's rate. We also prove that our functional bounds complementary graph entropy. Comparison between some of the aforementioned existing bounds are also given.

In Section 5.4, we give our second result, improved error exponents for the problem of lossless source coding with full side-information; examples and comparisons to previous known exponents are also given.

In Section 5.5 we use the ideas from Section 5.4 to give our third and fourth results, an improved error exponent for the lossy Wyner-Ziv problem and determination of the reliability function for the case when the side information is a deterministic function of the source.

In Section 5.6 we briefly give an application of our new bound to channel coding.

CHAPTER 2

CLASSIFICATION OF LARGE ALPHABET SOURCES

In this chapter we formulate and study the problem of classification of large alphabet sources. We introduce the class of α-large-alphabet sources and show that universal classification of such sources is possible when $0 \leq \alpha \leq 2$. We also show that common statistical tests such as chi-squared or likelihood ratio tests are consistent only for $0 \leq \alpha < 1$. We conclude with thoughts on classification of general sources and propose a new classifier that works well empirically.

2.1 Definitions and Problem Statement

Sets are usually denoted using calligraphic letters, e.g. $\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\}$. The set $\mathcal{A}^{\times n}$ is the n-fold cartesian product of $\mathcal{A}$. Strings are denoted in bold face, e.g. $\mathbf{x} = x_1 \cdots x_n$ (usually the blocklength is clear from the context). $\mathbf{1}\{A\}$ is the indicator function for event A and

$$N(a|\mathbf{x}) = \sum_{i=1}^{n} \mathbf{1}\{x_i = a\}.$$

We use $\Lambda_{\mathbf{x}}$ to denote the empirical distribution or *type* of string $\mathbf{x}$, i.e.

$$\Lambda_{\mathbf{x}} = n^{-1}\big[N(a_1|\mathbf{x}) \cdots N(a_{|\mathcal{A}|}|\mathbf{x})\big].$$

The set of all discrete distributions on alphabet $\mathcal{A}$ is denoted $\mathcal{P}(\mathcal{A})$. The set of all sequences of length n with type Q is denoted T_Q^n (again we usually omit n since it is clear from the context). The set of all type variables $Q \in \mathcal{P}(\mathcal{A})$, i.e. those for which $T_Q^n \neq \emptyset$, is denoted $\mathcal{P}^n(\mathcal{A})$. For other information theoretic notations we use the standard definitions, see e.g. [27]. If p is a distribution on

$\mathcal{A}$ then p^n is the n-fold i.i.d. product measure on $\mathcal{A}^{\times n}$, i.e.

$$p^n(\mathbf{x}) = \prod_{i=1}^{n} p(x_i).$$

For triangular arrays, $X_{n,m}$, $1 \leq m \leq n$, $n \geq 1$, the notation X^n refers to the rows of the array, i.e. $X^n = X_{n,1}, \ldots, X_{n,n}$. We use $\|\cdot\|_p$ to denote the pth Euclidean norm and $\langle\cdot\rangle$ to denote the standard inner product.

For any distribution p on a finite set $\mathcal{A}$, $\mathrm{supp}(p)$ denotes its support and we define

$$\check{p} = \min_{a \in (\mathcal{A} \cap \mathrm{supp}(p))} p(a) \text{ and } \hat{p} = \max_{a \in \mathcal{A}} p(a).$$

Our primary focus in the chapter will be the following class of distributions.

Definition 1. The sequence $\{p_n, q_n, \mathcal{A}_n\}$ is an α-large-alphabet source pair if for all n

$$\frac{\check{c}}{n^\alpha} \leq \min(\check{p}_n, \check{q}_n) \leq \max(\hat{p}_n, \hat{q}_n) \leq \frac{\hat{c}}{n^\alpha}, \tag{2.1}$$

where $\check{c}$ and $\hat{c}$ are positive constants independent of n; and where

$$\mathcal{A}_n = \mathcal{A}'_n \cup \mathcal{X}_n \cup \mathcal{Y}_n$$

with

$$\mathcal{A}'_n = \mathrm{supp}(p_n) \cap \mathrm{supp}(q_n)$$
$$\mathcal{X}_n = \mathrm{supp}(p_n) \cap \{a : q_n(a) = 0\}$$
$$\text{and } \mathcal{Y}_n = \mathrm{supp}(q_n) \cap \{a : p_n(a) = 0\}.$$

Note that for any α-large-alphabet source, $|\mathcal{A}_n| = \Theta(n^\alpha)$. This can easily be seen since

$$1 \geq \sum_{a \in \mathcal{A}'_n} p_n(a) \geq |\mathcal{A}'_n| \frac{\check{c}}{n^\alpha} \text{ and } 1 \leq |\mathcal{A}_n| \frac{\hat{c}}{n^\alpha}$$

which along with $1 \geq |\mathcal{X}_n|\frac{\check{c}}{n^\alpha}$ and $1 \geq |\mathcal{Y}_n|\frac{\check{c}}{n^\alpha}$ implies

$$\frac{3n^\alpha}{\check{c}} \geq |\mathcal{A}_n| \geq \frac{n^\alpha}{\hat{c}}.$$

Such distributions may arise from sampling a probability density. For example, suppose $f(x)$ is (almost everywhere) continuous on $[0,1]$ satisfying $\int f(x)dx = 1$ and $\check{c} \leq f(x) \leq \hat{c}$. If X is a random variable with density f and we define p_n as the distribution of $\lceil n^\alpha X \rceil$, then the sequence $\{p_n\}$ is α-large-alphabet with alphabet $\{1, \ldots, n^\alpha\}$. As we will see later studying this class sheds light on the general classification problem.

2.1.1 Problem Statement

For each n, let $X_{n,m}$, $1 \leq m \leq n$ be i.i.d. random variables with distribution p_n and similarly let $Y_{n,m}$, $1 \leq m \leq n$ be i.i.d. with distribution q_n. We assume that p_n and q_n are *unknown* distributions with a common finite alphabet $\mathcal{A}_n$. We also assume that p_n and q_n satisfy

$$\liminf_{n\to\infty} \|p_n - q_n\|_1 = \liminf_{n\to\infty} \sum_{a\in\mathcal{A}_n} |p_n(a) - q_n(a)| > 0. \tag{2.2}$$

For each n we observe independent realizations X^n and Y^n, the nth rows of the corresponding triangular arrays. Given a third independent row $Z_{n,m}$, $1 \leq m \leq n$ generated i.i.d, we wish to test which of hypotheses

$$\mathcal{H}_0 : Z^n \sim p_n^n \text{ for all } n,$$
$$\text{or } \mathcal{H}_1 : Z^n \sim q_n^n \text{ for all } n$$

is in effect. One may think of X^n and Y^n as being training data and the problem

is to determine whether Z^n came from the unknown distribution p_n or q_n. We refer to this problem as the *triangular array hypothesis testing problem*.

Let $P_n = p_n^n \times q_n^n \times p_n^n$ and $Q_n = p_n^n \times q_n^n \times q_n^n$. We will be concerned with the following asymptotic properties of tests.

Definition 2 (α-Universal Consistency)**.** For a given sequence of alphabets $\{\mathcal{A}_n\}_{n=1}^{\infty}$ with $|\mathcal{A}_n| = \Theta(n^\alpha)$, we say a sequence of tests $T_n : \mathcal{A}_n^{\times n} \times \mathcal{A}_n^{\times n} \times \mathcal{A}_n^{\times n} \to \{0, 1\}$ is α-*universally consistent* if for every sequence $\{p_n, q_n\}$ on $\{\mathcal{A}_n\}$ satisfying (2.1) and (2.2),

$$P_n(T_n(X^n, Y^n, Z^n) = 0) \to 1$$
$$\text{and } Q_n(T_n(X^n, Y^n, Z^n) = 1) \to 1 \text{ as } n \to \infty.$$

Definition 3 (Universal Consistency)**.** For a given sequence of alphabets $\{\mathcal{A}_n\}_{n=1}^{\infty}$ we say a sequence of tests $T_n : \mathcal{A}_n^{\times n} \times \mathcal{A}_n^{\times n} \times \mathcal{A}_n^{\times n} \to \{0, 1\}$ is *universally consistent* if for every sequence of distributions $\{p_n, q_n\}$ on $\{\mathcal{A}_n\}$ satisfying condition (2.2),

$$P_n(T_n(X^n, Y^n, Z^n) = 0) \to 1$$
$$\text{and } Q_n(T_n(X^n, Y^n, Z^n) = 1) \to 1 \text{ as } n \to \infty.$$

Note: Implicit in both definitions of universal consistency is that the classifier knows the underlying alphabet, however the classifiers considered in this work do not require knowledge of the symbols that do not appear in the training data. When proving impossibility results, however, we assume the classifier knows the alphabet.

2.2 Testing of α-large-alphabet sources

2.2.1 Achievability

In this subsection we show that α-large-alphabet sources can be handled with a simple test based on Euclidean geometric considerations. Loosely speaking, the idea is that under hypothesis $\mathcal{H}_0$, Λ_{Z^n} should be "closer" to Λ_{X^n} than it is to Λ_{Y^n}, despite the fact that $\|\Lambda_{X^n} - p_n\|_1$ need not tend to zero when $|\mathcal{A}_n|$ grows linearly or faster [65].

Theorem 1. *If $0 \leq \alpha < 2$ then the test*

$$\|\Lambda_{Z^n} - \Lambda_{X^n}\|_2^2 \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \|\Lambda_{Z^n} - \Lambda_{Y^n}\|_2^2 \tag{2.3}$$

is α-universally consistent.

To prove the result we need the following lemmas. Throughout we define

$$F = F(X^n, Y^n, Z^n) = \|\Lambda_{Z^n} - \Lambda_{X^n}\|_2^2 - \|\Lambda_{Z^n} - \Lambda_{Y^n}\|_2^2.$$

Lemma 1.

$$\mathbb{E}_0[F] = \sum_{a \in \mathcal{A}_n} -(p_n(a) - q_n(a))^2 + n^{-1}(q_n^2(a) - p_n^2(a))$$

$$\textit{and } \mathbb{E}_1[F] = \sum_{a \in \mathcal{A}_n} (p_n(a) - q_n(a))^2 + n^{-1}(q_n^2(a) - p_n^2(a))$$

Proof. Using $\mathbb{E}_i$ to denote expectation under $\mathcal{H}_i$, we now compute

$$\mathbb{E}_i[F(X^n, Y^n, Z^n)] = \mathbb{E}_i[\|\Lambda_{X^n}\|_2^2 - \|\Lambda_{Y^n}\|_2^2 - 2\langle \Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n} \rangle].$$

We start with the two-norm of the type

$$\mathbb{E}_i\Big[\|\Lambda_{X^n}\|_2^2\Big] = n^{-2}\sum_{a\in\mathcal{A}_n}\mathbb{E}_i\big[N^2(a|X^n)\big].$$

Since $N(a|X^n)$ is a binomial random variable with parameters $(n, p_n(a))$,

$$\begin{aligned}\mathbb{E}_i\Big[\|\Lambda_{X^n}\|_2^2\Big] &= n^{-2}\sum_{a\in\mathcal{A}_n} np_n(a)(1-p_n(a)) + n^2p_n^2(a)\\ &= n^{-1} + \sum_{a\in\mathcal{A}_n} p_n^2(a) - n^{-1}p_n^2(a)\end{aligned}$$

Similarly

$$\mathbb{E}_i\big[\|\Lambda_{Y^n}\|_2^2\big] = n^{-1} + \sum_{a\in\mathcal{A}_n} q_n^2(a) - n^{-1}q_n^2(a).$$

For the final term

$$\begin{aligned}&\mathbb{E}_i\big[\langle\Lambda_{Z^n}, (\Lambda_{X^n} - \Lambda_{Y^n})\rangle\big]\\ &= n^{-2}\sum_{a\in\mathcal{A}_n}\mathbb{E}_i\Big[N(a|Z^n)(N(a|X^n) - N(a|Y^n))\Big]\\ &= n^{-1}\sum_{a\in\mathcal{A}_n}\mathbb{E}_i\big[N(a|Z^n)](p_n(a) - q_n(a)).\end{aligned}$$

Under hypothesis $\mathcal{H}_0$, the previous line is

$$\sum_{a\in\mathcal{A}_n} p_n(a)^2 - p_n(a)q_n(a)$$

and under hypothesis $\mathcal{H}_1$ is

$$\sum_{a\in\mathcal{A}_n} -q_n(a)^2 + p_n(a)q_n(a).$$

Therefore

$$\begin{aligned}\mathbb{E}_0[F] &= \sum_{a\in\mathcal{A}_n} p_n^2(a) - n^{-1}p_n^2(a) - q_n^2(a) + n^{-1}q_n^2(a) - n^{-1}p_n^2(a) + 2p_n(a)q_n(a)\\ &= \sum_{a\in\mathcal{A}_n} -(p_n(a) - q_n(a))^2 + n^{-1}(q_n^2(a) - p_n^2(a)),\end{aligned}$$

and similarly

$$\mathbb{E}_1[F] = \sum_{a \in \mathcal{A}_n} (p_n(a) - q_n(a))^2 + n^{-1}(q_n^2(a) - p_n^2(a)).$$

□

Lemma 2. *For all* $0 < \alpha < 2$ *and for* $i = 0, 1$

$$\mathrm{Var}_i[n^\alpha F] \to 0$$

Proof. Follows from direct calculation using binomial moments. See Appendix A.1 for details. □

Lemma 3. *For any* α*-large-alphabet source pair* $\{p_n, q_n, \mathcal{A}_n\}$

$$\check{c}/3 \|p_n - q_n\|_1^2 \leq n^\alpha \|p_n - q_n\|_2^2$$

Proof. The result follows from the Cauchy-Schwarz inequality and the bound $|\mathcal{A}_n| \leq \frac{3n^\alpha}{\check{c}}$. □

We are now in a position to prove achievability.

Proof of Theorem 1. Case 1 : $0 < \alpha < 2$. Notice that the test $n^\alpha F \lessgtr 0$ makes the same decision as the test in the statement of the theorem. When hypothesis $\mathcal{H}_1$ is in effect (a subscript on operators denotes this) Lemma 1 tells us

$$\mathbb{E}_1[F] = \sum_{a \in \mathcal{A}_n} (p_n(a) - q_n(a))^2 + n^{-1}(q_n^2(a) - p_n^2(a)),$$

where both $\sum_{\mathcal{X}_n} p_n^2(a)$ and $\sum_{\mathcal{Y}_n} q_n^2(a)$ are $O(n^{-\alpha})$. Therefore by Lemma 3 we have

$$\begin{aligned} \liminf_{n\to\infty} \mathbb{E}_1[n^\alpha F] &= \liminf_{n\to\infty} n^\alpha \sum_{a \in \mathcal{A}_n} (p_n(a) - q_n(a))^2 \\ &\geq \liminf_{n\to\infty} \frac{\check{c}}{3} \|p_n - q_n\|_1^2, \end{aligned}$$

which is strictly positive by hypothesis. Invoking Lemma 2

$$\mathrm{Var}_1(n^\alpha F) \to 0$$

and the result follows from Chebyshev's inequality[1]. The hypothesis $\mathcal{H}_0$ is handled analogously.

Case 2: $\alpha = 0$. For this case we take square root of both sides of (2.3) so that we are working with norms. Now the result may be be proved using the weak law of large numbers (see for example Lemma 10 in Section 2.3). Suppose hypothesis $\mathcal{H}_0$ is in effect. The lefthand side of (2.3) is

$$\|\Lambda_{X^n} - \Lambda_{Z^n}\|_2 \leq \|\Lambda_{X^n} - p_n\|_2 + \|\Lambda_{Z^n} - p_n\|_2$$

and both terms on the right of the previous display tend to zero in probability. For the righthand side, note that by the reverse triangle inequality

$$\Big|\|\Lambda_{Y^n} - \Lambda_{Z^n}\|_2 - \|p_n - q_n\|_2\Big| \leq \|\Lambda_{Y^n} - q_n\|_2 + \|\Lambda_{Z^n} - p_n\|_2$$

and so for n large enough $\|\Lambda_{Y^n} - \Lambda_{Z^n}\|_2$ is as close to $\|p_n - q_n\|_2$ as we desire. Finally note that the hypothesis $\liminf_{n\to\infty} \|p_n - q_n\|_1 > 0$ implies $\liminf_{n\to\infty} \|p_n - q_n\|_2 > 0$ if the alphabet is not growing with n. □

2.2.2 Converse

We next show that the result in Theorem 1 cannot be improved.

Theorem 2 (Converse). *If $\alpha \geq 2$, then there are alphabets with growth rate $\Theta(n^\alpha)$ for which there are no α-universally consistent tests.*

[1]Sharper concentration results can be obtained using martingale techniques; see Theorem 25 in the Appendix for one such result.

To prove the result we need the following additional machinery.

Definition 4 (Testing Affinity)**.** Suppose P and Q are probability measures on some space $\mathbb{X}$ dominated by λ with densities f and g. Let the density $f \wedge g$ define the (sub-probability) measure $P \wedge Q$, i.e.

$$(P \wedge Q)(A) = \int_A (f \wedge g) d\lambda.$$

with $f \wedge g$ denoting the pointwise minimum of f and g.

Note that $2(a \wedge b) = a + b - |a - b|$, and so we may also write

$$\|P \wedge Q\|_1 = 1 - \frac{1}{2}\|P - Q\|_1. \tag{2.4}$$

Following Le Cam [66, Ch.16 §4] we associate with a hypothesis $\mathcal{H}_0$ (resp. $\mathcal{H}_1$) a set of measures, say A (resp. B). Let $0 \leq \phi \leq 1$ be a randomized test function, i.e. a function which gives the probability of accepting hypothesis $\mathcal{H}_0$. For a given ϕ and sets of measures A and B we define the worst case "average" error probability as follows

$$\mathfrak{R}(A, B, \phi) = \sup_{P \in A, Q \in B} \Big[\int (1 - \phi) dP + \int \phi dQ \Big],$$

and define the minimax error probability (or risk) as

$$\mathfrak{R}(A, B) = \inf_{\phi} \mathfrak{R}(A, B, \phi)$$

i.e. $\mathfrak{R}(A, B)$ is the best universally achievable risk. We recall the following result.

Lemma 4. *[Kraft [66, Ch.16 §4, Lem. 1]]*

$$\mathfrak{R}(A, B) = \sup_{P \in \mathrm{conv}(A), Q \in \mathrm{conv}(B)} \|P \wedge Q\|$$

where $\mathrm{conv}(A)$ *denote the convex hull of the set A.*

Equality (2.4) and Lemma 4 allow us to express minimax risk in terms of L_1 distances between convex hulls. We will also need the following result.

Lemma 5. *For any pair of probability measures P and Q, both dominated by a probability measure λ,*

$$\|P - Q\|_1^2 \leq \int \Big(\frac{dP}{d\lambda} - \frac{dQ}{d\lambda}\Big)^2 d\lambda.$$

Proof. Applying the Cauchy-Schwarz inequality gives

$$\begin{aligned}
\|P - Q\|_1 &= \int \Big|\frac{dP}{d\lambda} - \frac{dQ}{d\lambda}\Big| d\lambda \\
&\leq \sqrt{\int d\lambda \int \Big(\frac{dP}{d\lambda} - \frac{dQ}{d\lambda}\Big)^2 d\lambda} \\
&= \sqrt{\int \Big(\frac{dP}{d\lambda} - \frac{dQ}{d\lambda}\Big)^2 d\lambda}.
\end{aligned}$$

□

We now use these facts to establish a converse result. We first give a lower bound on the risk for a suitably chosen hypothesis testing problem on the sequence of alphabets $\mathcal{A}_n = \{1, \ldots, \lceil n^\alpha \rceil_2\}$, where $\lceil \cdot \rceil_2$ denotes rounding up to the next even integer. Define sets

$$\begin{aligned}
\mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}} = \{(p_n, q_n) \in \mathcal{P}(\mathcal{A}_n^{\times 2}) : &\|p_n - q_n\|_1 \geq \epsilon, \\
&\check{c} n^{-\alpha} \leq \min(\check{p}_n, \check{q}_n) \leq \max(\hat{p}_n, \hat{q}_n) \leq \hat{c} n^{-\alpha} \\
&\forall a \in \mathcal{A}_n : \max(p_n(a), q_n(a)) > 0\}, \\
A_{n,\alpha,\epsilon,\check{c},\hat{c}} = \{p_n^n \times q_n^n \times p_n^n : (p_n, q_n) \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}\}, \\
\text{and } B_{n,\alpha,\epsilon,\check{c},\hat{c}} = \{p_n^n \times q_n^n \times q_n^n : (p_n, q_n) \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}\}.
\end{aligned}$$

Observe that for any choice of $\epsilon > 0$ and constants $\check{c}, \hat{c}$ any sequence of pairs distributions $\{p_n, q_n\}$ with the nth chosen from $\mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}$ is by definition α-large

alphabet and moreover

$$\liminf_{n\to\infty} \|p_n - q_n\|_1 \geq \epsilon.$$

The following upper bound on the L_1 distance between the convex hulls of the sets for this testing problem combined with (2.4) give the aforementioned lower bound on the risk. The proof of the bound is similar in spirit to that of [13, Th. 4], which in turn borrows ideas from [67], using a so-called "mixture measure" to construct bad convex combinations. In our proof we apply the mixture measure idea to address the composite-versus-composite problem studied here.

Lemma 6. *Let* $0 < \epsilon < 1$. *For* $0 < \check{c} \leq \frac{1-\epsilon}{3} < 1 + \epsilon \leq \hat{c}$ *there exists* $P_n \in \mathrm{conv}(A_{n,\alpha,\epsilon,\check{c},\hat{c}})$ *and* $Q_n \in \mathrm{conv}(B_{n,\alpha,\epsilon,\check{c},\hat{c}})$ *so that*

$$\|Q_n - P_n\|_1 \leq \sqrt{2}\exp\left(\frac{n^2\epsilon^4}{2n^\alpha}\right).$$

Proof. Define $m = \lceil n^\alpha \rceil_2$. Let u_n be the uniform distribution on $\{1, \ldots, m\}$. Let $\Pi = \{-1, 1\}^{\times(m/2)}$ i.e. the set of all $\{-1, 1\}$ vectors of length $m/2$. For any $\pi \in \Pi$ let

$$\nu(i, \pi) = \begin{cases} \pi_{i/2} & i \text{ even} \\ -\pi_{(i+1)/2} & i \text{ odd,} \end{cases}$$

and define the distribution $q_{n,\pi}$ as

$$q_{n,\pi}(i) = (1 + \epsilon\nu(i, \pi))m^{-1} \text{ for } i \in \{1, \ldots, m\}.$$

We note that

$$\|q_{n,\pi} - u_n\|_1 = \epsilon \text{ for all } \pi. \tag{2.5}$$

Also since for all positive real x

$$x \leq \lceil x \rceil_2 \leq x + 2,$$

one has

$$\frac{1}{3} \leq \frac{n^\alpha}{m} \leq 1. \tag{2.6}$$

Define measures

$$P_{n,\pi} = u_n^n \times q_{n,\pi}^n \times u_n^n \text{ and } Q_{n,\pi} = q_{n,\pi}^n \times u_n^n \times u_n^n$$

and observe that (2.5), and (2.6) combined with

$$\frac{1-\epsilon}{m} \leq \min(\check{u}_n, \check{q}_{n,\pi}) \leq \max(\hat{u}_n, \hat{q}_{n,\pi}) \leq \frac{1+\epsilon}{m}$$

imply that $P_{n,\pi} \in A_{n,\alpha,\epsilon,\check{c},\hat{c}}$ and $Q_{n,\pi} \in B_{n,\alpha,\epsilon,\check{c},\hat{c}}$ for the $\epsilon, \check{c}$ and $\hat{c}$ of the theorem. Let μ denote the uniform distribution on the set Π and define mixtures

$$P_n = \sum_{\pi\in\Pi} P_{n,\pi}\mu(\pi) \text{ and } Q_n = \sum_{\pi\in\Pi} Q_{n,\pi}\mu(\pi).$$

Note that $P_n \in \text{conv}(A_{n,\alpha,\epsilon,\check{c},\hat{c}})$ and $Q_n \in \text{conv}(B_{n,\alpha,\epsilon,\check{c},\hat{c}})$ and further

$$P_n(\mathbf{x},\mathbf{y},\mathbf{z}) = m^{-2n} \sum_{\pi\in\Pi} \mu(\pi) q_{n,\pi}^n(\mathbf{y})$$

and

$$Q_n(\mathbf{x},\mathbf{y},\mathbf{z}) = m^{-2n} \sum_{\pi\in\Pi} \mu(\pi) q_{n,\pi}^n(\mathbf{x}).$$

We will now show that the stated L_1 bound holds for this choice of P_n and Q_n.

Taking $\lambda = u_n^n \times u_n^n \times u_n^n$ and invoking Lemma 5 we have

$$\begin{aligned}
\|P_n - Q_n\|_1^2 &\leq \sum_{\mathbf{x},\mathbf{y},\mathbf{z}} \Big(\frac{P_n(\mathbf{x},\mathbf{y},\mathbf{z}) - Q_n(\mathbf{x},\mathbf{y},\mathbf{z})}{\lambda(\mathbf{x},\mathbf{y},\mathbf{z})}\Big)^2 \lambda(\mathbf{x},\mathbf{y},\mathbf{z}) \\
&= E_\lambda\Big[\Big(\frac{P_n(X^n,Y^n,Z^n) - Q_n(X^n,Y^n,Z^n)}{\lambda(X^n,Y^n,Z^n)}\Big)^2\Big] \\
&= \mathbb{E}_\lambda\Big[\Big(\frac{m^{-2n}\sum_{\pi\in\Pi}\mu(\pi)q_{n,\pi}^n(Y^n) - m^{-2n}\sum_{\pi\in\Pi}\mu(\pi)q_{n,\pi}^n(X^n)}{m^{-3n}}\Big)^2\Big] \\
&= m^{2n}\mathbb{E}_\lambda\Big[\Big(\sum_{\pi\in\Pi}\mu(\pi)q_{n,\pi}^n(Y^n) - \sum_{\pi\in\Pi}\mu(\pi)q_{n,\pi}^n(X^n)\Big)^2\Big] \\
&\leq m^{2n}\mathbb{E}_\lambda\Big[\Big(\sum_{\pi\in\Pi}\mu(\pi)q_{n,\pi}^n(Y^n)\Big)^2\Big] + m^{2n}\mathbb{E}_\lambda\Big[\Big(\sum_{\pi\in\Pi}\mu(\pi)q_{n,\pi}^n(X^n)\Big)^2\Big].
\end{aligned}$$

Noting that under λ, Y^n and X^n have the same distribution and then expanding the square, we see that

$$\begin{aligned}
\|P_n - Q_n\|_1^2 &\leq 2m^{2n} \sum_{\pi \in \Pi} \sum_{\gamma \in \Pi} \mu(\pi)\mu(\gamma) \mathbb{E}_\lambda \left[q_{n,\pi}^n(Y^n) q_{n,\gamma}^n(Y^n) \right] \\
&= 2m^{2n} \sum_{\pi \in \Pi} \sum_{\gamma \in \Pi} \mu(\pi)\mu(\gamma) \mathbb{E}_\lambda \left[\prod_{i=1}^n q_{n,\pi}(Y_i) q_{n,\gamma}(Y_i) \right] \\
&= 2m^{2n} \sum_{\pi \in \Pi} \sum_{\gamma \in \Pi} \mu(\pi)\mu(\gamma) \Big(\mathbb{E}_{u_n} [q_{n,\pi}(Y_i) q_{n,\gamma}(Y_i)] \Big)^n, \qquad (2.7)
\end{aligned}$$

where on the previous line we used the fact under λ the Y_i are i.i.d. uniform random variables. Focusing on the expectation alone

$$\begin{aligned}
\mathbb{E}_{u_n} [q_{n,\pi}(Y_i) q_{n,\gamma}(Y_i)] &= \sum_i u_n(i)(1 + \epsilon\nu(i,\pi)) m^{-1} (1 + \epsilon\nu(i,\gamma)) m^{-1} \\
&= m^{-3} \sum_i 1 + \epsilon[\nu(i,\pi) + \nu(i,\gamma)] + \epsilon^2 \nu(i,\pi)\nu(i,\gamma) \\
&= m^{-3} \sum_i 1 + \epsilon^2 \nu(i,\pi)\nu(i,\gamma) \\
&= m^{-2} + m^{-3}\epsilon^2 \sum_{i \text{ even}} \pi_{i/2}\gamma_{i/2} + \sum_{i \text{ odd}} \pi_{(i+1)/2}\gamma_{(i+1)/2} \\
&= m^{-2} + 2m^{-3}\epsilon^2 \sum_{i=1}^{m/2} \phi(\pi_i, \gamma_i)
\end{aligned}$$

where $\phi(\pi_i, \gamma_i) = 1$ when $\pi_i = \gamma_i$ and $\phi(\pi_i, \gamma_i) = -1$ otherwise. Applying this calculation to (2.7) yields

$$\begin{aligned}
\|P_n - Q_n\|_1^2 &\leq 2m^{2n} \sum_{\pi \in \Pi} \sum_{\gamma \in \Pi} \mu(\pi)\mu(\gamma) \Big(m^{-2} + 2m^{-3}\epsilon^2 \sum_{i=1}^{m/2} \phi(\pi_i, \gamma_i) \Big)^n \\
&= 2 \sum_{\pi \in \Pi} \sum_{\gamma \in \Pi} \mu(\pi)\mu(\gamma) \Big(1 + 2m^{-1}\epsilon^2 \sum_{i=1}^{m/2} \phi(\pi_i, \gamma_i) \Big)^n \\
&\leq 2 \sum_{\pi \in \Pi} \sum_{\gamma \in \Pi} \mu(\pi)\mu(\gamma) \exp\Big(\frac{2n\epsilon^2}{m} \sum_{i=1}^{m/2} \phi(\pi_i, \gamma_i) \Big)
\end{aligned}$$

where we used the inequality $\log(1 + x) \leq x$. Recalling that μ is uniform over

$\{-1,1\}^{\times(m/2)}$ we may write

$$\begin{aligned}\|P_n - Q_n\|_1^2 &\le 2\mathbb{E}_{\pi,\gamma}\Big[\exp\Big(\frac{2n\epsilon^2}{m}\sum_{i=1}^{m/2}\phi(\pi_i,\gamma_i)\Big)\Big] \\ &= 2\Big(\frac{1}{2}\exp\Big(-\frac{2n\epsilon^2}{m}\Big)+\frac{1}{2}\exp\Big(\frac{2n\epsilon^2}{m}\Big)\Big)^{m/2}.\end{aligned}$$

Applying the inequality

$$\frac{1}{2}(\exp(u)+\exp(-u)) \le \exp\Big(\frac{u^2}{2}\Big),$$

which follows from Hoeffding's Lemma (or by simply comparing the series expansions), gives

$$\begin{aligned}\|P_n - Q_n\|_1^2 &\le 2\exp\Big(\frac{2n^2\epsilon^4}{m^2}\Big)^{m/2} \\ &= 2\exp\Big(\frac{n^2\epsilon^4}{m}\Big),\end{aligned}$$

i.e.

$$\|P_n - Q_n\|_1 \le \sqrt{2}\exp\Big(\frac{n^2\epsilon^4}{2m}\Big) \le \sqrt{2}\exp\Big(\frac{n^2\epsilon^4}{2n^\alpha}\Big).$$

□

We are now in a position to prove Theorem 2. Roughly the argument is as follows. Recall that the setup of Lemma 6 provides the tester with ϵ, the minimum L_1 distance between distributions and constants $\check{c}, \hat{c}$. But even for this "easier" problem, there is some choice of $\check{c}, \hat{c}, \epsilon$ and distributions $P_n \in \mathrm{conv}(A_{n,\epsilon,\check{c},\hat{c}})$ and $Q_n \in \mathrm{conv}(B_{n,\epsilon,\check{c},\hat{c}})$ so that when $\alpha \ge 2$

$$\limsup_{n\to\infty} \|P_n - Q_n\|_1 < 2$$

implying that no α-universally consistent test is exists.

Theorem (2). *If $\alpha \ge 2$, then there are alphabets with growth rate $\Theta(n^\alpha)$ for which there are no α-universally consistent tests.*

Proof. Let $\alpha \geq 2$, $\mathcal{A}_n = \{1, \ldots, \lceil n^\alpha \rceil_2\}$ and suppose by way of contradiction that there exists $\{T_n\}$, a universally consistent test for the α-large-alphabet hypothesis testing problem having alphabet $\mathcal{A}_n$. Now fix $0 < \epsilon < 1$, $0 < \check{c} \leq \frac{1-\epsilon}{3} < 1 + \epsilon \leq \hat{c}$ and choose $(p_n, q_n) \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}$ so that

$$p_n^n \times q_n^n \times p_n^n(T_n = 1) \geq \frac{1}{2} \sup_{\tilde{p}_n, \tilde{q}_n \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}} \tilde{p}_n^n \times \tilde{q}_n^n \times \tilde{p}_n^n(T_n = 1).$$

Since $\{T_n\}$ is α-universally consistent we have that

$$p_n^n \times q_n^n \times p_n^n(T_n = 1) \to 0$$

which in turn implies that

$$\sup_{\tilde{p}_n, \tilde{q}_n \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}} \tilde{p}_n^n \times \tilde{q}_n^n \times \tilde{p}_n^n(T_n = 1) \to 0. \tag{2.8}$$

We now choose $(r_n, s_n) \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}$ so that

$$r_n^n \times s_n^n \times s_n^n(T_n = 0) \geq \frac{1}{2} \sup_{\tilde{r}_n, \tilde{s}_n \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}} \tilde{r}_n^n \times \tilde{s}_n^n \times \tilde{s}_n^n(T_n = 0),$$

and therefore again by universality we must have

$$\sup_{\tilde{r}_n, \tilde{s}_n \in \mathcal{C}_{n,\alpha,\epsilon,\check{c},\hat{c}}} \tilde{r}_n^n \times \tilde{s}_n^n \times \tilde{s}_n^n(T_n = 0) \to 0. \tag{2.9}$$

Thus the existence of a α-universal test implies that

$$\sup_{P_n \in \mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}}} P_n(T_n = 1) \to 0$$

and

$$\sup_{Q_n \in \mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}}} Q_n(T_n = 0) \to 0$$

and therefore

$$\sup_{\substack{P_n \in \mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}} \\ Q_n \in \mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}}}} P_n(T_n = 1) + Q_n(T_n = 0) \to 0.$$

But

$$\sup_{\substack{P_n \in \mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}} \\ Q_n \in \mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}}}} P_n(T_n = 1) + Q_n(T_n = 0) \geq \inf_{\tilde{T}_n} \sup_{\substack{P_n \in \mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}} \\ Q_n \in \mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}}}} P_n(\tilde{T}_n = 1) + Q_n(\tilde{T}_n = 0) \tag{2.10}$$

$$= \mathfrak{R}(\mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}}, \mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}})$$

$$\geq 1 - \frac{\sqrt{2}}{2} \exp\left(\frac{n^2\epsilon^4}{2n^\alpha}\right) \tag{2.11}$$

where in (2.10) the infimum is over all (randomized) tests and where (2.11) follows from Lemma 6 and (2.4). Note when $\alpha > 2$ the exponential term goes to 1 as $n \to \infty$ and $1 - \sqrt{2}/2$ is strictly greater than zero. When $\alpha = 2$ taking $\epsilon = (1/2\log 2)^{1/4} > 0$ gives $1 - 2^{-1/4} > 0$. Thus for any $\alpha \geq 2$, choosing this ϵ and taking limits we obtain the inequality

$$\begin{aligned} 0 &= \lim_{n\to\infty} \sup_{\substack{P_n \in \mathcal{A}_{n,\alpha,\epsilon,\check{c},\hat{c}} \\ Q_n \in \mathcal{B}_{n,\alpha,\epsilon,\check{c},\hat{c}}}} P_n(T_n = 1) + Q_n(T_n = 0) \\ &\geq \lim_{n\to\infty} 1 - \frac{\sqrt{2}}{2} \exp\left(\frac{n^2\epsilon^4}{2n^\alpha}\right) \\ &> 0 \end{aligned}$$

a contradiction, and thus no such α-universal test $\{T_n\}$ exists. □

Although we used a particular choice $\{\mathcal{A}_n\}$ to prove the converse, a slight modification of Theorem 2 goes through for any $\{\mathcal{A}_n\}$ with $|\mathcal{A}_n| = \Theta(n^\alpha)$. Thus we can in fact state the following more general theorem.

Theorem 3. *Let $\{\mathcal{A}_n\}$ by any sequence of alphabets with $|\mathcal{A}_n| = \Theta(n^\alpha)$. Then there are no α-universal consistent tests for any $\alpha \geq 2$.*

2.3 Generalized Likelihood Ratio and Chi-Squared Tests

In this section we study the performance of two commonly used statistical tests: the generalized likelihood ratio and chi-squared tests. We show that both tests are α-universally consistent with sub-linear alphabet growth and that both tests are inconsistent with linear alphabet growth. Note that for both tests we actually prove *universal consistency* as opposed to merely α-universal consistency for up-to sub-linear alphabet growth, we return to this point in the conclusion.

2.3.1 GLRT and its Consistency

The GLRT is derived from the maximum likelihood method, which compares the likelihood functions evaluated with the most likely distribution in the hypothesis sets $\mathcal{H}_0$ and $\mathcal{H}_1$. This gives

$$\max_{p_n,q_n\in\mathcal{P}(\mathcal{A}_n)} p_n^n(X^n)q_n^n(Y^n)p_n^n(Z^n) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} \max_{p_n,q_n\in\mathcal{P}(\mathcal{A}_n)} p_n^n(X^n)q_n^n(Y^n)q_n^n(Z^n),$$

where the maximizations are over *arbitrary distributions* on the alphabet $\mathcal{A}_n$. (Recall that the constants $\check{c}, \hat{c}$ defining the α-large-alphabet sequence are unknown by the tester and the L_1 constraint is asymptotic in nature any so any p_n and q_n are feasible.)

The following Lemma allows us to rewrite the GLRT in terms of Kullback-Leibler divergences.

Lemma 7. *For any three probability distributions x, y and z on a common alphabet $\mathcal{A}$*

$$\min_{p,q\in\mathcal{P}(\mathcal{A})} D(x||p) + D(y||q) + D(z||p) = D(x||\hat{p}) + D(z||\hat{p}),$$

where

$$\hat{p} = (x + z)/2.$$

Proof. Choosing $q = y$ yields $D(y||q) = 0$. For the optimal p, the result follows from the parallelogram identity [27, Ex 1.3.19],

$$D(x||p) + D(z||p) = D(x||(x+z)/2) + D(z||(x+z)/2) + 2D((x+z)/2||p).$$

□

Using this Lemma combined with the well-known identity [27, Ch 1, Lemma 2.6]

$$p^n(\mathbf{x}) = \exp(-n[D(\Lambda_{\mathbf{x}}||p) + H(\Lambda_{\mathbf{x}})]) \tag{2.12}$$

we see that the GLRT test is equivalent to

$$D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n), \tag{2.13}$$

where $\hat{p}_n = (\Lambda_{X^n} + \Lambda_{Z^n})/2$ and $\hat{q}_n = (\Lambda_{Y^n} + \Lambda_{Z^n})/2$. Later it we will find the following useful. Define the functional

$$G(p, q, \mathcal{M}) = \sum_{a \in \mathcal{M}} p(a) \log\Big(\frac{2p(a)}{p(a)+q(a)}\Big) + q(a) \log\Big(\frac{2q(a)}{p(a)+q(a)}\Big)$$

and notice we may equivalently write the GLRT (2.13) as

$$G(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} G(\Lambda_{Y^n}, \Lambda_{Z^n}, \mathcal{A}_n).$$

We will also make use of the following result.

Lemma 8. *Suppose p and q are distributions on an alphabet $\mathcal{A}$, then*

$$G(p, q, \mathcal{A}) = \sum_{a \in \mathcal{A}} \sum_{i:\text{even}} \frac{1}{i(i-1)} \frac{(q(a) - p(a))^i}{(p(a) + q(a))^{i-1}}.$$

Further,

$$p(a)\log\frac{2p(a)}{p(a)+q(a)}+q(a)\log\frac{2q(a)}{p(a)+q(a)}\geq 0.$$

It turns out the growth-rate of the alphabet is of critical interest for proving consistency of the statistical tests. The following result allows us to prove a "weak law" for empirical distributions (to be used later) and Theorem 4, the consistency of the GLRT for sub-linear alphabet growth.

Lemma 9. *If* $|\mathcal{A}_n| = o(n)$ *then*[2]

$$n^{-1}\log|\mathcal{P}^n| \to 0 \text{ as } n\to\infty.$$

Proof. See [12, Lem. 1] □

Lemma 10 (Empirical Weak Law)**.** *Let* $X_{n,m}$, $1\leq m\leq n$ *be i.i.d. with distribution* p_n *on alphabet* $\mathcal{A}_n$*. If* $|\mathcal{A}_n| = o(n)$ *then for any* $\epsilon>0$

$$p_n^n(D(\Lambda_{X^n}||p_n)>\epsilon)\leq e^{-n(\epsilon-\delta_n)},$$

where $\delta_n(|\mathcal{A}_n|)\to 0$ *as* $n\to\infty$*.*

The final components of our proof of consistency of the GLRT (and chi-squared tests) are the following concentration results, which we include here for completeness.

Definition 5. A function $g:\mathcal{A}^n\to\mathbb{R}$ has the *bounded differences* property if for some non-negative constants $c_1,\ldots,c_n$,

$$\sup_{x_1,\ldots,x_n,x_i'\in\mathcal{A}}|g(x_1,\ldots,x_n)-g(x_1,\ldots,x_{i-1},x_i',x_{i+1},\ldots,x_n)|\leq c_i, \text{ for } 1\leq i\leq n. \tag{2.14}$$

[2]The sequence a_n has the property $a_n=o(b_n)$ iff $\lim\frac{a_n}{b_n}=0$.

Lemma 11 (Efron-Stein Inequality [68, 69])**.** *Let $\mathcal{A}$ be any set and let $g^n : \mathcal{A}^n \to \mathbb{R}$ be a function of n variables. Define $Z = g(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are arbitrary independent random variables taking values in $\mathcal{A}$. Let $X_1', \ldots, X_n'$ be independent copies of $X_1, \ldots, X_n$ and define*

$$Z_i' = g(X_1, \ldots, X_i', \ldots, X_n)$$

then

$$\mathrm{Var}(Z) \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[(Z - Z_i')^2].$$

Corollary 1. *Suppose g satisfies the hypothesis of Lemma 11 and has bounded differences with constant c. Then*

$$\mathrm{Var}(Z) \leq \frac{nc^2}{2}.$$

To establish consistency of the GLRT we also need

Lemma 12. *The quantity*

$$D(\Lambda_{\mathbf{x}} || (\Lambda_{\mathbf{x}} + \Lambda_{\mathbf{z}})/2)$$

viewed as a real-valued function of the vector $(\mathbf{x}, \mathbf{z}) = (x_1, \ldots, x_n, z_1, \ldots, z_n)$ satisfies the bounded differences property with the single constant

$$\frac{2}{n}(1 + \log 2 + \log(1 + n)).$$

Proof. See Appendix A.2. □

Theorem 4. *If $|\mathcal{A}_n| = o(n)$ then the GLRT* (2.13) *is universally consistent.*

Proof. Suppose hypothesis $\mathcal{H}_0$ is in effect. Define the set

$$\mathcal{D}_n^{\epsilon} = \{(\mathbf{x}, \mathbf{z}) : G(\Lambda_{\mathbf{x}}, \Lambda_{\mathbf{z}}) > \epsilon\}.$$

By definition

$$P_n((X^n, Z^n) \in \mathcal{D}_n^\epsilon) = \sum_{(\mathbf{x},\mathbf{z})\in\mathcal{D}_n^\epsilon} p_n^n(\mathbf{x}) p_n^n(\mathbf{z})$$
$$= \sum_{\substack{Q_X\in\mathcal{P}^n(\mathcal{A}_n) \\ Q_Z\in\mathcal{P}^n(\mathcal{A}_n):\\ G(Q_X,Q_Z)>\epsilon}} \sum_{\substack{\mathbf{x}\in T(Q_X) \\ \mathbf{z}\in T(Q_Z)}} p_n^n(\mathbf{x}) p_n^n(\mathbf{z}).$$

Using identity (2.12) and the bound [27, Ch 1, Lemma 2.5]

$$|T(Q_X)| \leq \exp(nH(Q_X)),$$

it follows that

$$\sum_{\mathbf{x}\in T(Q_X)} \sum_{\mathbf{z}\in T(Q_Z)} p_n^n(\mathbf{x}) p_n^n(\mathbf{z})$$
$$\leq \exp(-n[D(Q_X||p_n) + D(Q_Z||p_n)]).$$

Further, as in the proof of Lemma 7 we have for all distributions Q_X, Q_Z, p_n

$$D(Q_X||p_n) + D(Q_Z||p_n) \geq G(Q_X, Q_Z)$$

and therefore

$$P_n((X^n, Z^n) \in \mathcal{D}_n^\epsilon) \leq |\{\mathcal{P}(\mathcal{A}_n)\}|^2 e^{-n\epsilon}.$$

By way of Lemma 9 and the hypothesis, this implies that for all $\epsilon > 0$

$$P_n(D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) > \epsilon) \to 0 \text{ as } n \to \infty.$$

It remains to show that for some $\delta > 0$

$$\lim_{n\to\infty} P_n(D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n) > \delta) = 1. \tag{2.15}$$

Chebyshev's inequality tells us for any $\delta > 0$

$$P_n(|D(\Lambda_{Y^n}||\hat{q}_n) - \mathbb{E}[D(\Lambda_{Y^n}||\hat{q}_n)]| > \delta)$$
$$\leq \frac{\mathrm{Var}(D(\Lambda_{Y^n}||\hat{q}_n))}{\delta^2}.$$

The bounded differences property (Lemma 12) and the Efron-Stein inequality (Lemma 11) imply that this variance goes to zero. Thus it follows with probability tending to one, $D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n)$ 'concentrates' around $\mathbb{E}[D(\Lambda_{Y^n}||\hat{q}_n)] + \mathbb{E}[D(\Lambda_{Z^n}||\hat{q}_n)]$. Recalling $D(p||q)$ is convex in the pair (p, q), by Jensen's inequality

$$\begin{aligned}
&\mathbb{E}[D(\Lambda_{Y^n}||\hat{q}_n)] + \mathbb{E}[D(\Lambda_{Z^n}||\hat{q}_n)] \\
&\geq D(\mathbb{E}[\Lambda_{Y^n}]||\mathbb{E}[\hat{q}_n]) + D(\mathbb{E}[\Lambda_{Z^n}]||\mathbb{E}[\hat{q}_n]) \\
&= D(q_n||(p_n + q_n)/2) + D(p_n||(p_n + q_n)/2),
\end{aligned}$$

and from (2.2) and Pinsker's inequality [27, Ex 1.3.17]

$$\begin{aligned}
&\liminf_{n\to\infty} D(p_n||(p_n + q_n)/2) + D(q_n||(p_n + q_n)/2) \\
&\geq \liminf_{n\to\infty} \frac{1}{4\log 2} \left(\sum_{a\in\mathcal{A}_n} |p_n(a) - q_n(a)| \right)^2 \\
&> 0.
\end{aligned}$$

Thus for n sufficiently large $D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n)$ concentrates around a strictly positive quantity, which is enough to establish (2.15). Under hypothesis $\mathcal{H}_1$ the proof is similar. □

We now show that when the alphabet growth is linear, i.e. $\alpha = 1$, the GLRT is not α-universally consistent. We do this by means of a particular counterexample which we will refer to throughout the remainder of the chapter.

We first need the following technical result.

Lemma 13. *Let $\{p_n, q_n\}$ be a sequence of pairs of distributions and denote by $\mu_n^2(x, y)$ the shadow (see [20]), i.e. the distribution of the random vector $\big(np_n(X_n), nq_n(X_n)\big)$ when $X_n \sim p_n$. If $\mu_n^2(x, y)$ converges weakly to $\mu^2(x, y)$, then under hypothesis $\mathcal{H}_0$ (i.e.*

$Z^n \sim p_n^n$)

$$\mathbb{E}[D(\Lambda_{Z^n}||\hat{p}_n)] \to \int \Big[\sum_{j=1}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!} \log(2j) - \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-x)x^{k}}{k!} \log(j+k)\Big]\, d\mu^2(x,y)$$

and

$$\mathbb{E}[D(\Lambda_{Z^n}||\hat{q}_n)] \to \int \Big[\sum_{j=1}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!} \log(2j) - \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-y)y^{k}}{k!} \log(j+k)\Big]\, d\mu^2(x,y).$$

Proof. See Appendix A.2. □

Theorem 5. *There exists a sequence of alphabets having linear growth for which the GLRT* (2.13) *is not* α*-universally consistent.*

Proof. We let $\mathcal{A}_n = \{1, \ldots, 9n\}$ and will show there exists a pair of $\alpha = 1$ sources for which the GLRT fails. Define distributions

$$p_n(a) = \begin{cases} \frac{1}{2n} & \text{if } a \in \{1, \ldots, n\} \\ \frac{1}{16n} & \text{if } a \in \{n+1, \ldots, 9n\} \end{cases}$$

$$\text{and } q_n(a) = \begin{cases} \frac{5}{4n} & \text{if } a \in \{1, \ldots, n/2\} \\ \frac{1}{4n} & \text{if } a \in \{n/2+1, \ldots, n\} \\ \frac{1}{32n} & \text{if } a \in \{n+1, \ldots, 9n\}. \end{cases}$$

Using Lemma 13, and numerically evaluating the resulting integrals, we see that under hypothesis $\mathcal{H}_0$,

$$\lim_{n\to\infty} \mathbb{E}[D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n)] = 1.085$$

$$\lim_{n\to\infty} \mathbb{E}[D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n)] = 1.026$$

whereas under hypothesis $\mathcal{H}_1$,

$$\lim_{n\to\infty} \mathbb{E}[D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n)] = 1.026$$

$$\lim_{n\to\infty} \mathbb{E}[D(\Lambda_{Y^n}||\hat{q}_n) + D(\Lambda_{Z^n}||\hat{q}_n)] = 0.773.$$

From the Efron-Stein inequality and bounded differences property (Lemma 12), the random variables concentrate around their respective means, which by the previous calculation are converging to the values above. It follows that under hypothesis $\mathcal{H}_0$, the test incorrectly declares $\mathcal{H}_1$. This is illustrated in section 2.3.4. □

Another well-known statistical procedure is chi-squared testing and we turn to that next.

2.3.2 Chi-Squared Test and its Consistency

For any distributions p and q on alphabet $\mathcal{A}$, and any $\mathcal{M} \subseteq \mathcal{A}$ introduce the functional[3]

$$\chi^2(p, q, \mathcal{M}) = \sum_{a\in\mathcal{M}} \frac{(p(a) - q(a))^2}{p(a) + q(a)}.$$

We will usually write $\chi^2(p, q)$ when the set $\mathcal{M}$ is taken for the full alphabet $\mathcal{A}$. Following [71, Ch 17, Ex. 3], one can apply the following chi-squared procedure to the present problem

$$\sum_{a\in\mathcal{A}_n} \frac{(\Lambda_{X^n}(a) - \hat{p}_n(a))^2}{\hat{p}_n(a)} + \frac{(\Lambda_{Z^n}(a) - \hat{p}_n(a))^2}{\hat{p}_n(a)}$$

$$\begin{matrix}\mathcal{H}_1 \\ \gtrless \\ \mathcal{H}_0\end{matrix}$$

$$\sum_{a\in\mathcal{A}_n} \frac{(\Lambda_{Y^n}(a) - \hat{q}_n(a))^2}{\hat{q}_n(a)} + \frac{(\Lambda_{Z^n}(a) - \hat{q}_n(a))^2}{\hat{q}_n(a)}.$$

[3]For $\mathcal{M} = \mathcal{A}$ this functional is sometimes called the *triangular discrimination*, see [70].

After some manipulation, this yields

$$\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \chi^2(\Lambda_{Y^n}, \Lambda_{Z^n}), \tag{2.16}$$

which we will refer to as the chi-squared test (see also [66, Ch.4 §2]).

As with the GLRT, the chi-squared test is consistent with sublinear alphabet growth, in particular for $0 \leq \alpha < 1$. The proof is similar to that of the GLRT, and so only outline the argument.

Theorem 6. *Suppose $|\mathcal{A}_n| = o(n)$, then the chi-squared test* (2.16) *is universally consistent.*

Proof. Suppose hypothesis $\mathcal{H}_0$ is in effect, i.e. $X^n, Y^n, Z^n \sim P_n$. We will show the left side tends to zero in probability, while the other goes to something positive. For brevity we omit writing the alphabet argument in χ^2. Let $\epsilon > 0$. By Lemma taking the first term of the expansion from Lemma 8 we have that

$$D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) \geq \frac{1}{2}\chi^2(\Lambda_{X^n}, \Lambda_{Z^n})$$

therefore the event $\{D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) < \epsilon/2\}$ implies $\chi^2(p, q) < \epsilon$. Thus

$$P_n(\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}) > \epsilon) \leq P_n(D(\Lambda_{X^n}||\hat{p}_n) + D(\Lambda_{Z^n}||\hat{p}_n) > \epsilon/2)$$

which goes to zero according to the proof of Theorem 4.

An easy argument (see Lemma 35 in Appendix A.2) shows that $\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n})$ viewed as a function from $\mathbb{R}^{2n} \to \mathbb{R}$ has the bounded differences property with constant $8n^{-1}$. Also, Jensen's inequality and the joint convexity of the function

$(p-q)^2/(p+q)$ in p, q imply that

$$\begin{aligned}\mathbb{E}_{P_n}\Big[\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n})\Big] &= \sum_a \mathbb{E}_{P_n}\Big[\frac{(\Lambda_{Y^n}(a) - \Lambda_{Z^n}(a))^2}{\Lambda_{Y^n}(a) + \Lambda_{Z^n}(a)}\Big]] \\ &\geq \frac{(\mathbb{E}_{P_n}[\Lambda_{Y^n}(a)] - \mathbb{E}_{P_n}[\Lambda_{Z^n}(a)])^2}{\mathbb{E}_{P_n}[\Lambda_{Y^n}(a)] + \mathbb{E}_{P_n}[\Lambda_{Z^n}(a)]} \\ &= \sum_a \frac{(p_n(a) - q_n(a))^2}{p_n(a) + q_n(a)}.\end{aligned}$$

Now by Cauchy-Schwarz we have

$$\|p_n - q_n\|_1^2 = \Big(\sum_a \frac{|p_n(a) - q_n(a)|}{\sqrt{p_n(a) + q_n(a)}}\sqrt{p_n(a) + q_n(a)}\Big)^2 \leq 2\chi^2(p_n, q_n),$$

therefore Efron-Stein implies the random variable $\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n})$ is concentrated around something strictly greater than $\frac{1}{2}\|p_n - q_n\|_1^2$, which is not tending to zero.

□

We also have a corresponding result about inconsistency of the chi-squared test when $\alpha = 1$.

Lemma 14. *Let $\{p_n, q_n\}$ be a sequence of pairs of distributions and denote by $\mu_n^2(x, y)$ the shadow (see [20]), i.e. the distribution of the random vector $\big(np_n(X_n), nq_n(X_n)\big)$ when $X_n \sim p_n$. If $\mu_n^2(x, y)$ converges weakly to $\mu^2(x, y)$, then under hypothesis $\mathcal{H}_0$ (i.e. $Z^n \sim p_n^n$)*

$$\mathbb{E}[\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n)] \to 2\int \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-x)x^k}{k!}\frac{(j-k)}{j+k}\, d\mu^2(x, y)$$

and

$$\begin{aligned}\mathbb{E}[\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n}, \mathcal{A}_n)] \to &\int \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-y)y^{j-1}}{(j-1)!}\frac{\exp(-x)x^k}{k!}\frac{(j-k)}{j+k}\frac{y}{x}\, d\mu^2(x, y) \\ &+ \int_{C^2} \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-y)y^k}{k!}\frac{(j-k)}{j+k}\, d\mu^2(x, y).\end{aligned}$$

Proof. See Appendix A.2. □

Theorem 7. *There exists a sequence of alphabets having linear growth for which the chi-squared test* (2.16) *is not α-universally consistent.*

Proof. Using the distributions from the proof of Theorem 5, applying Lemma 14, and numerically evaluating the resulting integrals, we see that under hypothesis $\mathcal{H}_0$,

$$\lim_{n\to\infty} \mathbb{E}[\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n)] = 1.57$$
$$\lim_{n\to\infty} \mathbb{E}[\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n}, \mathcal{A}_n)] = 1.49$$

whereas under hypothesis $\mathcal{H}_1$,

$$\lim_{n\to\infty} \mathbb{E}[\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n)] = 1.49$$
$$\lim_{n\to\infty} \mathbb{E}[\chi^2(\Lambda_{Y^n}, \Lambda_{Z^n}, \mathcal{A}_n)] = 1.14.$$

By a similar argument as used in the proof of Theorem 5, it follows that under hypothesis $\mathcal{H}_0$, the test incorrectly declares $\mathcal{H}_1$. □

2.3.3 Understanding the Inconsistency

The inconsistency of both the GLRT and chi-squared test for linear alphabets can be explained neatly by relating these tests to the L_2-norm test $nF \lessgtr 0$, where

$$F = \sum_a n(\Lambda_{X^n}(a) - \Lambda_{Z^n}(a))^2 - \sum_a n(\Lambda_{Y^n}(a) - \Lambda_{Z^n}(a))^2.$$

Recall, from Lemmas 1 and 2 we know that the random variable nF concentrates around values which guarantee consistent detection, i.e. asymptotically $-\mathbb{E}_0[nF] = \mathbb{E}_1[nF] > 0$. But unlike our L_2-norm test, which weights all terms equally (by n), the χ^2 test weights the terms in the first sum of F by

$(\Lambda_{X^n}(a) + \Lambda_{Z^n}(a))^{-1}$ and those in the second sum by $(\Lambda_{Y^n}(a) + \Lambda_{Z^n}(a))^{-1}$. There is no guarantee that the inequality

$$\mathbb{E}_0[\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}) - \chi^2(\Lambda_{Y^n}, \Lambda_{Z^n})] < 0$$

should hold for such weights.

For the case of the GLRT the same reasoning applies by reducing the GLRT to a chi-squared test via a Taylor series expansion, see Lemma 8. For these distributions, numerical calculations show it suffices to restrict attention to the case where the symbol count is zero in the training string and is positive in the test string or vice versa (in fact with high probability $N(a|X^n) = 0$ and $N(a|Z^n) \in \{1, 2, 3\}$ or vice-versa). This observation about the counts combined with Lemma 8 implies

$$G(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}) \approx \log(2)\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}),$$

(Lemma 36 in Appendix A.2 makes this slightly more rigorous).

Another frequently used test in statistics is the Hellinger metric, $h(p, q)$, which for two mass functions p and q is defined via

$$h^2(p, q) = \frac{1}{2}\sum_{a\in\mathcal{A}}(\sqrt{p(a)} - \sqrt{q(a)})^2. \tag{2.17}$$

At first glance one may be tempted to think that the test

$$h^2(\Lambda_{X^n}, \Lambda_{Z^n}) \lessgtr h^2(\Lambda_{Y^n}, \Lambda_{Z^n}) \tag{2.18}$$

would not suffer from the same problems as the chi-squared test and GLRT since it does not involve divisions by empirical distributions. However since $(p - q)^2 = (\sqrt{p} - \sqrt{q})^2(\sqrt{p} + \sqrt{q})^2$, $h(p, q)$ may also be written as

$$h^2(p, q) = \frac{1}{2}\sum_{a\in\mathcal{A}}\frac{(p - q)^2}{(\sqrt{p} + \sqrt{q})^2},$$

and again the test involves divisions by counts. We conjecture (for evidence see the next sub-section) that the Hellinger test is not universally consistent for $\alpha = 1$[4].

2.3.4 Simulation ($\alpha = 1$ case)

In Figure 2.1 we show the empirical performance (over 10000 trials) of the L_2-norm classifier (2.3), the GLRT classifier (2.13), the chi-squared classifier (2.16) and the Hellinger classifier (2.18) for increasing n and a uniform prior on the two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. The alphabet is $\mathcal{A}_n = \{1, \ldots, 9n\}$; Example A refers to the distributions p_n, q_n appearing in the proof of Theorem 5; Example B is the same sequence p_n versus $r_n = 1/(9n)$, the uniform distribution. We see that in Example A the average error probability of the GLRT and chi-squared classifier tends to 1/2, as predicted by Theorems 5 and 7; we also notice the apparent inconsistency of the Hellinger test previously mentioned. In Example B, even though all tests seem to be consistent, the fraction of errors for our new classifier converges to zero more quickly than does the GLRT.

2.4 Testing with Infinite Training Data

In this section we suppose that the tester is given access to an "infinite" amount of training data, i.e. for each n he or she knows $(p_n, q_n, \mathcal{A}_n)$, the underlying

[4]The missing ingredient is the concentration of the random variable $h^2(\Lambda_{X^n}, \Lambda_{Z^n})$ about its mean. Once this is established one can readily verify using a calculation similar to Lemma 14 that the numerical values of the means imply the inconsistency. Concentration would also establish the consistency of the Hellinger test for sub-linear alphabet growth, since the inequality $\chi^2(p, q) \geq 2h^2(p, q)$ [66, Ch.4 §2] implies a proof along the lines of Theorem 6.

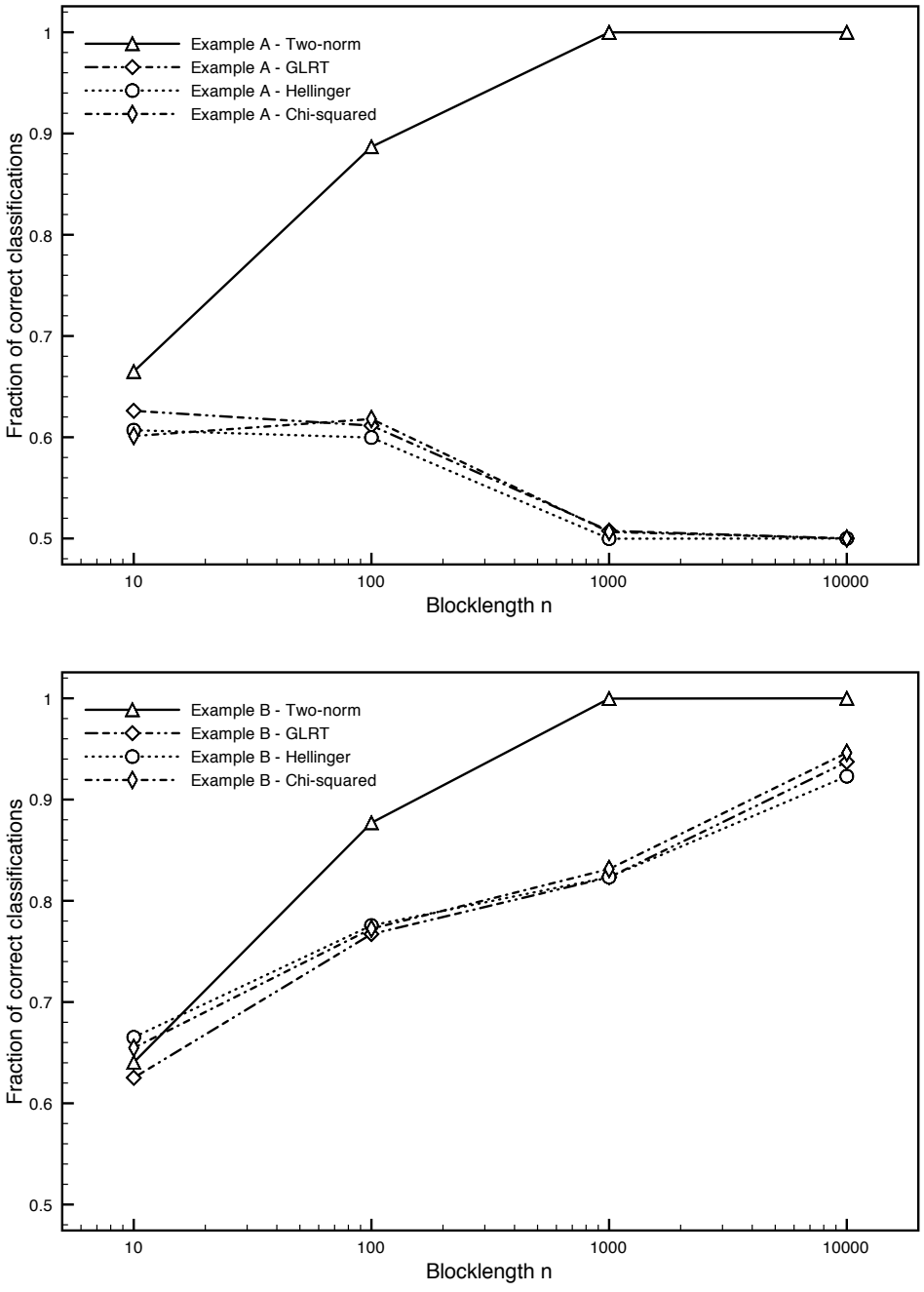

Figure 2.1: Simulation of the performance of L_2-norm versus statistical tests. Example A illustrates the inconsistency of GLRT and Chi-squared (Theorems 5 and 7) and suggests inconsistency of Hellinger test.

distributions and alphabets. The following theorem answers the question for a sequence $\{p_n, q_n, \mathcal{A}_n\}$ satisfying

$$\liminf_{n\to\infty} \sum_{a\in\mathcal{A}_n} |p_n(a) - q_n(a)| > 0,$$

what, if any, are the conditions on the growth rate of the alphabet guaranteeing consistent testing between

$$\mathcal{H}_0 : Z^n \sim p_n^n$$
$$\mathcal{H}_1 : Z^n \sim q_n^n \text{ for all } n.$$

Theorem 8. *For any sequence of alphabets $\{\mathcal{A}_n\}$ and sequence of distributions $\{p_n\}, \{q_n\}$ satisfying*

$$\liminf_{n\to\infty} \sum_{a\in\mathcal{A}_n} |p_n(a) - q_n(a)| > 0$$

the likelihood ratio test

$$p_n^n(X^n) \lessgtr_{\mathcal{H}_0}^{\mathcal{H}_1} q_n^n(X^n)$$

is exponentially consistent, i.e. if

$$P_{e,n} = p_n^n(p_n^n(X^n) < q_n^n(X^n)) + q_n^n(p_n^n(X^n) > q_n^n(X^n))$$

denotes the sum of the type I and type II errors, then

$$\liminf -\frac{1}{n}\log\big(P_{e,n}\big) \geq \liminf \frac{1}{8}\Big(\sum_{a\in\mathcal{A}_n} |p_n - q_n|\Big)^2.$$

Proof. By the Neyman Pearson theory the optimum test is the likelihood ratio test. Invoking Lemma 4 with the point sets $A_n = \{p_n^n\}, B_n = \{q_n^n\}$, we find the minimum error probability for this problem is

$$\mathfrak{R}(A_n, B_n) = 1 - \frac{1}{2}\|p_n^n - q_n^n\|_1.$$

To bound this probability, we follow [72, Cor. 13.1.1] and again make use of the Hellinger metric (2.17). First we recall the inequality (see [73, Ch.3])

$$h^2(p,q) \leq \frac{1}{2}\|p-q\|_1 \leq \sqrt{2}h(p,q). \tag{2.19}$$

For product measures it is well known that the Hellinger metric factorizes (see [73, Ch.3]). Thus in the i.i.d. case

$$h^2(p^n,q^n) = 1-(1-h^2(p,q))^n.$$

Applying these results allows us to write the following chain of inequalities

$$\begin{aligned}
\mathfrak{R}(A_n,B_n) &= 1-\frac{1}{2}\|p_n^n - q_n^n\|_1 \\
&\leq 1-h^2(p_n^n,q_n^n) \\
&= (1-h^2(p_n,q_n))^n \\
&\leq \exp(-nh^2(p_n,q_n)),
\end{aligned}$$

where on the previous line we used the inequality $1+x \leq \exp(x)$. Finally we use the right side of inequality (2.19) to give

$$\mathfrak{R}(A_n,B_n) \leq \exp(-n\frac{1}{8}\|p_n-q_n\|_1^2).$$

But by hypothesis

$$\liminf_{n\to\infty} \|p_n-q_n\|_1 > 0,$$

which gives the result. □

Note that this result extends the classical fixed distribution, fixed alphabet i.i.d. case which states that the testing error, $\mathfrak{R}(\{p^n\},\{q^n\})$, decays exponentially fast with the blocklength n when $p \neq q$ [72, Cor. 13.1.1]. In fact examining the proof we see that $nh^2(p_n,q_n) \to \infty$ is sufficient.

2.5 Beyond α-large-alphabet model

We conclude with some comments on the general-source triangular array hypothesis testing problem (i.e. removing the α-source assumption). Firstly, Theorems 4 and 6 show that the GLRT and chi-squared tests are *universally* consistent (i.e. can handle non-homogeneous sources) provided that the underlying alphabet grows sub-linearly. Using Lemma 10 and bounding the L_2 distances by relative entropies (via Pinkser's inequality), one can also show that the L_2-test (2.3) is also universally consistent with sub-linear alphabet growth, provided that the asymptotic separation occurs in L_2, i.e. the assumption (2.2) is replaced by

$$\liminf_{n\to\infty} \|p_n - q_n\|_2^2 > 0.$$

The counterexample from the proof of Theorem 5 shows that neither the GLRT nor chi-squared test are universally consistent with linear alphabet growth. The following Lemma shows that the L_2-test (2.3) is also inconsistent for inhomogeneous sources with linear alphabet growth.

Lemma 15. *Let $\tilde{p}_n$ and $\tilde{q}_n$ be a sequence of $\alpha = 1$ large alphabet sources, defined on alphabet $\tilde{\mathcal{A}}_n$ such that $n\|\tilde{p}_n - \tilde{q}_n\|_2^2 = \epsilon$ for every n. Denote by ω a special symbol that does not occur in any of $\tilde{\mathcal{A}}_n$ and define*

$$\mathcal{A}_n = \tilde{\mathcal{A}}_n \cup \{\omega\}.$$

Let δ_x denote a point-mass at x and define $p_n = \frac{1}{2}\tilde{p}_n + \frac{1}{2}\delta_\omega$ and $q_n = \frac{1}{2}\tilde{q}_n + \frac{1}{2}\delta_\omega$. Then the test

$$\|\Lambda_{X^n} - \Lambda_{Z^n}\|_2^2 \lesseqgtr \|\Lambda_{Y^n} - \Lambda_{Z^n}\|_2^2$$

is inconsistent.

Proof. See Appendix A.3. □

Dataset	Description	Number of classes
20ng	Usenet articles	20
r52	Reuters newswires	52
r8	Same data as r52, grouped more coarsely	8
webkb	Webpages from a CS department	4

Table 2.1: Datasets used for comparison of classification methods

Roughly speaking the proof uses the fact that the L_2 distance for the $\alpha = 1$ component converges in probability to either $\frac{\epsilon}{4}$ or $-\frac{\epsilon}{4}$, but the variance for the symbol ω is order 1 in probability, and so reliable detection is impossible. Here the problem is that the L_2 test relies on the *unnormalized* counts, and a symbol with large probability can dominate the overall statistic. The GLRT and chi-squared test avoid this problem by using normalized counts, but as we have seen, this normalization can eliminate the bias necessary to ensure consistency.

2.6 Results with real-world datasets

We conclude this chapter by applying the various tests considered here to some real world data. We used the datasets summarized in Table 2.1, which were taken from [74]. Dealing with real-world data requires us to make some departures from the model considered in this chapter. Firstly, each of the datasets comprise multi-class classification problems as opposed to the binary classification considered here. Secondly, the training and test datasets are not all of a common length.

Fortunately each of the tests we considered can be viewed as minimum distance test, e.g. in the case of the GLRT we decide the class i for which

Dataset	L_2	GLRT	Chi-squared	Hellinger
20ng	**22**	15.7	19	16
r52	**6.7**	0	5.6	4.4
r8	**5.0**	1.5	2.6	2.6
webkb	**19**	11.8	13.6	14

Table 2.2: Classification results for "rare" words (words occurring at-most 20 times) only. Figures are percentage of correct classifications

$d(\Lambda_{\mathbf{X}_i}, \hat{P}_i) + d(\Lambda_{\mathbf{Z}}, \hat{P}_i)$ is smallest, where $\hat{P}_i$ is the weighted sum of $\Lambda_{\mathbf{X}_i}$ and $\Lambda_{\mathbf{Z}}$ (with the weighing given by the length of the training document); or in the case of the L_2-norm test we decide the class i for which $\|\Lambda_{\mathbf{X}_i} - \Lambda_{\mathbf{Z}}\|_2^2$ is smallest.

Table 2.2 shows the results when applied to data that loosely "fit" the $\alpha = 1$-large-alphabet model. To obtain these results we took the real-world data and kept only those words that occurred fewer than 20 times. This meant that some common words with possibly high discriminatory power were removed from the test and training sets. The results show the L_2 norm test performing the best of the various distance metrics.

Table 2.3 shows the results using all of the available data. Also included are results for support vector machines (SVM) [23] reported by [74]. The column GLRT(b) corresponds to a tweaked version of the GLRT we devised to correct the poor performance on the `r52` dataset. We observed that when dealing with skewed training sets (i.e. where the lengths of the training data are very different), the GLRT is systematically biased towards the shorter class. For example suppose we have training lengths n_x and n with $n_x \ll n$ and the test string is also length n. The GLRT first forms the quantities

$$\hat{p}(\cdot) = \frac{N(\cdot|X^{n_x}) + N(\cdot|Z^n)}{n + n_x}, \hat{q}(\cdot) = \frac{N(\cdot|Y^n) + N(\cdot|Z^n)}{n + n}$$

Dataset	SVM	L_2	GLRT	GLRT(b)	Chi-squared	Hellinger
20ng	80.8	52	81.7	**82.7**	74.8	60.4
r52	92	84.1	1.9	91.2	86.1	74.8
r8	**94.5**	91.2	87.5	**96.2**	94.2	90.9
webkb	87.9	74.3	91.1	**91.7**	82.5	78.7

Table 2.3: Classification results for full datasets. Figures are percentage of correct classifications.

and then carries out the test

$$\frac{n_x}{n} D(\Lambda_{X^{n_x}} || \hat{p}) + D(\Lambda_{Z^n} || p) \lessgtr D(\Lambda_{Y^n} || \hat{p}) + D(\Lambda_{Z^n} || p).$$

When the true hypothesis is that Y^n and Z^n are from the same class (i.e. have the same distribution) we observed that the GLRT incorrectly decided for the case that X^{n_x} and Z^n were from the same class. A reason for this appeared to be that $D(\Lambda_{Z^n} || \hat{p})$ was small because $\hat{p} \approx \Lambda_{Z^n}$. By "repeating" the training data, so that all strings were the same length, e.g. by forming

$$\tilde{p}(\cdot) = \frac{\frac{n}{n_x} N(\cdot | X^{n_x}) + N(\cdot | Z^n)}{n + n}$$

we found the bias disappeared, and these are reported as GLRT(b) in Table 2.3. As can be seen from the table, GLRT(b) performs quite well, outperforming the published SVM results in all but one example.

CHAPTER 3

COMPRESSION OF OF LARGE ALPHABET SOURCES

In this chapter we formulate and study the problem of compression of large alphabet sources. A connection between the results of the previous chapter and the present chapter is apparent if one notices that the set of distributions used in the converse part Theorem 11 are in-fact $\alpha = 1$ large-alphabet distributions. Thus another interpretation for the results in this chapter is that for $0 \leq \alpha < 1$ universal compression of α-large-alphabet sources is possible; for $\alpha = 1$ it is not.

3.1 Notation and Preliminaries

Throughout logarithms and exponents are in base e. For a distribution p on a finite alphabet $\mathcal{A}$, we use $H(p) = \sum_{a \in \mathcal{A}} -p(a) \log p(a)$ to denote entropy. The notation $\mathcal{A}^{\times n}$ is the n-fold Cartestian product of $\mathcal{A}$. We use bold type to denote strings (or vectors), e.g. $\mathbf{x} = x_1 \cdots x_n$, usually the length is clear from the context and will be omitted. We use $\Lambda_{\mathbf{x}}$ to denote the empirical distribution or type of the string $\mathbf{x}$. $H_2(x)$ denotes the binary entropy function. For a probability distribution p, $\operatorname{supp} p$ denotes the support of p i.e. the set of symbols having positive probability. $\mathcal{P}(\mathcal{A})$ denotes the set of all distributions on the set $\mathcal{A}$. $\mathcal{P}^n(\mathcal{A})$ denotes the set of possible empirical distributions for a string of length n on the alphabet $\mathcal{A}$. For a type $Q \in \mathcal{P}^n(\mathcal{A})$, we use T_Q^n to denote the typeclass of Q, i.e. the set of strings with type Q.

We mainly consider sequences of alphabets $\{\mathcal{A}_n\}$ and distributions $\{p_n \in \mathcal{P}(\mathcal{A}_n)\}$. In this case, unless specified otherwise, when we write the random variable X^n, we mean the nth row of a corresponding triangular array

$\{X_{n,m}, 1 \leq m \leq n\}_{n\geq 1}$, so that $X^n = X_{n,1}, \ldots, X_{n,n}$ and $X_{n,i} \sim p_n$.

We first formalize the notation of an achievable rate sequence.

Definition 6. Let $\{\mathcal{A}_n\}$ be a sequence of finite alphabets. For a sequence of distributions $\{Q_n\}$, where Q_n is defined on the product space $\mathcal{A}_n^{\times n}$, we say a sequence of rates $\{R_n\}$ is *achievable (for source coding)* if for every $\delta > 0, \epsilon > 0$, there exist a sequence of sets $\{\mathcal{M}_n\}$ and a sequence of deterministic maps $\{f_n : \mathcal{A}_n^{\times n} \to \mathcal{M}_n, g_n : \mathcal{M}_n \to \mathcal{A}_n^{\times n}\}$ satisfying

$$\frac{1}{n} \log |\mathcal{M}_n| < R_n + \delta$$

and

$$Q_n(g_n(f_n(X^n)) \neq X^n) \leq \epsilon$$

for all n sufficiently large.

Remark: For a given sequence of distributions $\{Q_n\}$, it is straightforward to verify that a sequence of rates $\{R_n\}$ is achievable with deterministic maps iff $\{R_n\}$ is achievable with randomized maps.

Using information-spectrum methods [33], the following theorem provides a second characterization of an achievable rate sequence.

Theorem 9. *Let $\{\mathcal{A}_n\}$ be a sequence of finite alphabets. Let $\{Q_n\}$ be a sequence of probability measures such that Q_n is a measure on the product space $\mathcal{A}_n^{\times n}$. Suppose $X^n \sim Q_n$. Then the sequence $\{R_n\}$ is achievable for source coding if and only if for every $\delta > 0$*

$$\lim_{n\to\infty} Q_n\Big(-\frac{1}{n} \log Q_n(X^n) - R_n > \delta \Big) = 0. \tag{3.1}$$

Proof. Achievability: Suppose that (3.1) holds. Let $G_n^\delta = \{\mathbf{x} \in \mathcal{A}_n^{\times n} : -n^{-1}\log Q_n(\mathbf{x}) \leq R_n + \delta\}$. Then

$$1 \geq Q_n(G_n^\delta) \geq |G_n^\delta| \exp(-n(R_n + \delta))$$

which implies

$$\frac{1}{n}\log |G_n^\delta| \leq R_n + \delta.$$

Furthermore, by hypothesis, as $n \to \infty$

$$Q_n({G_n^\delta}^c) \to 0,$$

so that defining f_n, g_n to identify those sequences in G_n^δ suffices.

Converse: Assume that there exists $\delta > 0$ and $\epsilon > 0$ so that

$$\limsup_{n\to\infty} Q_n\Big(-\frac{1}{n}\log Q_n(X^n) - R_n > \delta\Big) > \epsilon.$$

Let

$$B_n = \{\mathbf{x} : -n^{-1}\log Q_n(\mathbf{x}) > R_n + \delta\}.$$

By Definition 6, the achievability of R_n implies the existence of a sequence of sets $A_n \triangleq \{\mathbf{x} : g_n(f_n(\mathbf{x})) = \mathbf{x}\}$ satisfying $Q_n(A_n^c) \leq \epsilon/4$ for all n sufficiently large. Now, $Q_n(A_n \cap B_n) \geq Q_n(B_n) - Q_n(A_n^c)$ and $Q_n(A_n \cap B_n) < |A_n \cap B_n| \exp(-n(R_n + \delta))$, which together gives

$$\begin{aligned} n^{-1}\log |A_n| &\geq n^{-1}\log |A_n \cap B_n| \\ &> n^{-1}\log[Q_n(A_n \cap B_n)] + R_n + \delta \\ &\geq n^{-1}\log[Q_n(B_n) - Q_n(A_n^c)] + R_n + \delta. \end{aligned}$$

However, for a subsequence $\{n_k\}$ we have that $Q_{n_k}(B_{n_k}) \geq \epsilon/2$, thus $Q_{n_k}(B_{n_k}) - Q_{n_k}(A_{n_k}^c) > \epsilon/4$ for all $n_k \geq n_0$. Therefore for all $n_k \geq n_0$

$${n_k}^{-1}\log |\mathcal{M}_{n_k}| \geq {n_k}^{-1}\log |A_{n_k}| > R_{n_k} + \delta/2,$$

i.e. $n^{-1}\log|\mathcal{M}_n| > R_n+\delta/2$ for infinitely many n, contradicting the achievability of $\{R_n\}$. □

Corollary 1. *If there is a code $f : \mathcal{A}_n^{\times n} \to \mathcal{M}_n, g : \mathcal{M}_n \to \mathcal{A}_n^{\times n}$ with rate $n^{-1}\log|\mathcal{M}_n| \leq R_n$ and probability of error $Q_n(g(f(X^n)) \neq X^n) \leq \epsilon$ then*

$$Q_n(-n^{-1}\log Q_n(X^n) > R_n + \delta) \leq \epsilon + \exp(-n\delta).$$

Proof. This is implied by the calculations in the converse part of the proof Theorem 9. Adopting the definitions from that proof we saw that

$$Q_n(B_n) \leq Q_n(A_n^c) + Q_n(A_n \cap B_n).$$

By hypothesis $Q_n(A_n^c) \leq \epsilon$. Furthermore

$$\begin{aligned} Q_n(A_n \cap B_n) &= \sum_{\mathbf{x}\in A_n\cap B_n} Q_n(\mathbf{x}) \\ &\leq \sum_{\mathbf{x}\in A_n\cap B_n} \exp(-n[R_n+\delta]) \\ &\leq \exp(-n\delta) \end{aligned}$$

where the final equality uses the fact that the range of f is at most $\exp(nR_n)$. □

3.2 Universal Compression of Large Alphabet Sources

We begin by defining universal compression and then study the cases of sublinear and linear alphabet growth.

Definition 7 ($\{\mathcal{A}_n, R_n\}$-Universal Codes)**.** The pair $\{\mathcal{A}_n, R_n\}$ admits a deterministic (respectively random) universal code if for every $\delta > 0$ and $\epsilon > 0$, there

exists an integer n_0 such that for all $n > n_0$ there is a deterministic (resp. random) encoder/decoder pair $(f_n : \mathcal{A}_n^{\times n} \to \mathcal{M}_n, g_n : \mathcal{M}_n \to \mathcal{A}_n^{\times n})$ with rate $R_n + \delta$ such that for all $p_n \in \mathcal{P}(\mathcal{A}_n)$, when $X^n \sim p_n^n$

$$\Pr(g_n(f_n(X^n)) \neq X^n) \leq \epsilon + \min_{\tilde{f}_n, \tilde{g}_n} \Pr(\tilde{g}_n(\tilde{f}_n(X^n)) \neq X^n)$$

where the minimum is over all deterministic codes with rate R_n.

Definition 8. A sequence of alphabets $\{\mathcal{A}_n\}$ supports *deterministic (resp. random) universal compression* if for every rate sequence $\{R_n\}$ the sequence $\{\mathcal{A}_n, R_n\}$ admits a deterministic (resp. random) universal code.

3.2.1 Sublinear Alphabet Growth

For sub-linear alphabet growth we have the following positive result.

Theorem 10. *If $\{\mathcal{A}_n\}$ is a sequence of alphabets satisfying $|\mathcal{A}_n| = o(n)$ then $\{\mathcal{A}_n\}$ supports deterministic universal compression.*

Proof. Let $\delta > 0$ and $\epsilon > 0$ be arbitrary. Define

$$\mathcal{M}_n = \{1, \ldots, |\mathcal{P}^n(\mathcal{A}_n)|\} \times \{1, \ldots, \lceil \exp(n[R_n + \delta/2]) \rceil\}$$

Define the encoder f_n as follows. Let f_n first send the type Λ_{X^n} and then for types Λ_{X^n} satisfying

$$H(\Lambda_{X^n}) \leq R_n + \delta/2$$

send the index of the sequence within the typeclass (recall that the number of sequences in a typeclass, T_Q^n is at most $\exp(nH(Q))$ [27]); otherwise send an arbitrary index. The decoder, g_n, declares an arbitrary output if the type is such

that $H(\Lambda_{X^n}) > R_n + \delta/2$, otherwise it can decode X^n unambiguously. In total the scheme requires

$$\begin{aligned} n^{-1}\log M_n &\leq R_n + \delta/2 + o(n) + n^{-1}\log|\mathcal{P}^n(\mathcal{A}_n)| \\ &\leq R_n + \delta \text{ nats}, \end{aligned}$$

for all n sufficiently large, using the fact that $n^{-1}\log|\mathcal{P}^n(\mathcal{A}_n)| \to 0$ if $|\mathcal{A}_n| = o(n)$ [12, Lem. 1]. Therefore the inequality in the previous display holds for all $n \geq n_1$ (and n_1 depends only on $\mathcal{A}_n$).

Let us now fix a sequence of distributions $\{p_n \in \mathcal{P}(\mathcal{A}_n)\}$. Notice that an error can occur only when $\{H(\Lambda_{X^n}) > R_n + \delta/2\}$. Therefore

$$\begin{aligned} p_n^n(g_n(f_n(X^n)) \neq X^n) &\leq p_n^n(H(\Lambda_{X^n}) > R_n + \delta/2) \\ &= p_n^n(H_n - R_n > \delta/2 + H_n - H(\Lambda_{X^n})), \end{aligned}$$

where we introduced the term $H_n = -n^{-1}\log p_n^n(X^n)$. Notice that the identity [27, Lem. 2.6]

$$-\frac{1}{n}\log p^n(\mathbf{x}) = D(\Lambda_{\mathbf{x}}||p) + H(\Lambda_{\mathbf{x}})$$

implies that

$$H_n - H(\Lambda_{X^n}) = D(\Lambda_{X^n}||p_n),$$

and therefore since $D(p||q) \geq 0$,

$$p_n^n(H_n - R_n > \delta/2 + H_n - H(\Lambda_{X^n})) \leq p_n^n(H_n - R_n > \delta/2).$$

Now let

$$e_n^* = \min_{\tilde{f}_n, \tilde{g}_n} p_n^n(\tilde{g}_n(\tilde{f}_n(X^n)) \neq X^n),$$

where the minimum is over all rate R_n codes. Then Corollary 1 implies

$$p_n^n(H_n - R_n > \delta/2) \leq e_n^* + \exp(-n\delta/2).$$

Finally, since $\exp(-n\delta/2) \leq \epsilon$ for all $n \geq n_2$ we have shown that for all $n \geq n_0 \triangleq \max(n_1, n_2)$

$$p_n^n(g_n(f_n(X^n)) \neq X^n) \leq \epsilon + e_n^*.$$

This completes the proof. □

3.2.2 Linear Alphabet Growth

We next show that linear growth is the best possible.

Theorem 11. *The sequence of alphabets* $\{\mathcal{A}_n \triangleq \{1, 2, \ldots, n\}\}$, $n = 2, 4, \ldots$ *does not support random universal compression.*

To prove this result we will exhibit a collection of i.i.d. sources on $\{\mathcal{A}_n\}$, each of which is individually compressible at the same rate R_n, but for which universal compression demands rates strictly bounded away from R_n.

Throughout this subsection we will take n to be an even integer, $\mathcal{A}_n = \{1, \ldots, n\}$, and $\mathcal{E}_n$ will denote the collection of subsets of $\mathcal{A}_n$ having size exactly $n/2$, i.e.

$$\mathcal{E}_n = \{E \subset \mathcal{A}_n : |E| = n/2\}.$$

To each $E \in \mathcal{E}_n$, we can associate a probability measure on $\mathcal{A}_n$ in the following manner

$$u_{n,E}(a) = \begin{cases} \frac{2}{n} & \text{if } a \in E \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

We will let $\mathcal{U}_n$ denote the set of all such measures. Observe that by counting arguments

$$|\mathcal{E}_n| = |\mathcal{U}_n| = \binom{n}{\frac{n}{2}}.$$

An object of key interest will be the probability measure Q_n defined on the space $\mathcal{A}_n^{\times n}$ via

$$Q_n(\mathbf{x}) = \sum_{p_n \in \mathcal{U}_n} \frac{p_n^n(\mathbf{x})}{|\mathcal{U}_n|}.$$

This measure has the following Bayesian interpretation. Let π_n be the uniform measure on $\mathcal{E}_n$. Choose an $E \in \mathcal{E}_n$ according to π_n and then generate a string X^n i.i.d. according to $u_{n,E}$ (cf. (3.2)). The marginal distribution of X^n under this scheme is then given by Q_n.

To prove Theorem 11, we require the following lemmas. The proofs of some results are omitted due to space constraints.

Lemma 16. *Let $p_n \in \mathcal{U}_n$, suppose that $X^n \sim p_n^n$ and let $J_n = |\operatorname{supp} \Lambda_{X^n}|$. Then*

$$p_n^n\Big(J_n < \frac{n}{2}\Big[1 - \Big(1 - \frac{2}{n}\Big)^n - e^{-3}\Big]\Big) \leq \exp(-ne^{-6}/2).$$

Proof. Notice that changing any symbol in X^n changes J_n by at most one. Therefore J_n satisfies the hypotheses of McDiarmid's inequality [75]. Note that

$$\begin{aligned}
\mathbb{E}[J_n] &= \sum_{a \in \mathcal{A}_n} p_n^n(N(a|X^n) > 0) \\
&= \sum_{a \in \mathcal{A}_n} \Big[1 - \Big(1 - \frac{2}{n}\Big)^n\Big] \mathbf{1}\{a \in \operatorname{supp} p_n\} \\
&= \frac{n}{2}\Big[1 - \Big(1 - \frac{2}{n}\Big)^n\Big].
\end{aligned}$$

Applying McDiarmid's inequality, with $\epsilon = ne^{-3}/2$ gives the result. □

Lemma 17. *Let X, Y be two random variables on finite sets $\mathcal{X}$ and $\mathcal{Y}$ with joint distribution p_{XY}. Then for every $y \in \mathcal{Y}$*

$$H(X|Y = y) \leq \log |\{x : p(x|y) > 0\}|.$$

Proof. Fix $y \in \mathcal{Y}$. Let u be the uniform probability distribution over the set $\{x : p(x|y) > 0\}$. Observe that

$$\begin{aligned}\sum_{x\in\mathcal{X}} P_{X|Y}(x|y) \log \frac{P_{X|Y}(x|y)}{u(x)} &= \sum_{x\in\{x:p(x|y)>0\}} P_{X|Y}(x|y) \log \frac{P_{X|Y}(x|y)}{u(x)} \\ &= \log |\{x : p(x|y) > 0\}| - H(X|Y = y)\end{aligned}$$

Jensen's inequality implies that

$$0 \leq \sum_{x\in\{x:p(x|y)>0\}} P_{X|Y}(x|y) \log \frac{P_{X|Y}(x|y)}{u(x)}$$

giving the result. □

Lemma 18. *Let $E_n \sim \pi_n$ and $X^n|E_n \sim u_{E_n,n}$ and define $\gamma_n = \left(1 - \frac{2}{n}\right)^n + e^{-3}$. Then*

$$\frac{1}{n} H(E_n|X^n) \leq H_2(\gamma_n/2) + (\log 2) \exp(-ne^{-6}/2).$$

for all $n \geq 2$.

Proof. Define the set

$$C_n = \left\{\mathbf{x} : |\operatorname{supp} \Lambda_{\mathbf{x}}| < \frac{n}{2}(1 - \gamma_n)\right\}.$$

Notice that under the scheme of the Lemma $X^n \sim Q_n$, therefore

$$\frac{1}{n} H(E_n|X^n) = \frac{1}{n} \sum_{\mathbf{x}\in C_n} H(E_n|X^n = \mathbf{x}) Q_n(\mathbf{x}) \tag{3.3}$$

$$+ \frac{1}{n} \sum_{\mathbf{x}\in C_n^c} H(E_n|X^n = \mathbf{x}) Q_n(\mathbf{x}). \tag{3.4}$$

Applying Lemma 17 to the first term of (3.3) gives

$$\begin{aligned}\frac{1}{n} \sum_{\mathbf{x}\in C_n} H(E_n|X^n = \mathbf{x}) Q_n(\mathbf{x}) &\leq \frac{1}{n} (\log |\mathcal{E}_n|) Q_n(C_n) \\ &\leq (\log 2) Q_n(C_n),\end{aligned}$$

where on the previous line we used the fact that for all $1 \leq k \leq n$

$$\binom{n}{k} \leq \exp(nH_2(k/n)). \tag{3.5}$$

Now applying Lemma 16 to bound $Q_n(C_n)$ gives

$$\frac{1}{n}\sum_{\mathbf{x}\in C_n} H(E_n|X^n=\mathbf{x})Q_n(\mathbf{x}) \leq (\log 2)\exp(-ne^{-6}/2).$$

We now turn our attention to the second term of (3.3). Notice that for $\mathbf{x} \in C_n^c$, the event $\{X^n = \mathbf{x}\}$ implies that the random support set E_n must contain at least those elements that occur in $\mathbf{x}$. On the set C_n^c the "sparsest" $\mathbf{x}$ contains

$$\left\lceil \frac{n}{2}\Big(1-\gamma_n\Big)\right\rceil$$

distinct values. Therefore conditional on $\{X^n = \mathbf{x}\}$, E_n can take on at most

$$\binom{n-\lceil \frac{n}{2}(1-\gamma_n)\rceil}{\frac{n}{2}-\lceil \frac{n}{2}(1-\gamma_n)\rceil} \leq \binom{n}{\frac{n}{2}-\lceil \frac{n}{2}(1-\gamma_n)\rceil}$$

values and applying Lemma 17 gives

$$\begin{aligned}\frac{1}{n}\sum_{\mathbf{x}\in C_n^c} H(E_n|X^n=\mathbf{x})Q_n(\mathbf{x}) &\leq \frac{1}{n}\log\binom{n}{\frac{n}{2}-\lceil \frac{n}{2}(1-\gamma_n)\rceil}\sum_{\mathbf{x}\in C_n^c} Q_n(\mathbf{x}) \\ &\leq \frac{1}{n}\log\binom{n}{\frac{n}{2}-\lceil \frac{n}{2}(1-\gamma_n)\rceil}.\end{aligned}$$

Again using (3.5) to bound the binomial coefficient gives

$$\frac{1}{n}\sum_{\mathbf{x}\in C_n^c} H(E_n|X^n=\mathbf{x})Q_n(\mathbf{x}) \leq H_2\Big(\frac{\frac{n}{2}-\lceil \frac{n}{2}(1-\gamma_n)\rceil}{n}\Big).$$

Since $0 < 1-\gamma_n < 1$ for all $n \geq 2$, applying the inequality $\lceil x \rceil \geq x$ and using the monotonicity of $H_2(p)$ for $0 \leq p \leq 1/2$ we get

$$H_2\Big(\frac{\frac{n}{2}-\lceil \frac{n}{2}(1-\gamma_n)\rceil}{n}\Big) \leq H_2(\gamma_n/2).$$

□

Lemma 19. *Let* $H_n = -\frac{1}{n}\log Q_n(X^n)$. *Then*

$$H_n \leq \log n + \frac{\log|\mathcal{U}_n|}{n} - \log 2 \quad \textit{a.s.}$$

Proof. Let $\mathbf{x} \in \mathcal{A}_n^{\times n}$ be a string with positive Q_n probability. Then

$$\begin{aligned} Q_n(\mathbf{x}) &= \frac{1}{|\mathcal{U}_n|}\sum_{p\in\mathcal{U}_n} p^n(\mathbf{x}) \\ &\geq \frac{1}{|\mathcal{U}_n|}\Big(\frac{2}{n}\Big)^n, \end{aligned}$$

where we used the fact that the string has probability $(2/n)^n$ under at least one measure in the collection $\mathcal{U}_n$. Therefore

$$\begin{aligned} -\frac{1}{n}\log Q_n(X^n) &\leq -\frac{1}{n}\Big[\log\frac{1}{|\mathcal{U}_n|} + \log\Big(\frac{2}{n}\Big)^n\Big] \\ &= \frac{1}{n}\log|\mathcal{U}_n| - \log 2 + \log n. \end{aligned}$$

□

Proof of Theorem 11. Let $\beta > 0$. Then observe that for any $\{p_n \in \mathcal{U}_n\}$ there is a sequence of codes with rate sequence $\{R_n = \log n - \log 2 + \beta\}$ that makes no error at each n: with this much rate we can simply map each possible string onto a unique element in the message set. Now let $\delta > 0, \epsilon > 0$ be given. The existence of a $\{\mathcal{A}_n, R_n\}$-random universal code implies that there exists random maps $\{f_n, g_n\}$ and n_0 so that for all $n > n_0$ both

$$\frac{1}{n}\log|\mathcal{M}_n| < R_n + \delta \tag{3.6}$$

and

$$\Pr(g_n(f_n(X^n)) \neq X^n) \leq \epsilon, X^n \sim p_n^n, \text{ for all } p_n \in \mathcal{U}_n. \tag{3.7}$$

Let $\gamma_n(f,g)$ denote the probability distribution of the random code f_n, g_n. Notice that (3.7) implies that for all $n > n_0$

$$\begin{aligned}\epsilon &\geq \frac{1}{|\mathcal{U}_n|}\sum_{p_n\in\mathcal{U}_n}\sum_{\tilde{f}_n,\tilde{g}_n}\gamma_n(\tilde{f}_n,\tilde{g}_n)p_n^n(\tilde{g}_n(\tilde{f}_n(X^n))\neq X^n)\\ &= \sum_{\tilde{f}_n,\tilde{g}_n}\gamma_n(\tilde{f}_n,\tilde{g}_n)Q_n(\{\mathbf{x}:\tilde{g}_n(\tilde{f}_n(\mathbf{x}))\neq\mathbf{x}\}).\end{aligned}$$

Viewing the last display as an expectation, it follows there must be at least one deterministic code f_n^*, g_n^* for which

$$Q_n(\{\mathbf{x}: g_n^*(f_n^*(\mathbf{x}))\neq\mathbf{x}\})\leq\epsilon \text{ for all } n>n_0. \tag{3.8}$$

Recalling Definition 6, we see that (3.6) and (3.8) tell us that the existence of $\{f_n, g_n\}$ (and in particular the specific deterministic maps $\{f_n^*, g_n^*\}$) implies that $\{R_n\}$ is achievable for the mixture probability measure Q_n. However, we will proceed to show that for some $\beta > 0$ and all n sufficiently large

$$Q_n\Big(-\frac{1}{n}\log Q_n(X^n) - R_n > \beta\Big) > C(\beta) > 0,$$

i.e. that $\{R_n\}$ is not achievable for $\{Q_n\}$ by Theorem 9.

Suppose that $E_n \sim \pi_n$ and $X^n|E_n \sim p_{E_n,n}$. By the chain rule for entropy we see that

$$H(X^n) = H(E_n) + H(X^n|E_n) - H(E_n|X^n). \tag{3.9}$$

Let $H_n = -\frac{1}{n}\log Q_n(X^n)$ and observe that $\mathbb{E}[H_n] = \frac{1}{n}H_{Q_n}(X^n)$. From equation (3.9) and Lemma 18 we have

$$\frac{1}{n}H_{Q_n}(X^n)\geq\frac{1}{n}\log\binom{n}{\frac{n}{2}}+\log\frac{n}{2}-H_2(\gamma_n/2)-\log 2\exp(-ne^{-6}/2),$$

where $\gamma_n = \left(1-\frac{2}{n}\right)^n + e^{-3}$. We now recall the reverse Markov inequality: for any $X \leq a$ a.s. and $d < \mathbb{E}[X]$

$$\Pr(X>d)\geq\frac{\mathbb{E}[X]-d}{a-d}.$$

Applying this inequality with $X = H_n$, $d = \log n - \log 2 + 2\beta$ and $a = \log n + \frac{\log |\mathcal{U}_n|}{n} - \log 2$ (Lemma 19), where β will be chosen to ensure $d < \mathbb{E}[X]$, we see that for $n \geq 2$

$$\begin{aligned} Q_n(H_n > d) &\geq \frac{n^{-1} \log \binom{n}{\frac{n}{2}} + \log \frac{n}{2} - H_2(\gamma_n/2) - (\log 2) \exp(-ne^{-6}/2) - \log \frac{n}{2} - 2\beta}{\log n + \frac{\log |\mathcal{U}_n|}{n} - \log 2 - \log n + \log 2 - 2\beta} \\ &= \frac{n^{-1} \log \binom{n}{\frac{n}{2}} - H_2(\gamma_n/2) - (\log 2) \exp(-ne^{-6}/2) - 2\beta}{\frac{\log |\mathcal{U}_n|}{n} - 2\beta} \end{aligned}$$

Taking the liminf as $n \to \infty$ gives

$$\liminf_{n \to \infty} Q_n(H_n > d) \geq \frac{H_2(1/2) - H_2((e^{-2} + e^{-3})/2) - 2\beta}{H_2(1/2) - 2\beta},$$

which is strictly positive provided that $0 < 2\beta < H_2(1/2) - H_2((e^{-2} + e^{-3})/2) \approx 0.385$. We conclude that $\mathcal{A}_n$ does not support random universal compression.

□

3.3 Universal Compression with Distributional Side Information

We next show that informing the decoder of the true source distribution is as good as informing both the encoder and the decoder. That is, there exist codes with a universal encoder that perform asymptotically as well as codes for which both the encoder and decoder are tailored for the particular source distribution. Our result requires random codes; it would be interesting to determine whether this is also possible with deterministic codes, or whether there is a fundamental difference between deterministic and randomized codes, as occurs in, for example, arbitrarily varying channels [27, Sec. 2.6].

Theorem 12. *For any sequences $\{\mathcal{A}_n\}$, $\{R_n\}$, and real numbers $\delta > 0$ and $\epsilon > 0$ there exists an integer n_0 such that for all $n > n_0$ there is a random encoder/decoder pair $(f_n : \mathcal{A}_n^{\times n} \to \mathcal{M}_n, g_n : \mathcal{M}_n \times \mathcal{P}(\mathcal{A}_n) \to \mathcal{A}_n^{\times n})$ with rate $R_n + \delta$ such that for all $p_n \in \mathcal{P}(\mathcal{A}_n)$, when $X^n \sim p_n^n$*

$$\Pr(g_n(f_n(X^n), p_n) \neq X^n) \leq \epsilon + \min_{\tilde{f}_n, \tilde{g}_n} \Pr(\tilde{g}_n(\tilde{f}_n(X^n)) \neq X^n)$$

where the minimum is over all deterministic codes with rate R_n.

Proof. Let $\delta > 0, \epsilon > 0$ be given. The encoder assigns each to $\mathbf{x} \in \mathcal{A}_n^{\times n}$ an index in $\{1, \ldots, \lceil \exp(n[R_n + \delta]) \rceil\}$ uniformly at random. Let $B(i)$ denote the set of sequences assigned to index i and $U(\mathbf{x})$ denote the index assigned to sequence $\mathbf{x}$. The encoder sends $i = U(X^n)$. Using the received i, the decoder declares its output as the $\mathbf{x} \in B(i) \cap G_n(p_n)$ if $|B(i) \cap G_n(p_n)| = 1$, where

$$G_n(p_n) = \{\mathbf{x} : -n^{-1} \log p_n^n(\mathbf{x}) \leq R_n + \delta/2\};$$

and declares an arbitrary string otherwise.

By the union bound it follows that the error probability is upper-bounded by

$$\Pr(\{\exists \tilde{\mathbf{x}} \neq X^n \in G_n(p_n) : U(X^n) = U(\tilde{\mathbf{x}})\}) + \Pr(X^n \notin G_n(p_n)). \tag{3.10}$$

Applying the union bound to the first term of (3.10) gives

$$\begin{aligned}
&\sum_{\mathbf{x} \in G_n(p_n)} \Pr(U(X^n) = U(\mathbf{x})) \\
&\leq |G_n(p_n)| \frac{1}{\lceil \exp(n[R_n + \delta]) \rceil} \\
&\leq \exp(n[R_n + \delta/2]) \frac{1}{\lceil \exp(n[R_n + \delta]) \rceil} \\
&= \exp(-n\delta/2).
\end{aligned}$$

Note that this term is smaller than $\epsilon/2$ for all n larger than some n_1. Now we turn to the second term of (3.10). Let

$$e_n^* = \min_{\tilde{f}_n, \tilde{g}_n} p_n^n(\tilde{g}_n(\tilde{f}_n(X^n)) \neq X^n),$$

where the minimum is over all rate R_n codes. Corollary 1 implies

$$\Pr(X^n \notin G_n(p_n)) = p_n^n(H_n - R_n > \delta/2) \leq e_n^* + \exp(-n\delta/2).$$

Notice that $\exp(-n\delta/2) \leq \epsilon/2$ for all n larger than n_1. Therefore for any $p_n \in \mathcal{P}(\mathcal{A}_n)$

$$\Pr(\{\exists \tilde{\mathbf{x}} \in G_n(p_n) : U(X^n) = U(\mathbf{x})\}) + \Pr(X^n \notin G_n(p_n)) \leq \epsilon + e_n^*$$

for $n \geq n_1$, which concludes the proof. □

3.4 Non-universal compression

In the classical fixed-distribution setting, the entropy of an i.i.d. source with distribution p, $H(p)$, is the fundamental limit as far as compression is concerned. In the large alphabet setting, a natural question is when the sequence of rates $\{H(p_n)\}$ is achievable. We show that it is provided $n^{-1/2} \log |\mathcal{A}_n| \to 0$.

Lemma 20. *Suppose* $|\mathcal{A}| = 2$, *then*

$$\max_{p \in \mathcal{P}(\mathcal{A})} \sum_{a \in \mathcal{A}} p(a) \log^2 p(a) \leq 1.$$

Suppose $|\mathcal{A}| \geq 3$, *then*

$$\max_{p \in \mathcal{P}(\mathcal{A})} \sum_{a \in \mathcal{A}} p(a) \log^2 p(a) = \log^2 |\mathcal{A}|.$$

Lemma 21. *For any distribution p on a finite alphabet $\mathcal{A}$ and any $\epsilon > 0$*

$$p^n\Big(|-\frac{1}{n}\log p^n(X^n) - H(p)| > \epsilon\Big) \leq \frac{\max(\log^2|\mathcal{A}|, 1)}{n\epsilon^2}$$

Proof. We start by noticing that $E[H_n] = H(p)$. Applying Chebyshev's inequality to the random variable H_n gives

$$p^n\Big(|H_n - \mathbb{E}[H_n]| > \epsilon\Big) \leq \frac{\mathrm{Var}(H_n)}{\epsilon^2}.$$

Next we notice that

$$\begin{aligned}\mathrm{Var}(H_n) &= n^{-2}\mathrm{Var}\Big(\sum_{i=1}^{n} -\log p(X_i)\Big)\\ &\leq n^{-1}\mathbb{E}[(-\log p(X_i))^2].\end{aligned}$$

Finally, since

$$\mathbb{E}[(-\log p(X_i))^2] = \sum_{a\in\mathcal{A}} p(a)\log^2 p(a),$$

invoking Lemma 20 gives the result. □

This immediately implies:

Theorem 13. *For any sequence of distributions $\{p_n\}$ on alphabets $\{\mathcal{A}_n\}$ satisfying*

$$n^{-\frac{1}{2}}\log|\mathcal{A}_n| \to 0$$

the sequence of rates $\{H(p_n)\}$ is achievable.

The growth rate of Theorem 13 is indeed the best we can do.

Theorem 14. *Suppose $\{\mathcal{A}_n = \{1, 2, \ldots, |\mathcal{A}_n|\}\}$ satisfies*

$$\limsup_{n\to\infty} n^{-\frac{1}{2}}\log|\mathcal{A}_n| = c > 0.$$

Define $m_n = |\mathcal{A}_n| - 1$ *and*

$$p_n(a) = \begin{cases} \frac{1}{2} & \text{if } a = 1 \\ \\ \frac{1}{2m_n} & \text{otherwise.} \end{cases}$$

Then for any $\delta > 0$*,*

$$\liminf_{n\to\infty} p_n^n\Big(-\frac{1}{n}\log p_n^n(X^n) - H(p_n) > \delta\Big) > 1 - \Phi((4/c)\delta) > 0,$$

where $\Phi(\cdot)$ *is the CDF of a standard normal random variable.*

Proof. Define $I_{n,i} = -\log p_n(X_{n,i})$ and $H_n = n^{-1}\sum_{i=1}^n I_{n,i}$. Notice that $I_{n,i}$ is a random variable with distribution

$$\Pr(I_{n,i} = x) = \begin{cases} \frac{1}{2} & \text{if } x = \log 2 \\ \\ \frac{1}{2} & \text{if } x = \log 2 + \log m_n. \end{cases}$$

Furthermore we have that

$$H(p_n) = \log 2 + \frac{1}{2}\log m_n$$

Let $J_n = |\{i : I_{n,i} = \log 2 + \log m_n\}|$. The event of interest maybe be written $\{J_n(\log 2 + \log m_n) + (n - J_n)\log 2 > nH(p_n) + n\delta\}$. After substituting the value of $H(p_n)$, dividing through by $(1/2)\sqrt{n}$ and simplifying, this is equivalent to the event

$$\Big\{\frac{J_n - n/2}{(1/2)\sqrt{n}} > \frac{\sqrt{n}2\delta}{\log m_n}\Big\}.$$

Let $\{n_k\}$ be a subsequence such that $\lim_{k\to\infty} n_k^{-\frac{1}{2}}\log|\mathcal{A}_{n_k}| = c$. Then it follows that for all k sufficiently large $\frac{\sqrt{n_k}}{\log m_{n_k}} < 2/c$ and therefore

$$p_{n_k}^{n_k}\Big(\frac{J_{n_k} - n_k/2}{(1/2)\sqrt{n_k}} > \frac{\sqrt{n_k}2\delta}{\log m_n}\Big) \geq p_{n_k}^{n_k}\Big(\frac{J_{n_k} - n_k/2}{(1/2)\sqrt{n_k}} > (4/c)\delta\Big).$$

Now since J_n is a sum of n independent Bernoulli(1/2) random variables, the central limit theorem implies the result. □

CHAPTER 4

RELIABILITY IN SOURCE CODING WITH SIDE INFORMATION

In this chapter we study error exponents in the classical fixed distribution (and fixed alphabet) asymptotic. As mentioned in Chapter 1, our goal is to understand optimal communication/compression schemes, where optimality refers to the best decay of the error probability.

4.1 Definitions and Notations

We use $\mathcal{P}(\mathcal{X})$ to denote the set of discrete probability distributions on $\mathcal{X}$ and $\mathcal{C}(\mathcal{X} \to \mathcal{Y})$ to denote all channels from $\mathcal{X}$ to $\mathcal{Y}$. For $P \in \mathcal{P}(\mathcal{X})$ and $V \in \mathcal{C}(\mathcal{X} \to \mathcal{Y})$, we write $P \times V$ to denote the distribution of the pair $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ in which X is generated according to $P(\cdot)$ and Y is taken as the output of the channel V whose input is X. For $P \in \mathcal{P}(\mathcal{X})$ and $P_{Y|X} \in \mathcal{C}(\mathcal{X} \to \mathcal{Y})$ we use P_{XY} as shorthand for $P_X \times P_{Y|X}$.

We use $\mathbf{x}$ to denote vectors in $\mathcal{X}^n$; usually the length of the vector is clear from the context. For any $\mathbf{x} \in \mathcal{X}^n$ we write $Q_{\mathbf{x}}(\cdot)$ as the empirical distribution or *type* of $\mathbf{x}$. The set of all sequences of length n with type Q is denoted T_Q^n. The set of all type variables $Q \in \mathcal{P}(\mathcal{X})$, i.e. those for which $T_Q^n \neq \emptyset$, is denoted $\mathcal{P}^n(\mathcal{X})$. For $Q \in \mathcal{P}^n(\mathcal{X})$, we let $\mathcal{C}^n(Q, \mathcal{Y})$ denote the set of all $W \in \mathcal{C}(\mathcal{X} \to \mathcal{Y})$ for which (1) $T_{Q \times W}^n$ is non-empty; and (2) in the case that $Q(x) = 0$, $W(\cdot|x)$ takes the form $W(y|x) = \xi(y)/2^{-n}$, for any choice of $\xi(y)$ so that $\sum_y \xi(y) = 2^n$. For $\mathbf{x} \in \mathcal{X}^n$ and $V \in \mathcal{C}(\mathcal{X} \to \mathcal{Y})$ we denote by $T_V^n(\mathbf{x})$, the set of sequences in $\mathcal{Y}^n$ having conditional type V given $\mathbf{x}$. For a type Q_Y, $k(Q_Y)$ returns a unique index for that type.

Throughout, when dealing with discrete random variables, all logarithms and exponents have base 2. We take $0 \log 0 = 0$ and $\log 0 = -\infty$ based on continuity arguments. For a distribution or type P we let $H(P)$ denote entropy. For strings $\mathbf{x}, \mathbf{y}$, we write $H(\mathbf{x}|\mathbf{y})$ as the conditional empirical entropy. For a distribution P_X and a channel $P_{Y|X}$ we write $I(P_X; P_{Y|X})$ for the mutual information between X and Y supposing that $P_X \times P_{Y|X}$ governs the pair. $D(P||Q)$ denotes the Kullback-Leibler (KL) divergence between distributions P and Q. We also use the standard definitions of conditional entropy, conditional mutual information, and conditional KL divergence.

Whenever the range of a summation, maximization or minimization is clear we will use shorthand, e.g. $\sum_{Q_X \in \mathcal{P}^n(\mathcal{X})} = \sum_{Q_X}$. We define $(x)^+ \triangleq \max(0, x)$.

For the Gaussian Wyner-Ziv problem logarithms and exponents have base e. For K a variance or covariance matrix, we write f_K as a shorthand for a $\mathcal{N}(0, K)$ Gaussian random density. For $(X, Y) \sim f_K$, we write $f_{K_{Y|X}}$ for the conditional distribution of Y given X and write $K_{Y|X}$ for the conditional covariance (matrix). $h(K)$ denotes the differential entropy of a Gaussian random variable with distribution f_K. A subscript K denotes that expectation or mutual information should be computed using f_K. Additional definitions, facts and proofs for the Gaussian version of the Wyner-Ziv problem can be found in Appendix B.4.

4.2 SCPSI Results and Discussion

Let (X_i, Y_i) be the output of a memoryless source with distribution $P_{XY}(x, y)$ on a finite alphabet $\mathcal{X} \times \mathcal{Y}$. The first encoder observes only the i.i.d sequence X^n, the second encoder observes only the i.i.d sequence Y^n. The decoder, g^n :

$\mathcal{M}_1 \times \mathcal{M}_2 \to \mathcal{X}^n$ must reproduce X^n using the messages from the encoders. The encoders are deterministic functions $f_1^n : \mathcal{X}^n \to \mathcal{M}_1$ and $f_2^n : \mathcal{Y}^n \to \mathcal{M}_2$.

For this set-up the rate region was determined by Ahlswede and Körner [40] and by Wyner [39] who showed that R_1, R_2 are achievable if

$$\exists S - Y - X \text{ s.t. } R_1 \geq H(X|S), R_2 \geq I(Y;S).$$

The closure of the union of the pairs over all such S gives the entire rate region.

Let the decoder output be denoted $\hat{X}^n = g^n(f_1^n(X^n), f_2^n(Y^n))$, then error probability is

$$P_e(f_1^n, f_2^n, g^n) = P_{XY}^n(\hat{X}^n \neq X^n),$$

and we define the source coding with partial side information error exponent as

$$\eta(P_{XY}, R_1, R_2) = \lim_{\epsilon \downarrow 0} \limsup_{n \to \infty} -\frac{1}{n} \log \left[\min_{f_1^n, f_2^n, g^n} P_e(f_1^n, f_2^n, g^n) \right], \tag{4.1}$$

where the minimization ranges over all encoders and decoders f_1^n, f_2^n, g^n, such that

$$\log \mathcal{M}_i \leq n(R_i + \epsilon). \tag{4.2}$$

Our main results for SCPSI are as follows.

Theorem 15. *Let $R_1, R_2, P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be given. Then*

$$\eta(P_{XY}, R_1, R_2) \geq \eta_L(P_{XY}, R_1, R_2) \triangleq \tag{4.3}$$

$$\inf_{Q_Y} \sup_{\substack{Q_{S|Y}: \\ I(Q_Y;Q_{S|Y}) \leq R_2}} \inf_{\substack{Q_{X|YS}: \\ H(Q_X) > R_1}} D(Q_{XYS}||P_{XY}Q_{S|Y}) + [R_1 - H(Q_{X|S}|Q_S)]^+ \tag{4.4}$$

where the joint distribution of X, Y, S is $Q_Y Q_{S|Y} Q_{X|YS}$ and S takes finitely many values.

The scheme to achieve this exponent is explained in detail in the appendix. In brief, operating on a type by type basis, the second encoder generates a codebook with 2^{nR_2} codewords chosen uniformly from $T^n_{Q_S}$ and uses this to compress the side information. The primary encoder bins the X sequences and transmits the index of the bin containing the sequence. The decoder declares its output as the source sequence in the received bin that has the smallest empirical entropy conditional on the compressed side information.

Theorem 16. *Let $R_1, R_2, P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be given, and suppose that $P_{XY}(x, y) > 0$ for all x and y. Then*

$$\eta(P_{XY}, R_1, R_2) \leq \eta_U(P_{XY}, R_1, R_2) \triangleq \inf_{Q_Y} \sup_{\substack{Q_{S|Y}: \\ I(Q_Y;Q_{S|Y}) \leq R_2}} \inf_{\substack{Q_{X|Y}: \\ H(Q_{X|S}|Q_S) > R_1}} D(Q_{XY}||P_{XY}) \tag{4.5}$$

where the joint distribution of X, Y, S is $Q_Y Q_{X|Y} Q_{S|Y}$, i.e. X, Y and S form a Markov chain in that order, and S satisfies

$$|\mathcal{S}| \leq |\mathcal{X}| \cdot |\mathcal{Y}| + |\mathcal{Y}| + 2. \tag{4.6}$$

4.2.1 Discussion

Both theorems can be viewed as a competitive game between two players, nature and the code designer. Nature's goal is to minimize the exponent and the code designer's goal is to maximize it. The particular problem under consideration determines the parameters and order of the plays. For example in Theorem 15, nature plays first, choosing a "worst-case" side information distribution. Then knowing nature's choice the code designer picks the best codebook (via its choice of test channel). Nature plays last, choosing the worst possible consistent

joint distribution. Notice that the choices at each step match the "information" available to the players.

A standard application of the change-of-measure argument [27, p.g. 268] provides the following upper-bound on the SCPSI exponent

$$\eta(P_{XY}, R_1, R_2) \leq \eta_{SP}(P_{XY}, R_1, R_2) \triangleq \tag{4.7}$$

$$\inf_{Q_{XY}} \sup_{\substack{Q_{S|Y}: \\ I(Y;S)\leq R_2}} \begin{cases} D(Q_{XY}||P_{XY}) & H(Q_{X|S}|Q_S) > R_1 \\ \infty & H(Q_{X|S}|Q_S) \leq R_1. \end{cases} \tag{4.8}$$

One can show that $\eta_U \leq \eta_{SP}$, and so formally η_U provides an improvement upon the standard sphere-packing upper bound. In the game theoretic interpretation the η_{SP} exponent is obtained by letting nature's play reveal the *joint distribution* of the source and side information, and then the code designer plays, choosing the best codebook. But in the SCPSI problem, the code designer knows only the marginal type of the side information, i.e. our improved upper bound better captures the inherent structure of the problem.

The optimizations in Theorems 15 and 16 differ in several respects. Foremost, in Theorem 16 the inner-most optimization is over $Q_{X|Y}$, so that X, Y, S adhere to the Markov structure, yet in the achievable exponent this Markov constraint is not present. This differing Markov structure is also present in the partial Wyner-Ziv exponent results of Jayaraman and Berger [47, 76] who attribute the gap between sphere packing and random exponents (present even at low rates) to this type of difference in the Markov structure. The other differences between η_L and η_U are the range of the inner most optimization and the presence of the binning term in the achievable exponent.

Despite of these differences, the bounds provided by the theorems do allow

us to determine the error exponent exactly in some special cases. When $R_2 = 0$, there is no possibility of encoding the side information. Taking $S = 0$ a.s., in both exponents, one recovers the standard point to point exponent

$$\inf_{\substack{Q_X: \\ H(X) \geq R_1}} D(Q_X || P_X).$$

Similarly, when $R_2 \geq \log |\mathcal{Y}|$, the second encoder may send all of the side information to the decoder. In this case our achievable exponent recovers the Oohama and Han [57] exponent, which has been shown to be tight near the boundary of the Slepian-Wolf rate region (see also Gallager [52]).

4.3 Wyner-Ziv Results and Discussion

Let (X_i, Y_i) be the output of a memoryless source with distribution $P_{XY}(x, y)$ on a finite alphabet $\mathcal{X} \times \mathcal{Y}$. Let $\hat{\mathcal{X}}$ be the reproduction alphabet and $d : \mathcal{X} \to \hat{\mathcal{X}}$ a single letter distortion measure. Define the distortion between two strings as $d(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$.

An encoder observes the i.i.d. source sequence, X^n and communicates a message using nR bits (or nats) to the decoder. The decoder combines the message with the side information Y^n to give its reproduction $\hat{X}^n$. The encoder/decoder pair are functions $\psi : \mathcal{X}^n \to \mathcal{M}$ and $\varphi : \mathcal{M} \times \mathcal{Y}^n \to \hat{\mathcal{X}}^n$, where $\mathcal{M}$ is a fixed set.

The rate region was determined by Wyner and Ziv [77], who showed that if the allowable distortion is Δ, then the required rate is given by

$$R_{WZ}(P_{XY}, \Delta) = \inf I(X; Z) - I(Y; Z),$$

where the infimum is over all auxiliary random variables Z such that (1) Z, X, and Y form a Markov chain in this order and (2) there exists a function λ such

that

$$\mathbb{E}[d(X, \lambda(Y, Z))] \leq \Delta.$$

Let $\hat{X}^n = \varphi(\psi(X^n), Y^n)$ be the decoder's output and define the error probability

$$P_e(\psi, \varphi, \Delta, d) = \Pr\left(d(X^n, \hat{X}^n) > \Delta\right). \tag{4.9}$$

We define the Wyner-Ziv error exponent to be

$$\theta(R, \Delta, P_{XY}, d) = \lim_{\epsilon \downarrow 0} \limsup_{n \to \infty} -\frac{1}{n} \log \left[\min_{(\psi, \varphi)} P_e(\psi, \varphi, \Delta, d)\right] \tag{4.10}$$

where the minimization ranges over all encoder/decoder pairs satisfying

$$\log |\mathcal{M}| \leq n(R + \epsilon). \tag{4.11}$$

Our main results for the Wyner-Ziv problem are as follows.

Discrete Memoryless Case

Theorem 17. *Let $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and $R > 0$, $\Delta > 0$, $d(\cdot, \cdot)$ be given. Then*

$$\theta(R, \Delta, P_{XY}, d) \geq \inf_{Q_X} \sup_{Q_{Z|X}} \inf_{Q_Y} \sup_{f \in \mathcal{F}} \inf_{Q_{XYZ}} G_D\left[Q_{XYZ}, P_{XY}, f, d, \Delta, R\right] \tag{4.12}$$

where

$$G_D\left[Q_{XYZ}, P_{XY}, f, d, \Delta, R\right] = \begin{cases} D(Q_{XYZ}||P_{XY}Q_{Z|X}) & \mathbb{E}_Q[d(X, f(Y,Z))] \geq \Delta \\ D(Q_{XYZ}||P_{XY}Q_{Z|X}) & \\ \quad +\big(R - I(Q_X; Q_{Z|X}) & \\ \quad\quad +I(Q_Y; Q_{Z|Y})\big)^+ & \mathbb{E}_Q[d(X, f(Y,Z))] < \Delta \\ & I(Q_X; Q_{Z|X}) \geq R \\ \infty & \textit{otherwise}, \end{cases}$$

$\mathcal{F} = \{f | f : \mathcal{Y} \times \mathcal{Z} \to \hat{\mathcal{X}}\}$, and Z takes finitely many values. Note in the final minimization over Q_{XYZ}, Q_{XZ} and Q_Y are fixed to be those specified earlier in the optimization.

For completeness, we state the upper bound, which can be proved easily following Marton's [58] sphere-packing/change-of-measure proof for the point-to-point case.

Theorem 18. *Let $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and $R > 0$, $\Delta > 0$, $d(\cdot, \cdot)$ be given. Then*

$$\theta(R, \Delta, P_{XY}, d) \leq \inf_{Q_{XY}: R_{WZ}(Q_{XY}, \Delta) > R} D(Q_{XY}||P_{XY}).$$

This result is analogous to the upper bound in (4.7) and is therefore not as strong as its SCPSI counterpart (cf. (4.5)). We expect that this bound can be improved, although the technique used to obtain Theorem 16 does not seem to be applicable here. If this bound can be strictly improved in the binary erasure case, it would imply an exponent loss (see Section 4.4.1).

Gaussian Case

Theorem 19. *Let (X_i, Y_i) be jointly Gaussian with zero means and covariance matrix*

$$\Sigma = \begin{bmatrix} 1 & \zeta_{XY} \\ \zeta_{XY} & 1 \end{bmatrix}, \tag{4.13}$$

and let $d(x, \hat{x}) = (x - \hat{x})^2$. Then for any $R > 0$, $\Delta > 0$, and Σ as in (4.13)*,*

$$\theta(R, \Delta, f_\Sigma, d) \geq \inf_{\sigma_X^2} \sup_{\rho_{xz}} \inf_{\sigma_Y^2} \sup_{\lambda \in \Lambda} \inf_{\rho_{yz}, \rho_{xy}} G_G\left[K, \Sigma, \lambda, \Delta, R\right] \tag{4.14}$$

where

$$G_G[K, \Sigma, \lambda, \Delta, R] = \begin{cases} D(K||\bar{K}) & \mathbb{E}_K[(X - \lambda(Y, Z))^2] \geq \Delta \\ D(K||\bar{K}) & \\ \quad +\big(R - I_K(X; Z) & \mathbb{E}_K[(X - \lambda(Y, Z))^2] < \Delta \\ \quad +I_K(Y; Z)\big)^+ & I_K(X; Z) \geq R \\ \infty & \text{otherwise,} \end{cases} \tag{4.15}$$

$\Lambda = \{\lambda : \mathbb{R} \times \mathbb{R} \to \mathbb{R} : \lambda(y, z) = \alpha y + \beta z, \alpha, \beta \in [-M_\lambda, M_\lambda]\}$*, the covariance matrix of (X, Y, Z) is*

$$K = \begin{bmatrix} \sigma_X^2 & \sigma_X \sigma_Y \rho_{xy} & \sigma_X \rho_{xz} \\ \sigma_X \sigma_Y \rho_{xy} & \sigma_Y^2 & \sigma_Y \rho_{yz} \\ \sigma_X \rho_{xz} & \sigma_Y \rho_{yz} & 1 \end{bmatrix}$$

and

$$\bar{K} = \begin{bmatrix} 1 & \zeta_{XY} & \frac{\rho_{xz}}{\sigma_X} \\ \zeta_{XY} & 1 & \zeta_{XY} \frac{\rho_{xz}}{\sigma_X} \\ \frac{\rho_{xz}}{\sigma_X} & \zeta_{XY} \frac{\rho_{xz}}{\sigma_X} & \frac{\rho_{xz}^2}{\sigma_X^2} + 1 - \rho_{xz}^2 \end{bmatrix}. \tag{4.16}$$

$M_\lambda > 0$ is an arbitrary real number. The covariance matrix $\bar{K}$ corresponds to a source (X, Y, Z), where $X, Y \sim \mathcal{N}(0, \Sigma)$, Z, X and Y form a Markov chain in that order, and the distribution of Z conditional on X is taken from K.

Theorem 20. *Let (X_i, Y_i) be jointly Gaussian with zero means and covariance Σ as in* (4.13)*. Let $R_{X|Y}(f_\Sigma, \Delta)$ denote the conditional rate distortion function. Let $\tilde{\theta}$ denote the error exponent for a modified Gaussian Wyner-Ziv problem in which the side information is also available at the encoder. Then for any, $\Delta > 0$, $R > R_{X|Y}(f_\Sigma, \Delta)$*

$$\tilde{\theta}(R, \Delta, f_\Sigma, d) \leq \inf_{\Pi: R_{X|Y}(f_\Pi, \Delta) \geq R} D(\Pi||\Sigma) \tag{4.17}$$

where Π is a 2×2 positive definite covariance matrix and

$$R_{X|Y}(f_\Pi, \Delta) = R_{WZ}(f_\Pi, \Delta) = \frac{1}{2} \log^+ \left(\frac{\mathrm{Var}_\Pi(X|Y)}{\Delta} \right).$$

Corollary 2. *Under the assumptions of Theorem 20, we have that*

$$\theta(R, \Delta, f_\Sigma, d) \leq \tilde{\theta}(R, \Delta, f_\Sigma, d).$$

Proof. Any code that works for the Wyner-Ziv problem will work when the encoder also sees the side information. This implies that the error exponent for the Wyner-Ziv problem is upper bounded by the error exponent for the two-sided problem. □

The upper bound in the Corollary is identical to the change-of-measure upper bound obtained via Theorem 18. As with that bound, we believe that this upper bound can be improved, and showing a strict improvement would establish an exponent loss.

4.3.1 Discussion

As in the SCPSI case, in the Wyner-Ziv case the same game-theoretic interpretation holds, but there are more parameters and the game becomes more

elaborate. Nature plays first, choosing the most "difficult" source distribution. The code designer plays next, selecting the "best" test channel for that difficult source. Nature plays again choosing the worst marginal distribution for the side information. Then, knowing everything so far, the code designer chooses the estimation function. Nature has the final play, choosing the worst consistent joint distribution for triple X, Y, Z. Once again the choices and order of plays match the problem.

The nature of the optimizations in Theorems 17 and 19 give us some insight into the design of practical coding schemes by revealing a tension, which we examine in detail in the next section for the binary erasure and Gaussian problems. Briefly we see that the objective functions G_D (resp. G_G) contain three cases which correspond to

- a violation of the distortion constraint even when the codeword is decoded correctly;
- the use of binning, leading to the potential for decoding the wrong codeword;
- no possibility for error.

A large codebook allows for a cleaner quantization and hence lower chance of the first kind of event. But this large codebook comes with the requirement of binning, leading to the potential for the second kind of event. Thus these two kinds of errors are in tension.

Theorem 17 allows us to determine a portion of the reliability function for a certain functional source coding problem. If we wish to reproduce a function $g(X)$ of the source X losslessly at the decoder, then the rate required is

$H_P(g(X)|Y)$, which follows from the results of Orlitsky and Roche [78]. Setting the distortion measure to be

$$d(X, f(Y, Z)) = d_H(g(X), f(Y, Z))$$

(d_H is the hamming measure) and evaluating Theorem 17 in the limit as $\Delta \to 0$ provides an achievable exponent for this problem. This can be seen by always choosing $Q_{Z|X}$ so that $Z = g(X)$ and letting the reproduction function be $f(Y, Z) = Z$. Using the fact that $Z \leftrightarrow X \leftrightarrow Y$, one can show that limit as $\Delta \to 0$ of the righthand-side of equation (4.12) is

$$\xi_L(R, P_{XY}) = \inf_{Q_{XY}: H_Q(g(X)) \geq R} D(Q_{XY}||P_{XY}) + (R - H_Q(g(X)|Y))^+. \tag{4.18}$$

One can prove a change-of-measure/sphere-packing argument (say, again along the lines of Marton [58]), which yields the following upper bound on the error exponent for this problem

$$\xi_U(R, P_{XY}) = \inf_{Q_{XY}: H_Q(g(X)|Y) \geq R} D(Q_{XY}||P_{XY}). \tag{4.19}$$

On account of the fact that both (4.18) and (4.19) are optimizations of a continuous function over a compact sets, the $\inf$ is attained. Therefore it follows that the relationship between these two functions is analogous to the relationship between the sphere-packing and random coding exponents in channel coding [27, Lemma 2.5.4]. Thus for $R \geq 0$ until some critical rate R_c the reliability function for the functional source coding problem is given exactly by

$$\inf_{Q_{XY}: H_Q(g(X)|Y) \geq R} D(Q_{XY}||P_{XY}).$$

4.4 Examples

4.4.1 Binary Erasure Case

As an application of Theorem 17, we turn to the binary erasure version of the Wyner-Ziv problem. In this case, X is uniformly distributed over the set $\{-1, +1\}$, and Y equals X passed through a binary erasure channel with erasure probability p

$$P(Y = 0|X = 1) = p = 1 - P(Y = 1|X = 1)$$
$$P(Y = 0|X = -1) = p = 1 - P(Y = -1|X = -1).$$

We would like to permit the reconstruction string to have erasures but not errors. The reconstruction alphabet is thus

$$\hat{\mathcal{X}} = \{-1, 0, 1\}.$$

One way to avoid errors in the reconstruction string is to use the "erasure" distortion measure

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

This distortion measure is overly harsh, however, in that it prohibits all errors. For the Wyner-Ziv problem, higher rates can be achieved if one tolerates a vanishing probability of error. We will therefore consider a finite approximation of

this distortion measure,

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

where K is a large but fixed constant. We will examine the rate-distortion and reliability functions in the limit as K tends to infinity.

To determine the rate-distortion function in this case, let Z be the output of a binary erasure channel with input X and erasure probability δ. If Z, X, and Y form a Markov chain in this order, then it follows that

$$I(X;Z) - I(Y;Z) = p(1-\delta).$$

There is a natural choice of f for this case

$$f(y,z) = \begin{cases} 1 & \text{if } z = 1 \text{ or } y = 1 \\ 0 & \text{if } z = 0 \text{ and } y = 0 \\ -1 & \text{otherwise.} \end{cases} \tag{4.20}$$

Then $\mathbb{E}[d(X, f(Y,Z))] = p\delta$, and so any rate

$$R \geq (p - \Delta)^+$$

is achievable. To see that this is in fact the best possible, consider the problem in which the side information Y^n is available to both the encoder and the decoder. The rate-distortion function for this problem is given by

$$\min_{p(\hat{x}|x,y)} I(X;\hat{X}|Y).$$

such that

$$\mathbb{E}[d(X,\hat{X})] \leq \Delta.$$

This minimization can be computed using classical techniques and shown in the limit as K tends to infinity to equal $(p-\Delta)^+$. It follows that $(p-\Delta)^+$ is the rate-distortion function for both problems. In particular, there is no "rate loss" in the sense that the rate-distortion function is the same whether the side information is available at both the encoder and decoder or at the decoder only.

We note that for the problem with side information at both the encoder and decoder, there is a simple scheme that achieves the rate-distortion function $(p-\Delta)^+$. Since the encoder knows the locations of the erasures in Y^n, it can simply communicate the value of X^n in the first nR erased locations.

We turn to the application of Theorem 17 to this set-up. For simplicity of exposition, we will consider the optimization problem in (4.12) with two restrictions: (1) Q_X is fixed to be the uniform distribution over $\{-1,+1\}$; and (2) we optimize $Q_{Z|X}$ over the class of binary erasure channels, instead of optimizing over the class of all test channels from $\mathcal{X}$ to $\mathcal{Z}$. The optimization problem in (4.12) then reduces to

$$\sup_{Q_{Z|X}} \min_{Q_{Y|XZ}} G[Q_{XYZ}, P_{XY}, f, \Delta, R].$$

This optimization problem can be written in the following alternative form

$$\sup_{Q_{Z|X}} \min(G_1(Q_{Z|X}), G_2(Q_{Z|X})), \tag{4.21}$$

where

$$G_1(Q_{Z|X}) = \min_{Q_{Y|XZ}} D(Q_{XYZ}||P_{XY}Q_{Z|X})$$

with the minimization being over all $Q_{Y|XZ}$ such that

$$\mathbb{E}_Q[d(X, f(Y,Z))] \geq \Delta,$$

and

$$G_2(Q_{Z|X}) = \min_{Q_{Y|XZ}} D(Q_{XYZ}||P_{XY}Q_{Z|X}) + (R - I_Q(X;Z) + I_Q(Y;Z))^+,$$

with the optimization being over all $Q_{Y|XZ}$ such that

$$\mathbb{E}_Q[d(X, f(Y, Z))] < \Delta,$$

and

$$I_Q(X; Z) \geq R.$$

This last condition, of course, either holds for all choices of $Q_{Y|XZ}$ or for none of them.

The alternative form of the optimization problem given in (4.21) is useful because it shows that maximizing over the binary erasure test channel amounts to maximizing the minimum of the exponents of two error events: the first, $G_1(Q_{Z|X})$, is the exponent on the event that Y^n and Z^n together provide insufficient information about X^n to enable the decoder to meet the distortion constraint. Thus an error will occur even if the codeword Z^n is decoded correctly. The second, $G_2(Q_{Z|X})$, is the exponent on the probability of a binning error.

These two error exponents are in tension in the following sense. Choosing $Q_{Z|X}$ to have a low probability of erasure communicates many of the bits in X^n to the decoder via Z^n. This makes it unlikely that Y^n and Z^n will reveal too few bits about X^n for the decoder to meet the distortion constraint, meaning that $G_1(Q_{Z|X})$ will be large. At the same time, choosing $Q_{Z|X}$ to have a low probability of erasure requires the use of large codebook, which makes the binning error probability high, leading to a small $G_2(Q_{Z|X})$. On the other hand, choosing $Q_{Z|X}$ to have a high probability of erasure leads to exactly the opposite behavior: the binning error probability is small since little information is being communicated through Z^n, but it is much more likely that the realization of Y^n and Z^n do not collectively reveal enough of the bits in X^n to meet the distortion constraint.

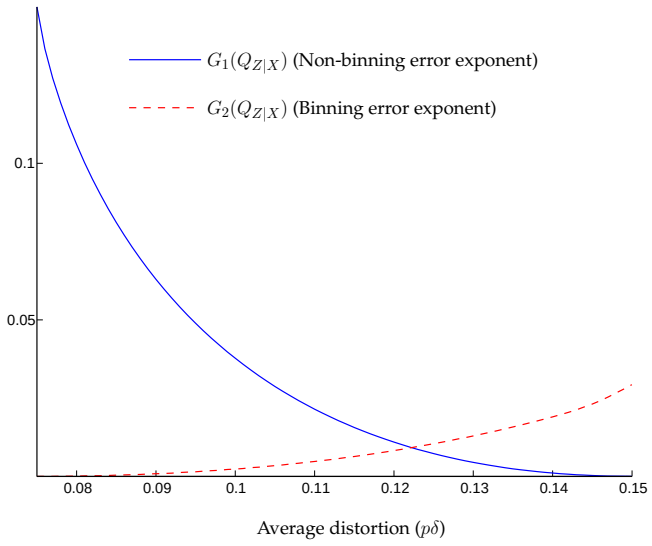

Figure 4.1: Tension in choice of the test channel erasure probability δ, revealed by Theorem 17. Note that $p\delta$ is the average distortion of the system. Here $\Delta = 0.15$, $p = 0.5$, and $R = 0.425$.

This tension is illustrated in Fig. 4.1. The optimum choice of $Q_{Z|X}$ is given by a moderate erasure probability that balances the exponents of the two error probabilities. With this choice, both are dominant error events.

The exponent itself is shown for various R in Fig. 4.2. Since we have not optimized over Q_X, this is properly interpreted as an upper bound on the error exponent of the scheme. Fig. 4.2 also shows the error exponent of the simple scheme mentioned above for achieving the rate-distortion function when the side information is available at both the encoder and the decoder[1]. The error probability of this scheme is simply the probability that Y^n contains more than $n(R + \Delta)$ erasures. Assuming $R > p - \Delta$, the exponent of this event is equal to

$$D(R + \Delta||p),$$

i.e., the relative entropy between two Bernoulli distributions, one with success

[1]This is also the upper bound in Theorem 18.

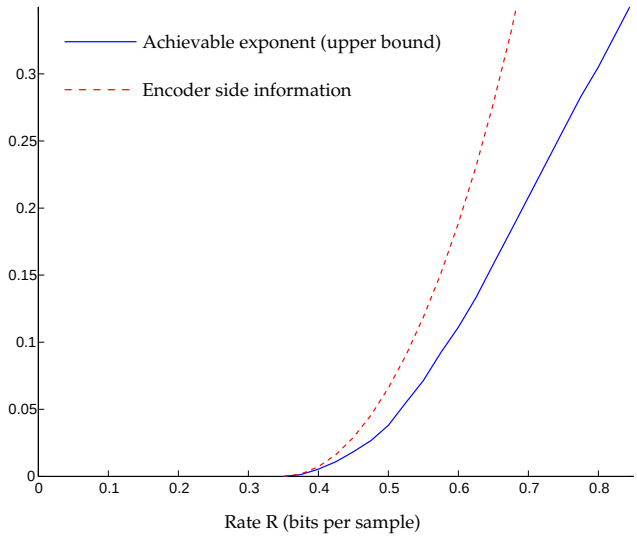

Figure 4.2: Upper bound on error exponent of Theorem 17, and the error exponent of the scheme that makes use of side information at the encoder. The parameters Δ, p are the same as those used in Fig. 4.1.

probability $R + \Delta$ and one with success probability p. Fig. 4.2 shows that when the side information is available at both the encoder and decoder the exponent is higher than for our one-sided scheme. This suggests that there is an exponent loss.

4.4.2 Gaussian Case

A similar test channel tension arises in the Gaussian case. This can be seen most clearly by considering the optimization problem over ρ_{xz} for fixed σ_X^2. In Fig. 4.3 we plot

$$G_3(\rho_{xz}) = \inf_{\sigma_Y^2} \sup_{\lambda \in \Lambda} \inf_{\rho_{xy}, \rho_{yz}} G\left[K, \Sigma, \lambda, \Delta, R\right]$$

where we hold $\sigma_X^2 = 1$, and $K = K(1, \sigma_Y, 1, \rho_{xy}, \rho_{yz}, \rho_{xz})$ is the covariance matrix of (X, Y, Z).

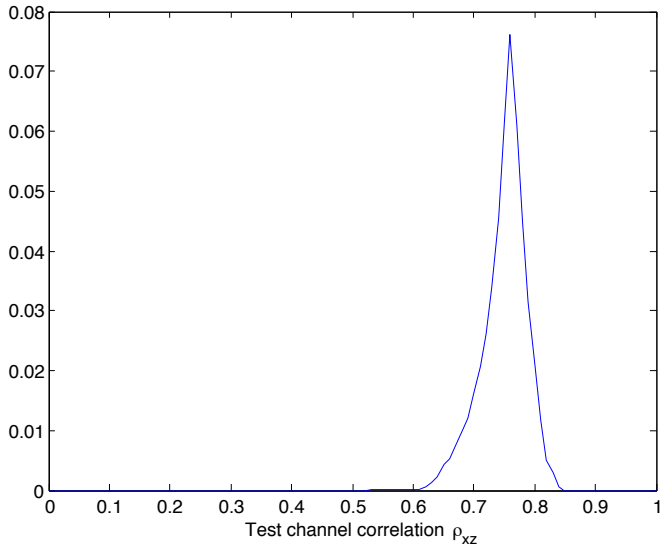

Figure 4.3: Test channel optimization for Theorem 19. The plot shows the exponent against ρ_{xz}, holding $\sigma_X^2 = 1$ fixed for $R = 0.4, \zeta_{xy} = 0.7$ and $\Delta = 0.4$.

Intuitively, ρ_{xz} controls the number of different codewords we use to cover the source sequences. At rate R the scheme allows us to identify at most $\exp(nR)$ codewords uniquely, and binning is required to go beyond this. A large codebook has the advantage that each source can be mapped to a better (i.e. closer) codeword. As we increase the size of the codebook beyond this point, the gains from having a "cleaner" codebook are outweighed by the penalty we pay for binning. From the plot we can see there is an optimum choice around $\rho_{xz} = 0.76$ for these parameters.

Figure 4.4 shows the exponent plotted (by numerically solving the optimization problem) against the rate. For comparison the upper bound of Theorem 20 is included, as is the exponent for the no side information case, corresponding a the continuous version of Marton's point-to-point exponent [58]. This result

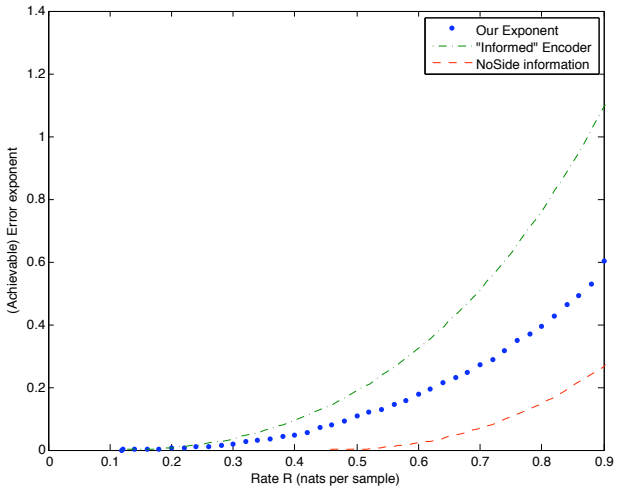

Figure 4.4: A plot of the achievable exponent of Theorem 19. Here $\zeta_{xy} = 0.7$ (the correlation coefficient between the source and side information) and $\Delta = 0.4$. $R(\Delta) = 0.121$ nats for these parameters.

was first proved by Ihara and Kubo [59], who showed the exponent is

$$\inf_{\sigma_X : \frac{1}{2}\log(\frac{\sigma_X^2}{\Delta}) > R} D(f_{\sigma_X}||f_1) = \frac{1}{2}\left(\Delta \exp(2R) - \log(\Delta \exp(2R)) - 1\right). \tag{4.22}$$

We can show our achievable exponent recovers (4.22) by taking the side information to be statistically independent i.e. $\zeta = 0$. In this case, one can show that $\rho_{xy} = \rho_{yz} = 0$ solve the inner optimization problem of (4.12). Further, since $X \perp\!\!\!\perp Y$, Y cannot help achieve the distortion constraint, choosing $\sigma_Y = 1$ is nature's best play. With these choices we see that $D(K||\bar{K}) = D(f_{\sigma_{\hat{X}}^2}||f_1)$ and we are left with the following equivalent optimization (where we have written $\hat{X} = \alpha Z$)

$$\inf_{\sigma_X} \sup_{\rho_{x\hat{x}}, \sigma_{\hat{X}}} \begin{cases} D(f_{\sigma_{\hat{X}}^2}||f_1) & \mathbb{E}[(X-\hat{X})^2] \geq \Delta \text{ or} \\ & I(X;\hat{X}) \geq R \\ \infty & \text{otherwise.} \end{cases}$$

As nature will always pick σ_X such that the supremum is finite, we are left with

$$\inf_{\sigma_X : R(\sigma_X^2, \Delta) \geq R} D(f_{\sigma_X^2} || f_1).$$

Expanding the divergence and appealing to the monotonicity of $x - \log x$ gives (4.22)[2].

Using equation (4.22) and Theorem 20 we can determine the error exponent exactly when the side information is available at both the encoder and decoder. In this case, Wyner [79, section 3] provides a simple scheme to achieve the rate distortion function. The encoder simply subtracts the conditional mean $\mathbb{E}[X|Y = y]$ from the source. An achievable exponent then follows by computing the point-to-point exponent for the random variable $X|Y = y$, which is again Gaussian, with mean $-\zeta y$ and variance $1 - \zeta^2$. Our achievable exponent in this case is

$$\inf_{\sigma_X : R(\sigma_X, \Delta) > R} D(f_{\sigma_X^2} || f_{1-\zeta^2}) = \frac{1}{2}\left(\frac{\Delta \exp(2R)}{1-\zeta^2} - \log\left(\frac{\Delta \exp(2R)}{1-\zeta^2}\right) - 1\right) \quad (4.23)$$

We now show that this is in fact the best we can do, by showing that (4.23) coincides with the upper bound of Theorem 20. The optimization problem of Theorem 20 can be solved as follows. We first note that if X, Y are zero mean with covariance matrix K, then $\mathrm{Var}(X|Y) = \frac{\det(K)}{\mathrm{Var} Y}$. Hence we may write the problem as

$$\inf_{K \succeq 0:\ g(K,\Delta,R) \leq 0} D(K||\Sigma)$$

where $g(K, \Delta, R) = -\log\det(K) + \log(\Delta) + \log e_2^T K e_2 + 2R$. The KKT conditions tell us the optimum K^* must satisfy

[2]Using a virtually identical argument one can show that exponent of Theorem 17 reduces to Marton's exponent for the discrete-memoryless case when the side information is independent of the source.

1. $-\frac{1}{2}(K^*)^{-1} + \frac{1}{2}\Sigma^{-1} + \lambda \left(-(K^*)^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & e_2^T K^* e_2 \end{bmatrix} \right) = 0$
2. $\lambda g(K^*) = 0.$

One can solve to this system to find

$$K^* = \begin{bmatrix} \zeta^2 + \Delta \exp(2R) & \zeta \\ \zeta & 1 \end{bmatrix}.$$

Evaluating $D(K^*||\Sigma)$ yields (4.23). Therefore, when the side information is available in both places we have determined the exponent exactly as (4.23).

CHAPTER 5

IMPROVED SOURCE CODING EXPONENTS VIA WITSENHAUSEN'S RATE

In this chapter we improve the results of the previous chapter, for the special case that the side information is availably fully (i.e. without being encoded) at the decoder, see Fig 5.1.

5.1 Notation and Preliminaries

For sets, types, etc., we use the same notations as the previous chapter. Unless specified, exponents and logarithms are taken in base 2. We use $\|\mathbf{x}\|_\infty$ to denote the supremum norm, i.e. $\|\mathbf{x}\|_\infty = \max_i |x_i|$. The notation $T_Q^{n,\epsilon}$ denotes the (Q, n, ϵ)-typical set, i.e. the set of $\mathbf{x} \in \mathcal{X}^n$ satisfying $\|Q_{\mathbf{x}} - Q\|_\infty \leq \epsilon$.

A graph $G = (V, E)$ is a pair of sets, where V is the set of vertices and $E \subset V \times V$ is the set of edges. Two vertices $x, y \in V$ are connected iff $(x, y) \in E$. In this chapter we need only consider simple graphs, i.e. undirected graphs without self-loops. The *degree of a vertex* v, $\Delta(v)$, is the number of other vertices

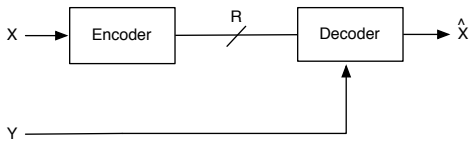

Figure 5.1: Source coding with full side information

to which v is connected. The *degree of a graph* G, denoted $\Delta(G)$ is defined as $\max_{v \in V} \Delta(v)$. A coloring of a graph is an assignment of colors to vertices so that no pair of adjacent vertices share the same color. The *chromatic number* of G, $\gamma(G)$, is defined to be the fewest number of colors needed to color G. For $U \subset V$, $G(U)$ is the *(vertex-) induced subgraph*, i.e. the graph with vertex set U and edge set $E \cap (U \times U)$. An *independent set* of G is a subgraph of G containing no edges. The graph $\bar{G}$ is the *graph complement* of G, which has the same vertex set of G and two vertices are connected in $\bar{G}$ if and only if they are not connected in G. A clique of G is a subset of the vertices of G such that every two vertices are connected. A graph G is called *perfect* if the chromatic number of every induced subgraph, $G(V')$ is equal to the size of the largest clique of $G(V')$.

Let $G = (V, E), H = (V', E')$ be two graphs. The *strong product* (also called the *and product* or *normal product*) $G \wedge H$ is a graph whose vertex set is $V \times V'$ and in which two vertices $(v, v'), (u, u')$ are connected iff

1. $v = u$ and $(v', u') \in E'$ or
2. $v' = u'$ and $(v, u) \in E$ or
3. $(v, u) \in E$ and $(v', u') \in E'$.

We will be interested in $G^n = G \wedge G \wedge \ldots \wedge G$ (n-factors), the n-fold strong product of G. One may think of the vertices of G^n as length n vectors $(v_1, \ldots, v_n)$ with two vertices are connected in G^n if each of the components of the vectors are either the same or connected in G. The *characteristic graph*, G_X, of a source P_{XY} is the graph whose vertex set is $\mathcal{X}$ and two vertices x, x' are connected if there is a $y \in \mathcal{Y}$ such that $P(y|x')P(y|x) > 0$. For a given $\mathbf{y}$, the set $Z(\mathbf{y}) = \{\mathbf{x} : P(\mathbf{x}|\mathbf{y}) > 0\}$ is the set of 'confusable' sequences, i.e. the set of $\mathbf{x}$s than can occur with a

given $\mathbf{y}$.

For a graph G and distribution Q on the vertices of G, we define the following functional.

Definition 9.

$$\kappa(G, Q) = \max_{\substack{W: W \ll G \\ QW = Q}} H(W|Q). \tag{5.1}$$

Note: whenever we write the graph G where a matrix is expected, we abuse notation and refer to the matrix $G = A + I$ where A is the adjacency matrix of graph G and I is the identity matrix.

A second equivalent definition of κ is

$$\kappa(G, Q) = \max_{X, \tilde{X}: Q_X = Q_{\tilde{X}} = Q} H(\tilde{X}|X) \tag{5.2}$$

where X and $\tilde{X}$ have common alphabet and $P(\tilde{x}|x) > 0$ only if $(\tilde{x}, x) \in E(G)$ or $x = \tilde{x}$.

We remark that similar optimizations arise in the determination of maximum entropy Markov chains subject to moment constraints [80].

We will also make use of the following graph functionals from graph/zero-error information theory.

Definition 10. The graph entropy [64], $H(G, Q)$, of a graph G and a distribution Q on the vertices of G is defined as

$$H(G, Q) = \min_{X \in Z \in \Gamma(G)} I(X; Z) \tag{5.3}$$

where X is a random node in the graph and has distribution Q, $\Gamma(G)$ denotes the set of all maximal independent sets of G, and the notation $X \in Z$ means $P_{Z|X}(z|x) = 0$ for $x \notin z$.

Definition 11. The complementary graph entropy (or co-entropy or π-entropy) [62, 63], $\bar{H}(G,Q)$ of a graph G with a distribution Q on the vertices of G is defined as

$$\bar{H}(G,Q) = \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{\log \gamma(G_X^n(T_Q^{n,\epsilon}))}{n}. \tag{5.4}$$

Graph entropy and complementary graph entropy are related as follows (see for example [81, Theorem 4])

$$\bar{H}(G,Q) \leq H(G,Q), \tag{5.5}$$

and equality holds in (5.5) if G is perfect [82, Corollary 12].

5.2 Properties of κ

In this section we give some properties of κ which will be used elsewhere in the chapter. Throughout this section G is a graph, Q is a distribution on the vertices of G and X is a random variable with distribution Q.

Property 1. $\kappa(G,Q) \leq H(Q) = H(X)$, where equality holds if G is fully connected.

Proof. Note that any valid choice of channel in the optimization defining $\kappa(G,Q)$ satisfies $QW = Q$, thus $H(W|Q) \leq H(Q)$, giving the first claim.

If G is fully connected then the constraint $W \ll G$ imposes no restriction on the choice of W. The problem is then to choose a W that produces the given output distribution Q. Setting the rows of W equal to Q gives $\kappa(G,Q) = H(Q)$.

□

Property 2. If G is the disjoint union of fully connected subgraphs then

$$\kappa(G, Q) = H(X|Y). \tag{5.6}$$

where

1. Y is a random variable with alphabet size $|\mathcal{Y}|$ equal to the number of disjoint subgraphs in G so that to each subgraph we associate a unique element $y \in \mathcal{Y}$; and
2. for the subgraph associated with y, the event $\{X = a, Y = y\}$ has probability $Q(a)$ if a is in the subgraph and probability zero otherwise.

Proof. Via (5.2), it follows that

$$\kappa(G, Q) = H(X) - \min_{X, \tilde{X}: Q_X = Q_{\tilde{X}} = Q} I(X; \tilde{X})$$

where X and $\tilde{X}$ have common alphabet and $P(\tilde{x}|x) > 0$ only if $(\tilde{x}, x) \in E(G)$ or $\tilde{x} = x$. Now, because Y can be determined from $\tilde{X}$ (due to the graph-based constraint on $P_{\tilde{X}|X}$), we have $I(X; \tilde{X}) = I(X; \tilde{X}, Y)$. By the chain rule,

$$I(X; \tilde{X}, Y) = I(X; Y) + I(X; \tilde{X}|Y),$$

which gives

$$\kappa(G, Q) = H(X|Y) - \min_{X, \tilde{X}: Q_X = Q_{\tilde{X}} = Q} I(X; \tilde{X}|Y).$$

Choosing $P_{\tilde{X}|X}$ so that $X \perp\!\!\!\perp \tilde{X}|Y$, i.e. setting $P_{\tilde{X}|X}(\tilde{x}|x) = Q(\{x' : x' = x \text{ or } (x, x') \in E(G)\})^{-1} Q(x)$ if $(x, \tilde{x}) \in E(G)$ or $x = \tilde{x}$, and $P_{\tilde{X}|X}(\tilde{x}|x) = 0$ otherwise shows the minimum is zero. □

Property 3. Let G be a graph and $Q^{(n)}$ be a sequence of distributions (on the vertices of G) converging to distribution Q^{∞}. Then

$$\limsup_{n \to \infty} \kappa(G, Q^{(n)}) \leq \kappa(G, Q^{\infty})$$

(I.e. $\kappa(G, \cdot)$ is upper semicontinuous in Q for a fixed G.)

Proof. Let

$$W^{(n)} = \arg \max_{\substack{W: W \ll G \\ Q^{(n)}W = Q^{(n)}}} H(W|Q^{(n)}),$$

where $W^{(n)}$ exists because we are maximizing a continuous function over a compact set. By choosing a subsequence and relabeling we may arrange it so that $H(W^{(n)}|Q^{(n)}) \to \limsup H(W^{(n)}|Q^{(n)})$ and $W^{(n)} \to W^{\infty}$, where both $W^{\infty} \ll G$ and $Q^{\infty}W^{\infty} = Q^{\infty}$ are true. Therefore, we have

$$\begin{aligned} \limsup_{n\to\infty} \kappa(G, Q^{(n)}) &= \limsup_{n\to\infty} H(W^{(n)}|Q^{(n)}) \\ &= H(W^{\infty}|Q^{\infty}) \leq \kappa(G, Q^{\infty}). \end{aligned}$$

□

5.3 Bounding Witsenhausen's Rate

Recall that in Witsenhausen's problem [60] the goal is communication of X^n to the decoder who has access to Y^n under the criterion $P_{XY}^n(X^n = \hat{X}^n) = 1$. This requirement is stricter than the vanishing error probability criterion of Slepian-Wolf and increases the required rate from $H(X|Y)$ to $R(G_X)$. As mentioned in Chapter 1, an alternative characterization of Witsenhausen's rate is via complementary graph entropy.

Lemma 22.

$$R(G_X) = \max_{Q_X \in \mathcal{P}(\mathcal{X})} \bar{H}(G_X, Q_X). \tag{5.7}$$

Proof. This identity is easily established using the fact that there are only polynomially many types. See [63, Lemma 3] for a recent proof of a more general result. □

Our new bound on $R(G_X)$ stems from a new bound on $\bar{H}(G_X, Q_X)$. We derive this bound in two steps. First we provide a degree bound on the type-induced subgraph $G_X^n(T_Q^n)$. We then we pass from a degree bound to a chromatic number bound.

Lemma 23. *Let $Q_X \in \mathcal{P}^n(\mathcal{X})$. Then*

$$\begin{aligned} &(n+1)^{-|\mathcal{X}||\mathcal{X}|} \exp(n\kappa_n(G_X, Q_X)) - 1 \\ &\leq \Delta(G_X^n(T_{Q_X}^n)) \\ &\leq (n+1)^{|\mathcal{X}||\mathcal{X}|} \exp(n\kappa_n(G_X, Q_X)) \end{aligned} \tag{5.8}$$

where

$$\kappa_n(G_X, Q_X) = \max_{\substack{W \in \mathcal{C}^n(Q_X, \mathcal{X}) \\ W \ll G_X \\ Q_X W = Q_X}} H(W|Q_X). \tag{5.9}$$

Note: κ_n maximizes over channels giving rise to types rather than distributions, but of course we may replace κ_n by κ in the right-hand inequality of (5.8) *to get another valid upper bound.*

Proof. Suppose $\mathbf{x} \in T_{Q_X}^n$, and let $N(\mathbf{x})$ denote the neighbors of $\mathbf{x}$ in the induced subgraph $G_X^n(T_{Q_X}^n)$. We partition the set $\{(\mathbf{x}, \mathbf{x}') : \mathbf{x}' \in N(\mathbf{x})\}$ by joint type $Q_{XX'}$ and observe that each joint type can be written as $Q_X \times W$ for some W. One may verify that $W \ll G_X$ and $Q_X W = Q_X$.

For any $\mathbf{x} \in T_{Q_X}^n$ we can count the number of sequences in $N(\mathbf{x})$ by decomposing $\{(\mathbf{x}, \mathbf{x}') : \mathbf{x}' \in N(\mathbf{x})\}$ into joint types, determining a W for each joint type

and using the standard cardinality bounds for type classes. Thus

$$\begin{aligned}\Delta(G_X^n(T_{Q_X}^n)) &\leq \sum_{\substack{W:W\ll G\\ Q_XW=Q_X}} |T_W^n(\mathbf{x})| \\ &\leq \sum_{\substack{W:W\ll G\\ Q_XW=Q_X}} \exp(nH(W|Q_X)) \\ &\leq (n+1)^{|\mathcal{X}||\mathcal{X}|} \max_{\substack{W:W\ll G\\ Q_XW=Q_X}} \exp(nH(W|Q_X)).\end{aligned}$$

For the reverse inequality, we let $\Delta(\mathbf{x})$ denote the degree of vertex $\mathbf{x}$ in the induced subgraph. Then

$$\Delta(\mathbf{x}) = \sum_{\substack{W\in\mathcal{C}^n(Q_X,\mathcal{X})\\ W\neq I, W\ll G\\ Q_XW=Q_X}} |T_W^n(\mathbf{x})|.$$

To see this, note first that if W arises by selecting a $\mathbf{x}' \in N(\mathbf{x})$, then $T_W(\mathbf{x}) \subset N(\mathbf{x})$. This is the case because $N(\mathbf{x})$ is simply a union of W-shells. And second, that any $W \neq I$ with $W \ll G$ and $Q_XW = Q_X$ gives rise to a neighbor. Then because $\Delta(G_X^n(T_{Q_X}^n)) = \max_{\mathbf{x}\in T_{Q_X}} \Delta(\mathbf{x})$, we have

$$\Delta(G_X^n(T_{Q_X}^n)) = \max_{\mathbf{x}\in T_{Q_X}} \sum_{\substack{W\in\mathcal{C}^n(Q_X,\mathcal{X})\\ W\neq I, W\ll G\\ Q_XW=Q_X}} |T_W^n(\mathbf{x})|$$

$$\Delta(G_X^n(T_{Q_X}^n)) = \max_{\mathbf{x}\in T_{Q_X}} \sum_{\substack{W\in\mathcal{C}^n(Q_X,\mathcal{X})\\ V\ll G\\ Q_XV=Q_X}} |T_V^n(\mathbf{x})| - 1.$$

Using the cardinality bound for typeclasses we get

$$\begin{aligned}
&\Delta(G_X^n(T_{Q_X}^n)) \\
&\geq \max_{\mathbf{x}\in T_{Q_X}} \max_{\substack{W:W\ll G\\ Q_X W=Q_X}} |T_W^n(\mathbf{x})| - 1 \\
&\geq \max_{\mathbf{x}\in T_{Q_X}} (n+1)^{-|\mathcal{X}||\mathcal{X}|} \max_{\substack{W:W\ll G\\ Q_X W=Q_X}} \exp(n(H(W|Q_X))) - 1 \\
&= (n+1)^{-|\mathcal{X}||\mathcal{X}|} \max_{\substack{W:W\ll G\\ Q_X W=Q_X}} \exp(n(H(W|Q_X))) - 1
\end{aligned}$$

where we implicitly assumed we still have $W \in \mathcal{C}^n(Q_X, \mathcal{X})$. □

Using this result we now bound $\bar{H}(G_X, Q_X)$.

Lemma 24. *For any $Q_X \in \mathcal{P}(\mathcal{X})$ and graph G_X with vertex set $\mathcal{X}$*

$$\bar{H}(G_X, Q_X) \leq \kappa(G_X, Q_X).$$

Proof. We may upper bound $\bar{H}(G_X, Q_X)$ by coloring each typeclass separately. This gives

$$\begin{aligned}
&\bar{H}(G_X, Q_X) \\
&= \lim_{\epsilon\to 0} \limsup_{n\to\infty} \frac{\log \gamma(G_X^n(T_{Q_X}^{n,\epsilon}))}{n} \\
&\leq \lim_{\epsilon\to 0} \limsup_{n\to\infty} \frac{\log\left((n+1)^{|\mathcal{X}|} \max_{Q\in\mathcal{P}^n(\mathcal{X}):\|Q_X - Q\|_\infty\leq\epsilon} \gamma(G_X^n(T_Q^n))\right)}{n} \\
&= \lim_{\epsilon\to 0} \limsup_{n\to\infty} \frac{\log\left(\max_{Q\in\mathcal{P}^n(\mathcal{X}):\|Q_X - Q\|_\infty\leq\epsilon} \gamma(G_X^n(T_Q^n))\right)}{n}.
\end{aligned}$$

A well-known fact from graph theory tells us that $\gamma(G) \leq \Delta(G) + 1$ [83, §5.2].

Using this and Lemma 23 we obtain

$$\begin{aligned}
&\bar{H}(G_X, Q_X) \\
&\leq \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{\log\left(1 + \max_{Q \in \mathcal{P}^n(\mathcal{X}): \|Q_X - Q\|_\infty \leq \epsilon} \Delta(G_X^n(T_Q^n))\right)}{n} \\
&\leq \lim_{\epsilon \to 0} \limsup_{n \to \infty} \frac{1}{n} \log\left(1 + (n+1)^{|\mathcal{X}|^2} \max_{Q \in \mathcal{P}^n(\mathcal{X}): \|Q_X - Q\|_\infty \leq \epsilon} \exp(n\kappa(G_X, Q))\right).
\end{aligned}$$

We get a further upper bound by replacing $\mathcal{P}^n(\mathcal{X})$ in the maximization by $\mathcal{P}(\mathcal{X})$, which, after taking the $\limsup$, gives

$$\bar{H}(G_X, Q_X) \leq \lim_{\epsilon \to 0} \sup_{Q \in \mathcal{P}(\mathcal{X}): \|Q_X - Q\|_\infty \leq \epsilon} \kappa(G_X, Q).$$

Because κ is upper semicontinuous, it attains its supremum and therefore we can consider any sequence $\epsilon_m \downarrow 0$, and let $Q^{(m)}$ be the corresponding maximizer. Observe that $\limsup_{m \to \infty} \|Q^{(m)} - Q_X\|_\infty \leq 0$, i.e. $Q^{(m)} \to Q_X$. Thus, by κ property 3 we have

$$\bar{H}(G_X, Q_X) \leq \limsup_{m \to \infty} \kappa(G_X, Q^{(m)}) \leq \kappa(G_X, Q_X).$$

□

We can now state our bound on Witsenhausen's rate.

Theorem 21.

$$R(G_X) \leq \max_{Q_X \in \mathcal{P}(\mathcal{X})} \kappa(G_X, Q_X).$$

Proof. An immediate consequence of Lemma 24 and (5.7). □

5.3.1 Comparison of Bounds

Since both $\kappa(G_X, Q_X)$ and $H(G_X, Q_X)$ provide upper bounds on $\bar{H}(G_X, Q_X)$, and therefore on Witsenhausen's rate, it is natural to ask which of these two

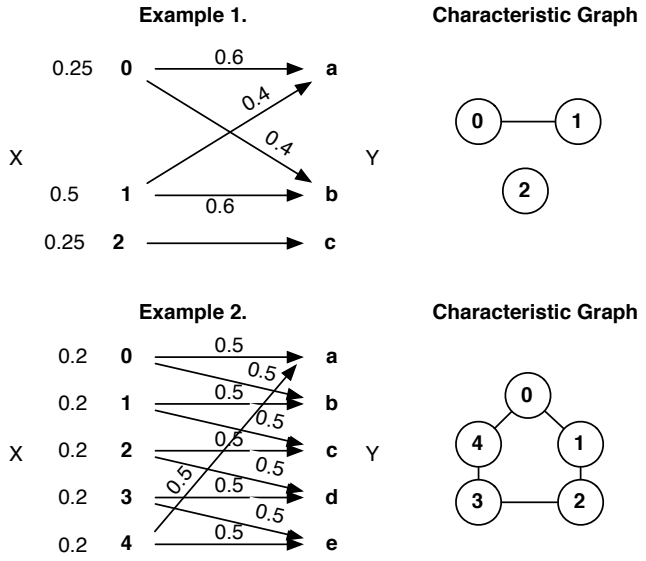

Figure 5.2: Example 1 and 2: Two source distributions and their characteristic graphs

bounds is best. It is currently unknown whether one bound dominates the other. For all of the examples we studied, $\kappa(G_X, Q_X) \geq H(G_X, Q_X)$. However, $\kappa(G_X, Q_X)$ does have one important advantage in that can be computed efficiently by convex programming techniques whereas computation of $H(G_X, Q_X)$ involves a maximization over all distributions on the potentially exponentially many (in $|\mathcal{X}|$) independent sets [84]. Additionally even computation of the simple bound $R(G_X) \leq \gamma(G_X)$ requires finding the chromatic number of G_X which is NP-complete. Therefore using κ could be beneficial when dealing with large graphs.

We now provide three concrete examples to compare the bounds.

Example 1 (see Fig. 5.2)**.** For this example, the graph G_X is perfect and for perfect graphs it is known that $\bar{H}(G_X, Q_X) = H(Q_X) - H(\bar{G}_X, Q_X)$ [82, Theorem

2]. To compute

$$\begin{aligned}\bar{H}(G_X, Q_X) &= H(Q_X) - H(\bar{G}_X, Q_X) \\ &= H(Q_X) - \min_{X \in Z \in \Gamma(\bar{G}_X)} I(X;Z) \\ &= \max_{X \in Z \in \Gamma(\bar{G}_X)} H(X|Z)\end{aligned}$$

it suffices to notice the set $\Gamma(\bar{G}_X) = \{\{0,1\},\{2\}\}$ forces $P_{Z|X}$ to be deterministic. Therefore, using h_2 to denote the binary entropy function, we have

$$\begin{aligned}\bar{H}(G_X, Q_X) &= P(Z = \{0,1\}) h_2\Big(\frac{Q_X(0)}{Q_X(0)+Q_X(1)}\Big) \\ &\quad + P(Z = \{2\}) h_2(0) \\ &= [Q_X(0)+Q_X(1)] h_2\Big(\frac{Q_X(0)}{Q_X(0)+Q_X(1)}\Big).\end{aligned}$$

For $\kappa(G_X, Q_X)$ we use the fact that G_X is the disjoint union of fully connected subgraphs, and therefore κ property 2 gives

$$\kappa(G_X, Q_X) = [Q_X(0)+Q_X(1)] h_2\Big(\frac{Q_X(0)}{Q_X(0)+Q_X(1)}\Big).$$

Since the graph is perfect, recall that we have equality in (5.5), and therefore in this example

$$\kappa(G_X, Q_X) = H(G_X, Q_X) = \bar{H}(G_X, Q_X).$$

Example 2 (see Fig. 5.2)**.** Here the characteristic graph is the well-known pentagon graph, C_5. For this case, with Q_X denoting the uniform input distribution as depicted, both $\bar{H}(G_X, Q_X)$ and $H(G_X, Q_X)$ are known (e.g.[85, Example 1, pp. 105]),

$$\bar{H}(G_X, Q_X) = \frac{1}{2}\log 5 \text{ and } H(G_X, Q_X) = \log\frac{5}{2}.$$

To compute $\kappa(G_X, Q_X)$, we note the constraint $W \ll G_X$ implies that $H(W|Q_X) \leq \log 3$. But $\kappa(G_X, Q_X) = \log 3$ is achievable by setting $W(\tilde{x}|x) = 1/3$ whenever W can be non-zero. Therefore we have the strict inequalities

$$\bar{H}(G_X, Q_X) < H(G_X, Q_X) < \kappa(G_X, Q_X).$$

The final example shows that in certain cases the bound provided by Theorem 21 can be arbitrarily bad.

Example 3 (Looseness of the bound in Theorem 21)**.** Consider the graph G with $V(G) = \{0, 1, \ldots, 2^n\}$ and $E(G) = \{(n, n+1) : n \geq 0\} \cup \{(0, n) : n \geq 2\}$. It is clear that $\gamma(G) = 3$ for all n, and hence $R(G) \leq \log 3$. Yet, if we choose

$$W(b|0) = \begin{cases} 0 & \text{if } b = 0 \\ 2^{-n} & \text{otherwise} \end{cases}$$

$$W(b|a \neq 0) = \begin{cases} 1 & \text{if } b = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$Q = \left[\frac{1}{2}, \frac{1}{2^{n+1}}, \frac{1}{2^{n+1}} \cdots \frac{1}{2^{n+1}}\right]$$

one sees that $W \ll G$ and therefore that

$$\kappa(G, Q) \geq H(W|Q) = \frac{1}{2} \log 2^n = \frac{n}{2}.$$

5.4 Improved Exponents for Lossless Source Coding

We consider the setup depicted in Figure 5.1 under the vanishing error probability criterion. The encoder/decoder pair are functions $\psi : \mathcal{X}^n \to \mathcal{M}$ and $\varphi : \mathcal{M} \times \mathcal{Y}^n \to \hat{\mathcal{X}}^n$, where $\mathcal{M}$ is a fixed set. Define the error probability to be

$$P_e(\psi, \varphi) = \Pr(X^n \neq \hat{X}^n) \tag{5.10}$$

where $\hat{X}^n = \varphi(\psi(X^n), Y^n)$. Our focus is the asymptotic behavior of the error probability $P_e(\psi, \varphi)$ as n gets large. Define the *error exponent* (or *reliability function*) to be

$$\theta(R, P_{XY}) = \lim_{\epsilon \to 0} \liminf_{n \to \infty} -\frac{1}{n} \log \left[\min_{(\psi, \varphi)} P_e(\psi, \varphi)\right] \tag{5.11}$$

where the minimization ranges over all encoder/decoder pairs satisfying

$$\frac{1}{n}\log|\mathcal{M}| \leq R + \epsilon. \tag{5.12}$$

To state our result we need the following definitions. For any distribution $Q \in \mathcal{P}(\mathcal{X})$, channels $W, \tilde{W} \in \mathcal{C}(\mathcal{X} \to \mathcal{Y})$ and rate R define

$$\begin{aligned}\mathcal{Q}(W, \tilde{W}, Q, R) = \{Q_{X\tilde{X}Y} \in \mathcal{P}(\mathcal{X} \times \mathcal{X} \times \mathcal{Y}) : Q_{Y|X} = W,\\ Q_{Y|\tilde{X}} = \tilde{W}, Q_X = Q_{\tilde{X}} = Q,\\ I_{Q_{X\tilde{X}}}(X; \tilde{X}) \leq R\}.\end{aligned}$$

Let $\alpha(Q, W)$ be a real valued function and use the notation $\tilde{W} \leq_\alpha W$ to denote

$$\alpha(Q, \tilde{W}) \leq \alpha(Q, W) \text{ and } QW = \tilde{Q}W$$

and let

$$\mathcal{W}(\alpha, Q) = \{W \in \mathcal{C}(\mathcal{X} \to \mathcal{Y}), \tilde{W} \in \mathcal{C}(\mathcal{X} \to \mathcal{Y}) : \tilde{W} \leq_\alpha W\}.$$

Finally, define

$$\begin{aligned}&e(\alpha, Q, P_{XY}, R) =\\ &\min_{W, \tilde{W} \in \mathcal{W}(\alpha, Q)} \Big[D(W||P_{Y|X}|Q)\\ &\quad + \min_{Q_{X\tilde{X}Y} \in \mathcal{Q}(W, \tilde{W}, Q, R)} [I_{Q_{X\tilde{X}Y}}(X, Y; \tilde{X}) - R]^+\Big],\\ &e_r(Q, P_{XY}, R) =\\ &\min_{W \in \mathcal{C}(\mathcal{X} \to \mathcal{Y})} \Big[D(W||P_{Y|X}|Q) + [I(Q; W) - R]^+\Big],\end{aligned}$$

$$\begin{aligned}\text{and } &e_x(Q, P_{XY}, R) =\\ &\min_{\substack{Q_{X\tilde{X}} : Q_X = Q_{\tilde{X}} = Q\\ I(X;\tilde{X}) \leq R}} \Big[\mathbb{E}d_B(X, \tilde{X}) + I(X; \tilde{X}) - R\Big],\end{aligned}$$

where d_B is the Bhattacharyya distance

$$d_B(x, \tilde{x}) = -\log \sum_{y \in \mathcal{Y}} \sqrt{P_{Y|X}(y|x) P_{Y|X}(y|\tilde{x})}.$$

In [53], Csiszár and Körner construct source codes for the present problem by exhibiting schemes based on certain partitions $\{\mathcal{A}_i\}_{i=1}^q$ of $\mathcal{X}^n$. Let $\mathcal{M} = \{1, \ldots, q\}$ and for a given partition $\{\mathcal{A}_i\}_{i=1}^q$, define the encoder $\psi : \mathcal{X}^n \to \mathcal{M}$

$$\psi(\mathbf{x}) = i, \text{ if } \mathbf{x} \in \mathcal{A}_i.$$

A family of decoders $\{\varphi_i : \mathcal{Y}^n \to \mathcal{A}_i\}_{i=1}^q$ is specified via an α-decoding rule, which for a given $i \in \mathcal{M}$ and $\mathbf{y} \in \mathcal{Y}^n$, declares the output as $\mathbf{x} \in \mathcal{A}_i$ satisfying $\alpha(Q_{\mathbf{x}}, Q_{\mathbf{y}|\mathbf{x}}) \leq \alpha(Q_{\tilde{\mathbf{x}}}, Q_{\mathbf{y}|\tilde{\mathbf{x}}})$ for all $\tilde{x} \in \mathcal{A}_i$, with ties broken arbitrarily. The decoder is then specified by $\varphi(i, \mathbf{y}) = \varphi_i(\mathbf{y})$.

For this communication scheme, which we will call the CK scheme, the following result holds.

Theorem 22 (Csiszár and Körner [53, Theorem 2])**.** *For every $R \geq 0$, there exists a partition $\{\mathcal{A}_i\}_{i=1}^q$ of $\mathcal{X}^n$ with*

$$\frac{1}{n} \log q \leq R + \delta_n, \tag{5.13}$$

such that for every distribution $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ the encoder, ψ_{CK}, defined by the partition $\{\mathcal{A}_i\}_{i=1}^q$ and decoder, $\varphi_{CK,\alpha}$, defined via the α-decoder, has probability of error

$$P_e(\psi_{CK}, \varphi_{CK,\alpha}) \leq \exp\Bigg(- n \Big[\min_{Q_X \in \mathcal{P}^n(\mathcal{X})} D(Q_X || P_X) + e(\alpha, Q_X, P_{XY}, H(Q_X) - R) - \delta_n'' \Big] \Bigg) \tag{5.14}$$

where

$$\delta_n = \Big(|\mathcal{X}|^2 + |\mathcal{X}| \Big) \frac{\log(n+1)}{n} + \frac{1}{n} \tag{5.15}$$

and

$$\delta_n'' = \Big(|\mathcal{X}|^2|\mathcal{Y}| + |\mathcal{X}|\Big)\frac{\log(n+1)}{n}.$$

When specialized to particular α-decoders, namely the maximum likelihood decoder

$$\alpha_{ML}(Q, W) = D(W||P_{Y|X}|Q) + H(W|Q) \tag{5.16}$$

and the minimum entropy decoder

$$\alpha_{ME}(Q, W) = H(W|Q) \tag{5.17}$$

the following result provides two alternative bounds on the decoding error probability.

Lemma 25 (Csiszár and Körner [53, Lemma 4])**.** *Under the maximum likelihood decoder* (5.16)*,*

$$e(\alpha_{ML}, Q, P_{XY}, R) \geq \max(e_x(Q, P_{XY}, R), e_r(Q, P_{XY}, R)).$$

Under the minimum entropy decoder decoder (5.17)*,*

$$e(\alpha_{ME}, Q, P_{XY}, R) \geq e_r(Q, P_{XY}, R).$$

5.4.1 An Improved Scheme

Rather than encoding every sequence using the CK scheme, we propose to encode certain typeclasses using the Witsenhausen scheme whenever the rate allows it. The precise details are as follows.

The encoder and decoder agree on a coloring of every typeclass $T^n_{Q_X}$, such a coloring requires $\gamma(G^n_X(T^n_{Q_X}))$ colors.

Encoder: To communicate, the encoder first sends the type $Q_{\mathbf{x}}$ of the sequence $\mathbf{x}$, and if there is sufficient rate, i.e. $\log \gamma(G_X^n(T_{Q_{\mathbf{x}}}^n)) < nR$, sends the color of the sequence in the graph. Otherwise the encoder sends the index of the partition containing the sequence in the CK scheme.

Decoder: Using the type, the decoder knows whether the encoder is sending the color of the sequence or the partition index. In the former case it can decoder the sequence $\mathbf{x}$ without error, otherwise it uses the CK scheme to decode.

Using this scheme we will show:

Theorem 23. *For any $R > 0$, $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, and continuous $\alpha : \mathcal{P}(\mathcal{X}) \times \mathcal{C}(\mathcal{X} \to \mathcal{Y}) \to \mathbb{R}$,*

$$\theta(R, P_{XY}) \geq \min_{\substack{Q_X: \\ \min(\kappa(G_X,Q_X),H(G_X,Q_X))\geq R}} \Big[D(Q_X||P_X) + e(\alpha, Q_X, P_{XY}, H(Q_X) - R)\Big] \tag{5.18}$$

where G_X is the characteristic graph of the source P_{XY}.

Remark: One may replace the function $\min(\kappa(G_X, Q_X), H(G_X, Q_X)) \geq R$ restricting the feasible set in (5.18) by $\min(\kappa(G_X, Q_X), H(G_X, Q_X), f(G_X, Q_X)) \geq R$ where f is any upper semicontinuous (in Q_X) function satisfying $\bar{H}(G_X, Q_X) \leq f(G_X, Q_X)$.

Corollary 3. *Using maximum likelihood decoding (cf.* (5.16)*)*

$$\theta(R, P_{XY}) \geq \min_{\substack{Q_X: \\ \min(\kappa(G_X,Q_X),H(G_X,Q_X))\geq R}} \Big[D(Q_X||P_X) + \max(e_r(Q_X, P_{XY}, H(Q_X) - R), e_x(Q_X, P_{XY}, H(Q_X) - R))\Big]. \tag{5.19}$$

Using minimum entropy decoding (cf. (5.17)*)*

$$\theta(R, P_{XY}) \geq \min_{\substack{Q_X: \\ \min(\kappa(G_X,Q_X),H(G_X,Q_X))\geq R}} \Big[D(Q_X||P_X) + e_r(Q_X, P_{XY}, H(Q_X) - R)\Big].$$

Proof. The minimum entropy decoding rule (5.17) is clearly continuous. The ML decoding rule (5.16) is continuous provided that $W \ll P_{Y|X}$, but such a choice is guaranteed by the fact that $D(W||P_{Y|X}|Q)$ appears in the objective function defining $e(\alpha, Q, P_{XY}, R)$. The results then follows from Lemma 25 and Theorem 23. □

Before we can prove Theorem 23, we need to establish the following results.

Proposition 1. *For any distribution $Q \in \mathcal{P}(\mathcal{X})$, channels $W, \tilde{W} \in \mathcal{C}(\mathcal{X} \to \mathcal{Y})$ and rate R, the set $\mathcal{Q}(W, \tilde{W}, Q, R)$ is compact.*

Proof. The set is clearly bounded. To show that it is closed, let $Q^{(n)}_{X\tilde{X}Y} \to Q_{X\tilde{X}Y}$ and observe that for each n, $Q^{(n)}_{Y|X} = W, Q^{(n)}_{Y|\tilde{X}} = \tilde{W}, Q^{(n)}_X = Q^{(n)}_{\tilde{X}} = Q, I_{Q^{(n)}_{X\tilde{X}}}(X; \tilde{X}) \leq R$. Taking limits and using the continuity of mutual information gives the result. □

Proposition 2. *For any distribution $Q \in \mathcal{P}(\mathcal{X})$ and continuous $\alpha : \mathcal{P}(\mathcal{X}) \times \mathcal{C}(\mathcal{X} \to \mathcal{Y}) \to \mathbb{R}$, the set $\mathcal{W}(\alpha, Q)$ is compact.*

Proof. The set is clearly bounded. To see that it is closed, suppose $\{W^{(n)}\}$ and $\{\tilde{W}^{(n)}\}$ are two sequences in $\mathcal{W}(\alpha, Q)$. Suppose further that $W^{(n)} \to W$ and $\tilde{W}^{(n)} \to \tilde{W}$. Since for each $x \in \mathcal{X}$ and n we have

$$\sum_y Q(y)W^{(n)}(x|y) = \sum_y Q(y)\tilde{W}^{(n)}(x|y)$$

taking limits and then interchanging the order of limits and sums shows that $QW = Q\tilde{W}$. By continuity of α we also have that

$$\alpha(Q, W) = \lim_{n\to\infty} \alpha(Q, W^{(n)}) \leq \lim_{n\to\infty} \alpha(Q, \tilde{W}^{(n)}) = \alpha(Q, \tilde{W}).$$

□

Proposition 3. *Let $\{W^{(n)}\}, \{\tilde{W}^{(n)}\}$ be sequences of channels with $W^{(n)} \to W$ and $\tilde{W}^{(n)} \to \tilde{W}$. Then for any $Q \in \mathcal{P}(\mathcal{X})$, $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and R,*

$$\begin{aligned}
&\liminf_{n\to\infty} \left[D(W^{(n)}||P_{Y|X}|Q) + \min_{Q_{X\tilde{X}Y} \in \mathcal{Q}(W^{(n)}, \tilde{W}^{(n)}, Q, R)} [I_{Q_{X\tilde{X}Y}}(X, Y; \tilde{X}) - R]^+ \right] \\
&\geq D(W||P_{Y|X}|Q) + \min_{Q_{X\tilde{X}Y} \in \mathcal{Q}(W, \tilde{W}, Q, R)} [I_{Q_{X\tilde{X}Y}}(X, Y; \tilde{X}) - R]^+.
\end{aligned}$$

Proof. Let $Q^{(n)}_{X\tilde{X}Y} = Q^{(n)}_{X\tilde{X}Y}(W^{(n)}, \tilde{W}^{(n)}, R, Q)$ be a minimizer of

$$\min_{Q_{X\tilde{X}Y} \in \mathcal{Q}(W^{(n)}, \tilde{W}^{(n)}, Q, R)} [I_{Q_{X\tilde{X}Y}}(X, Y; \tilde{X}) - R]^+$$

if the set $\mathcal{Q}(W^{(n)}, \tilde{W}^{(n)}, Q, R)$ is non-empty and arbitrary otherwise. If the sets $\mathcal{Q}(W^{(n)}, \tilde{W}^{(n)}, Q, R)$ are empty for all n sufficiently large, then the result trivially holds. Therefore we can focus on the case for which there is a subsequence along which the sets are non-empty. Along this subsequence there is a further subsequence where $Q^{(n)}_{X\tilde{X}Y}$ converges, so that by relabeling we may assume that $Q^{(n)}_{X\tilde{X}Y} \to Q_{X\tilde{X}Y}$ and $\mathcal{Q}(W^{(n)}, \tilde{W}^{(n)}, Q, R)$ is non-empty for each n. Now by lower semicontinuity of the information measures it follows that

$$\begin{aligned}
&\liminf_{n\to\infty} \left[D(W^{(n)}||P_{Y|X}|Q) + \min_{Q_{X\tilde{X}Y} \in \mathcal{Q}(W^{(n)}, \tilde{W}^{(n)}, Q, R)} [I_{Q_{X\tilde{X}Y}}(X, Y; \tilde{X}) - R]^+ \right] \\
=&\liminf_{n\to\infty} D(W^{(n)}||P_{Y|X}|Q) + [I_{Q^{(n)}_{X\tilde{X}Y}}(X, Y; \tilde{X}) - R]^+ \\
&\geq D(W||P_{Y|X}|Q) + [I_{Q_{X\tilde{X}Y}}(X, Y; \tilde{X}) - R]^+.
\end{aligned}$$

The convergence of $W^{(n)}$ and $\tilde{W}^{(n)}$ and the continuity of mutual information imply that $Q_{X\tilde{X}Y} \in \mathcal{Q}(W, \tilde{W}, Q, R)$, which gives the result. □

Lemma 26. *Let $\{Q^{(n)}\}$, $Q^{(n)} \in \mathcal{P}(\mathcal{X})$ for each n, be a sequence of distributions converging to $Q \in \mathcal{P}(\mathcal{X})$ and $\alpha : \mathcal{P}(\mathcal{X}) \times \mathcal{C}(\mathcal{X} \to \mathcal{Y}) \to \mathbb{R}$ be continuous. Then for any R and $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$*

$$\liminf_{n\to\infty} e(\alpha, Q^{(n)}, P_{XY}, H(Q^{(n)}) - R) \geq e(\alpha, Q, P_{XY}, H(Q) - R).$$

Proof. If the sequence $Q^{(n)}$ is such that $e(\alpha, Q^{(n)}, P_{XY}, H(Q^{(n)}) - R) = \infty$ for all n sufficiently large, then the result trivially holds. Therefore it remains to handle the case for which there is a subsequence along which $e(\alpha, Q^{(n)}, P_{XY}, H(Q^{(n)}) - R)$ is finite for all n. Let us relabel so that $Q^{(n)}$ gives this property.

For any $W, \tilde{W}$, let $Q^{(n)}_{X\tilde{X}Y} = Q^{(n)}_{X\tilde{X}Y}(W, \tilde{W}, R, Q^{(n)})$ be a minimizer of

$$\min_{Q_{X\tilde{X}Y} \in \mathcal{Q}(W,\tilde{W},Q^{(n)},H(Q^{(n)})-R)} [I_{Q_{X\tilde{X}Y}}(X, Y; \tilde{X}) - H(Q^{(n)}) + R]^+$$

if the set $\mathcal{Q}(W, \tilde{W}, Q^{(n)}, H(Q^{(n)}) - R)$ is non-empty and arbitrary otherwise.

Similarly, let $W^{(n)} = W^{(n)}(R, Q^{(n)}), \tilde{W}^{(n)} = \tilde{W}^{(n)}(R, Q^{(n)})$ be minimizers of

$$\min_{W,\tilde{W} \in \mathcal{W}(\alpha,Q^{(n)})} \Big[D(W||P_{Y|X}|Q) + [I_{Q^{(n)}_{X\tilde{X}Y}}(X, Y; \tilde{X}) - H(Q^{(n)}) + R]^+ \Big].$$

The existence of such minimizers is guaranteed by the fact that we are minimizing a lower semicontinuous function over a non-empty compact set. Therefore, for the sequence $W^{(n)}, \tilde{W}^{(n)}, Q^{(n)}_{X\tilde{X}Y}(W^{(n)}, \tilde{W}^{(n)})$, we have

$$e(\alpha, Q^{(n)}, P_{XY}, H(Q^{(n)})-R) = D(W^{(n)}||P_{Y|X}|Q)+[I_{Q^{(n)}_{X\tilde{X}Y}}(X, Y; \tilde{X})-H(Q^{(n)})+R]^+, \tag{5.20}$$

and

$$Q^{(n)}_{Y|X} = W^{(n)}, Q^{(n)}_{Y|\tilde{X}} = \tilde{W}^{(n)}, Q^{(n)}_X = Q^{(n)}_{\tilde{X}} = Q^{(n)}, I^{(n)}_{Q_{X\tilde{X}}}(X; \tilde{X}) \leq H(Q^{(n)}) - R \tag{5.21}$$

and $\tilde{W}^{(n)} \leq_\alpha W^{(n)}$.

By compactness we can find a convergent subsequence, and by relabeling we may arrange it so that

$$W^{(n)} \to W, \tilde{W}^{(n)} \to \tilde{W}, Q^{(n)}_{X\tilde{X}Y} \to Q_{X\tilde{X}Y}.$$

Furthermore, taking the limits of both sides of each equation in (5.21) and recalling the continuity assumption for α, it follows that

$$Q_{Y|X} = W, Q_{Y|\tilde{X}} = \tilde{W}, Q_X = Q_{\tilde{X}} = Q, I_{Q_{X\tilde{X}}}(X;\tilde{X}) \leq H(Q) - R \text{ and } \tilde{W} \leq_\alpha W. \tag{5.22}$$

Now, taking the $\liminf$ of both sides of (5.20) gives

$$\begin{aligned}
&\liminf_{n\to\infty} E(\alpha, Q^{(n)}, P_{XY}, H(Q^{(n)}) - R) = \\
&\quad \liminf_{n\to\infty} D(W^{(n)}||P_{Y|X}|Q) + [I_{Q^{(n)}_{X\tilde{X}Y}}(X,Y;\tilde{X}) - R]^+ \\
&\overset{*}{\geq} D(W||P_{Y|X}|Q) + [I_{Q_{X\tilde{X}Y}}(X,Y;\tilde{X}) - H(Q) + R]^+
\end{aligned}$$

where $(^*)$ follows from the lower semicontinuity of the information measures. The result follows by noticing that (5.22) implies that both $Q_{X\tilde{X}Y} \in \mathcal{Q}(W, \tilde{W}, Q, H(P) - R)$ and $W, \tilde{W} \in \mathcal{W}(\alpha, Q)$. □

Lemma 27. *Let $Q^{(m)}_X$ be a sequence of distributions converging to Q_X, $\{\epsilon_m\}$ a sequence of positive reals converging to zero, $\omega(Q)$ be an upper semicontinuous function of Q, and*

$$F_\epsilon(Q_{XY}, \omega(\cdot), R) =$$
$$\begin{cases} D(Q_X||P_X) + e(\alpha, Q_X, P_{XY}, H(Q_X) - R) & \text{if } \omega(Q_X) \geq R - \epsilon \\ \infty & \text{otherwise,} \end{cases}$$

$$F(Q_{XY}, \omega(\cdot), R) =$$
$$\begin{cases} D(Q_X||P_X) + e(\alpha, Q_X, P_{XY}, H(Q_X) - R) & \text{if } \omega(Q_X) \geq R \\ \infty & \text{otherwise.} \end{cases}$$

Then

$$\liminf_{m\to\infty} F_{\epsilon_m}(Q_X^{(m)}, \omega(Q_X^{(m)}), R) \geq F(Q_X, \omega(Q_X), R) \tag{5.23}$$

Proof. We proceed by cases. Case 1: Q_X is such that $\omega(Q_X) \geq R$. If $\omega(Q_X^{(m)}) < R - \epsilon_m$ for all sufficiently large m, then the left-hand side is infinity and the result trivially holds. Otherwise we appeal to the semicontinuity of $D(Q||P)$ and the functional e (Lemma 26).

Case 2: Q_X is such that $\omega(Q_X) < R$. In this case by hypothesis we have that $\limsup \omega(Q_X^{(m)}) < R$, whence (5.23) holds with equality eventually, because both sides are infinity. □

We these facts established we now prove our main result in this section.

Proof of Theorem 23. We will show the scheme described at the start of subsection 5.4.1 has the performance specified by the theorem. Let $\epsilon > 0$. Note that for n sufficiently large the constraint (5.12) is met (cf. (5.13), (5.15) and the fact that there are only polynomially many types). Therefore using (5.14)

$$-n^{-1}\log P_e \geq \tag{5.24}$$

$$\min_{Q_X \in \mathcal{P}^n(\mathcal{X})} \begin{cases} D(Q_X||P_X) & \text{if } \log(\gamma(G_X^n(T_{Q_X}^n))) \geq nR \\ +e(\alpha, Q_X, P_{XY}, H(Q_X) - R) - \delta_n'' & \\ \infty & \text{otherwise.} \end{cases} \tag{5.25}$$

For each n, let $Q_X^{(n)}$ attain the minimum in the righthand side of (5.24). Along a subsequence where $Q_X^{(n)}$ is such that that the objective in (5.24) evaluated along this subsequence converges to the liminf of the righthand side of (5.24)

(cf. (5.11)) there is a further subsequence that converges to Q_X^∞. Let us relabel so that that $Q_X^{(n)} \to Q_X^\infty$, and so that the liminf is attained. Define

$$\omega(\cdot) = \min(H(G_X, \cdot), \kappa(G_X, \cdot)),$$

and note that ω is upper semicontinuous (using $\kappa(G_X, \cdot)$ property 3 and continuity of $H(G_X, \cdot)$ [86, Lemma 2.3]).

If the sequence of minimizers is such that $n^{-1}\log(\gamma(G_X^n(T^n_{Q_X^{(n)}}))) < R$ for all n sufficiently large then clearly we may write

$$\liminf_{n\to\infty} -n^{-1}\log P_e \geq \inf_{Q_X:\omega(Q_X,G_X)\geq R-\epsilon} D(Q_X||P_X) + e(\alpha, Q_X, P_{XY}, H(Q_X) - R), \tag{5.26}$$

because the righthand side of (5.24) is infinity for all n sufficiently large.

In the opposite case, i.e. there is a subsequence n_k for which

$$n_k{}^{-1}\log(\gamma(G_X^{n_k}(T^{n_k}_{Q_X^{(n_k)}}))) \geq R \text{ for all } k,$$

we argue as follows. For any $\delta > 0$, there exists an n_0 so that $\|Q_X^\infty - Q^{(n_k)}\|_\infty < \delta$ for every $n_k > n_0$. Thus for every $n_k > n_0$

$$n_k{}^{-1}\log(\gamma(G_X^{n_k}(T^{n_k,\delta}_{Q_X^\infty}))) \geq n_k{}^{-1}\log(\gamma(G_X^{n_k}(T^{n_k}_{Q_X^{(n_k)}}))) \geq R.$$

Thus, taking the $\limsup$ we conclude

$$\limsup_{k\to\infty} n_k{}^{-1}\log(\gamma(G_X(T^{n_k,\delta}_{Q_X^\infty}))) \geq R.$$

Now, for δ sufficiently small it follows from (5.4) that

$$\bar{H}(G_X, Q_X^\infty) + \epsilon \geq \limsup_{k\to\infty} n_k{}^{-1}\log(\gamma(G_X^{n_k}(T^{n_k,\delta}_{Q_X^\infty})))$$

and hence

$$\bar{H}(G_X, Q_X^\infty) + \epsilon \geq R. \tag{5.27}$$

Therefore the inequality $\omega(G_X, Q_X^\infty) \geq \bar{H}(G_X, Q_X^\infty)$ (cf. Lemma 5.7 and (5.5)) and (5.27) imply that (5.26) holds in this case.

To complete the proof let $\{\epsilon^{(m)}\}$ be any sequence tending to zero, and let $\tilde{Q}_X^{(m)}$ denote a minimum in (5.26). Taking a subsequence and relabeling we may assume that $\tilde{Q}_X^{(m)} \to \tilde{Q}_X^\infty$. Recalling the definitions from the statement of Lemma 27 gives

$$\begin{aligned}\lim_{m\to\infty} \liminf_{n\to\infty} -n^{-1} \log P_e &\geq \liminf_{m\to\infty} F_{\epsilon_m}(\tilde{Q}_X^{(m)}, \omega, R) \\ &\overset{*}{\geq} F(\tilde{Q}_X^\infty, \omega, R) \\ &\geq \inf_{Q_X \in \mathcal{P}(\mathcal{X})} F(Q_X, \omega, R),\end{aligned}$$

where * follows from Lemma 27. Since the sequence $\{\epsilon_m\}$ was arbitrary we are done.

□

5.4.2 Discussion and Comparisons

The achievable exponent provided by Theorem 23 is no worse than the exponent of Csiszár and Körner [53, Theorem 2] for any continuous α-decoder, which includes both maximum likelihood and minimum entropy decoders. To see this note that the minimization in (5.18) is over all $Q_X \in \mathcal{P}(\mathcal{X})$ satisfying $\min(\kappa(G_X, Q_X), H(G_X, Q_X)) \geq R$, whereas the minimization in (5.14) is over all $Q_X \in \mathcal{P}(\mathcal{X})$.

A second achievable exponent for the present problem is given by Oohama

and Han [57]:

$$e_{OH} \triangleq \min_{Q_{XY}: H(Q_X) \geq R} \Big[D(Q_{XY}||P_{XY}) + [R - H_{Q_{XY}}(X|Y)]^+ \Big]. \tag{5.28}$$

The exponent of Theorem 23 is no worse than (5.28). To see this we apply Corollary 3 (with minimum entropy decoding) to yield the following lower bound on the Theorem 23 exponent:

$$e_{ME} \triangleq \min_{\substack{Q_{XY}: \\ \min(\kappa(G_X,Q_X), H(G_X,Q_X)) \geq R}} \Big[D(Q_{XY}||P_{XY}) + [R - H_{Q_{XY}}(X|Y)]^+ \Big]. \tag{5.29}$$

The Oohama and Han exponent minimizes over all distributions $\{Q_{XY} : H(Q_X) \geq R\}$, whereas (5.29) minimizes over the smaller set of distributions $\{Q_{XY} : \min(\kappa(G_X, Q_X), H(G_X, Q_X)) \geq R\}$ (apply κ property 1 or notice that $H(G_X, Q_X) \leq H(Q_X)$).

For numerical comparisons we study several examples and evaluate e_{ME}, e_{OH} and the following lower bound on the Csiszár and Körner exponent of (5.14)

$$e_{CK} \triangleq \min_{Q_X} \Big[D(Q_X||P_X) + \max(e_r(Q_X, P_{XY}, H(Q_X) - R), e_x(Q_X, P_{XY}, H(Q_X) - R)) \Big]. \tag{5.30}$$

This bound is obtained by using maximum likelihood decoding and applying Lemma 25. It was necessary to use the bound because the complexity of the optimizations required to evaluate (5.14) made computation of (5.14) infeasible, even for the simple examples we study and exploiting convexity. A fairer comparison would be to replace e_{ME} with the stronger bound obtained using maximum likelihood decoding (Corollary 3). However, even the weaker e_{ME} is enough to show numerical improvements over both e_{CK} and e_{OH}.

Example 1, revisited. (Fig. 5.2)For this example we note that the calculations in Section 5.3.1 imply $\kappa(G_X, Q_X) = H(G_X, Q_X)$. In Figure 5.3 we plot e_{ME}

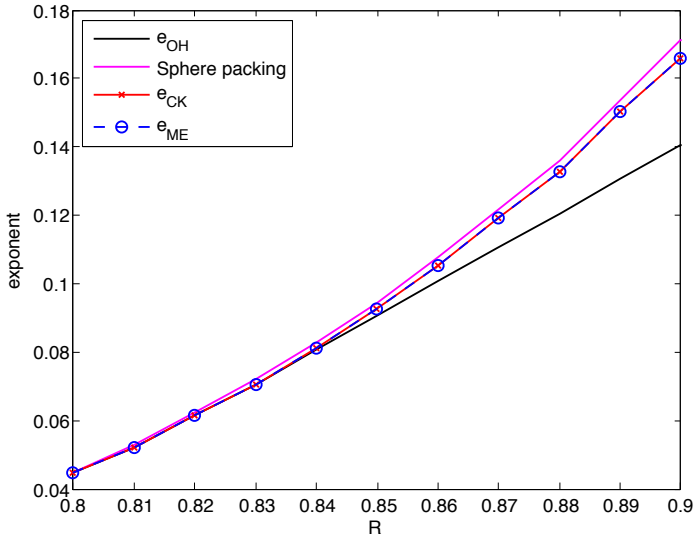

Figure 5.3: Comparing exponents for Example 1 of Figure 5.2. e_{ME} coincides with e_{CK} and both lie below the sphere packing exponent.

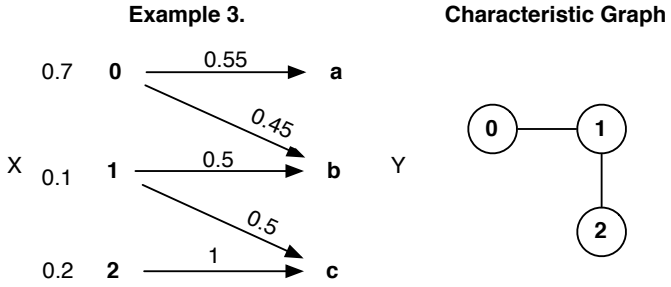

Figure 5.4: Example 3: A source distribution and its characteristic graph

(5.29) against e_{CK}, (5.30), and e_{OH} (5.28). From the figure we see that e_{ME} lies below the sphere packing exponent and above e_{OH}. When compared with e_{CK}, e_{ME} agrees (numerically) and was obtained using a universal minimum entropy decoder.

Example 3. (Fig. 5.4) In this example it is clear that any rate in excess of one bit allows the decoder to determine the source sequence without error. The various error exponents are plotted in Fig 5.5. From the figure we see that e_{ME}

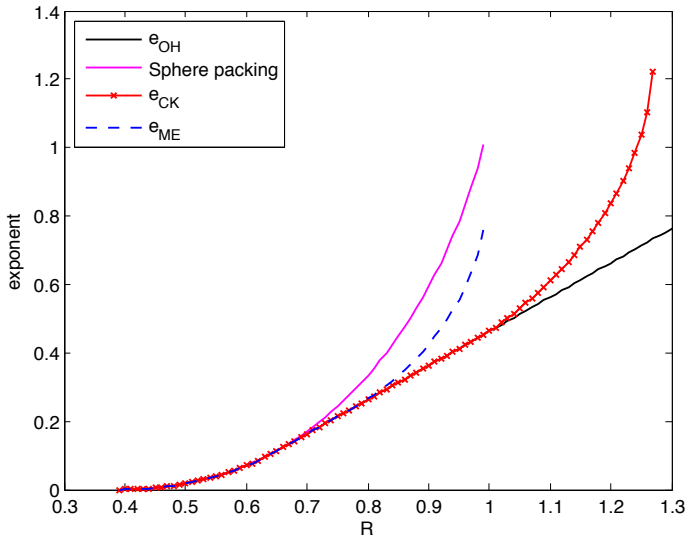

Figure 5.5: Comparing exponents for Example 4 (Figure 5.4). e_{ME} is infinite for all rates above 1 bit, whereas e_{CK} is finite for some rates above 1 bit. Interpret the exponent as infinite to the right of the point that the curve vanishes.

is infinite for all rates above 1 bit since $H(G_X, Q_X) \leq 1$. Notice, however, that e_{CK} is finite for some rates above one bit and therefore e_{CK} is dominated by e_{ME}. The e_{OH} exponent remains finite for all rates below $\log(3)$ bits and is also dominated by e_{ME}. Below 1 bit, e_{OH} and e_{CK} are dominated by e_{ME} in a certain region.

Note: As previously mentioned, the strongest results of Csiszár and Körner [53] are obtained by using the ML decoder in (5.14). However, in the particular case of Example 3, we note that if for some R the exponent e_{CK} is finite, then there exists a Q_X for which

$$\min_{\substack{Q_{\tilde{X}X}: H(\tilde{X}|X) \geq R \\ Q_{\tilde{X}} = Q_X}} \mathbb{E}[d_B(X, \tilde{X})] + R - H(\tilde{X}|X) < \infty.$$

Following the definitions (cf. equations (28) and (16) in [53]), this implies that the exponent of (5.14) would too be finite. Yet, as we see from Figure 5.4, e_{CK}

is finite for some rates above 1 bit and hence (5.14) is too. Thus, at least for Example 3, e_{ME} strictly improves upon the exponent of (5.14).

We conclude by observing that e_{ME} gives the optimal exponent in the case of deterministic side information.

Example 4, Deterministic side information. Suppose the side information is a deterministic function of the source, i.e. $Y = f(X)$ and let $P_{Y|X}$ denote the induced conditional distribution. In this case κ property 2 yields $\kappa(G_X, Q_X) = H_{Q \times P_{Y|X}} H(X|Y)$. Furthermore the minimization of (5.29) must select $Q_{Y|X} = P_{Y|X}$, i.e. the 'deterministic' side information. These observations imply that e_{ME} reduces to

$$e_{SP}(R, P_{XY}) = \min_{Q_{XY}: H_{Q_{XY}}(X|Y) \geq R} D(Q_{XY}||P_{XY}),$$

the sphere packing exponent for this problem. Thus the minimum entropy scheme is optimal for all rates and the reliability function is determined for this problem.

5.5 Improved Exponents for Wyner-Ziv

When dealing with lossy reproduction it is often convenient to use 'covering' (i.e. quantization) followed by binning and in this section we describe how use of the characteristic graph can yield improved error exponents in such scenarios. We focus on lossy compression with side information i.e. Wyner-Ziv [77]. Formally the error exponent problem in this case is as follows.

Let $\hat{\mathcal{X}}$ be the reproduction alphabet and $d : \mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}$ a single letter distortion measure. Define the distortion between two strings as $d(\mathbf{x}, \hat{\mathbf{x}}) =$

$\frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{x}_i)$. The encoder/decoder pair are functions $f^n : \mathcal{X}^n \to \mathcal{M}$ and $g^n : \mathcal{M} \times \mathcal{Y}^n \to \hat{\mathcal{X}}^n$, where $\mathcal{M}$ is a fixed set.

Let $\hat{X}^n = g^n(f^n(X^n), Y^n)$ be the decoder's output and define the error probability

$$P_e(f^n, g^n, \Delta, d) = \Pr\left(d(X^n, \hat{X}^n) > \Delta\right). \tag{5.31}$$

We define the Wyner-Ziv error exponent to be

$$\pi(R, \Delta, P_{XY}, d) = \lim_{\epsilon \downarrow 0} \liminf_{n \to \infty} -\frac{1}{n} \log \left[\min_{(f^n, g^n)} P_e(f^n, g^n, \Delta, d)\right] \tag{5.32}$$

where the minimization ranges over all encoder/decoder pairs satisfying

$$\log |\mathcal{M}| \leq n(R + \epsilon). \tag{5.33}$$

Before we state the result we define another graph functional.

Definition 12.

$$\kappa_2(P_{XY}, Q_{XYU}) = [\kappa(G_U, Q_U) - H(Q_{U|X}|Q_X)]^+,$$

where the graph G_U is defined from the distribution

$$Q_{UY}(u, y) = \sum_{x \in \mathcal{X}} P_{XY}(x, y) Q_{U|X}(u|x).$$

Note: Since P_{XY} will be fixed throughout, we will abbreviate to $\kappa_2(Q_{XYU})$ or even simply $\kappa_2(Q_X)$.

Our first result in this section is Theorem 24.

Theorem 24. *Let $P_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and $R > 0$, $\Delta > 0$, $d(\cdot, \cdot)$ be given. Then*

$$\pi(R, \Delta, P_{XY}, d) \geq \inf_{Q_X} \sup_{Q_{U|X}} \inf_{Q_Y} \sup_{\phi \in \mathcal{F}} \inf_{Q_{XYU}} \eta(R, P_{XY}, Q_{XYU}, \phi)$$

where

$$\eta(R, P_{XY}, Q_{XYU}, \phi) = \begin{cases} D(Q_{XYU}||P_{XY}Q_{U|X}) & \text{if } \mathbb{E}_Q[d(X, \phi(Y, U))] \geq \Delta \\ D(Q_{XYU}||P_{XY}Q_{U|X}) & \text{if } \mathbb{E}_Q[d(X, \phi(Y, U))] < \Delta \\ \quad +[R - I_Q(X;U) + I_Q(Y;U)]^+ & \quad \text{and } \kappa_2(P_{XY}, Q_{XYU}) \geq R \\ \infty & \text{otherwise} \end{cases}$$

and $\mathcal{F} = \{\phi | \phi : \mathcal{Y} \times \mathcal{U} \to \hat{\mathcal{X}}\}$. Note in the final minimization over Q_{XYU}, Q_{XU} and Q_Y are fixed to be those specified earlier in the optimization.

Proof. See Appendix. □

5.5.1 Discussion of Result

In the previous chapter, we determined an achievable exponent for the Wyner-Ziv problem, obtained by replacing η in Theorem 24 with

$$\eta_D(R, P_{XY}, Q_{XYU}, \phi) = \begin{cases} D(Q_{XYU}||P_{XY}Q_{U|X}) & \text{if } \mathbb{E}_Q[d(X, \phi(Y, U))] \geq \Delta \\ D(Q_{XYU}||P_{XY}Q_{U|X}) & \text{if } \mathbb{E}_Q[d(X, \phi(Y, U))] < \Delta \\ \quad +[R - I_Q(X;U) + I_Q(Y;U)]^+ & \quad \text{and } I(X;U) \geq R \\ \infty & \text{otherwise,} \end{cases}$$

the difference being the conditions under which we switch from case 2 to case 3. Theorem 24 is obtained by modifying the scheme in the previous chapter taking

into account the graph-based expurgation established in the previous section. Recalling κ property 1 we have the following inequality

$$\begin{aligned}\kappa_2(Q_{XYU}) &= [\kappa(G_U, Q_U) - H(U|X)]^+ \\ &\leq [H(U) - H(U|X)]^+ \\ &= I(X;U)\end{aligned}$$

therefore for any R, P_{XY}, ϕ and Q_{XYU} we see that $\eta_D(R, P_{XY}, Q_{XYU}, \phi) \leq \eta(R, P_{XY}, Q_{XYU}, \phi)$ and the present modification yields an achievable exponent that is never any worse than the result of the previous chapter.

5.5.2 Sketch of Scheme

Operating at blocks of length n, for each type Q_X, a test channel $Q^*_{U|X}(Q_X) = Q_{U^*|X}$ is selected. The test channel is used to generate a codebook, $B^n(Q_X)$, of approximately $2^{nI(U^*;X)}$ codewords. The key insight is that the (random) graph $B^n(Q_X) \cap G^n_{U^*}$, constructed from

$$Q_{U^*Y}(u, y) = \sum_{x \in \mathcal{X}} P_{XY}(x, y) Q_{U^*|X}(u|x)$$

plays the same role in this problem as did the graph characteristic graph of the source P_{XY} in the Slepian-Wolf problem.

In this modified scheme, the encoder first communicates the type of X^n and then if there is sufficient rate, i.e. $nR > \log \gamma(B^n(Q_X) \cap G^n_{U^*})$, rather than communicating a bin index the encoder may send the color of the codeword in the graph G_{U^*}. If there is insufficient rate, then the encoder communicates a bin index of the codeword. For each pair marginal types (Q_X, Q_Y) the decoder can

choose an estimation function ϕ and depending on the case, either decodes using the graph, or a minimum empirical entropy decoder. The estimation function is then used to combine the side information and the codeword to yield the reproduction.

The careful reader will notice our improvement makes use of our κ functional (via κ_2), but not graph entropy $H(G_X, Q_X)$, to bound the chromatic number. Primarily this is to keep the analysis shorter, and in principle there is no reason why a similar argument using $H(G_X, Q_X)$ and then taking the best bound would not work. As we shall see in the next sub-section, for improving exponents, using κ is enough.

5.5.3 Deterministic Side Information

We now use the result of Theorem 24 to determine the reliability function when the side information is a deterministic function of the source, i.e. $Y = f(X)$ a.s. for a deterministic f. We first note that in this case, the solution to the innermost optimization must be $Q_{Y|XU} = P_{Y|X}$ else the exponent is infinite. This reduces the problem to

$$\inf_{Q_X} \sup_{Q_{U|X},\phi} \eta(R, P_{XY}, Q_{XYU}, \phi)$$

where the distribution of Q_{XYU} is $Q_X P_{Y|X} Q_{X|U}$, i.e. U, X and Y form a Markov chain in that order. We can massage the exponent $\inf_{Q_X} \sup_{Q_{U|X},\phi} \eta(R, P_{XY}, Q_{XYU}, \phi)$ as follows

$$\inf_{Q_X} \sup_{Q_{U|X},\phi} \begin{cases} D(Q_{XYU}||P_{XY}Q_{U|X}) & \text{if } \mathbb{E}_Q[d(X,\phi(Y,U))] \geq \Delta \\ D(Q_{XYU}||P_{XY}Q_{U|X})+ & \text{if } \mathbb{E}_Q[d(X,\phi(Y,U))] < \Delta \\ \quad [R - I_Q(X;U) + I_Q(Y;U)]^+ & \text{and } \kappa_2(Q_{XYU}) \geq R \\ \infty & \text{otherwise} \end{cases}$$

$$\geq \inf_{Q_X} \sup_{Q_{U|X}:Y=\nu(U),\phi} \begin{cases} D(Q_{XYU}||P_{XY}Q_{U|X}) & \text{if } \mathbb{E}_Q[d(X,\phi(Y,U))] \geq \Delta \\ D(Q_{XYU}||P_{XY}Q_{U|X})+ & \text{if } \mathbb{E}_Q[d(X,\phi(Y,U))] < \Delta \\ \quad [R - I_Q(X;U) + I_Q(Y;U)]^+ & \text{and } [H(U|Y) - H(U|X)]^+ \geq R \\ \infty & \text{otherwise} \end{cases}$$

where the previous inequality follows because we maximize over a smaller set. The notation $Q_{U|X} : Y = \nu(U)$ means we consider only those test channels that result in Y being a deterministic function ν of U. By construction U, X and Y still form a Markov chain in that order, thus $H(U|X) = H(U|XY)$ and we can continue the chain of equalities with

$$= \inf_{Q_X} \sup_{Q_{U|X}:Y=\nu(U),\phi} \begin{cases} D(Q_{XYU}||P_{XY}Q_{U|X}) & \text{if } \mathbb{E}_Q[d(X,\phi(Y,U))] \geq \Delta \\ D(Q_{XYU}||P_{XY}Q_{U|X})+ & \text{if } \mathbb{E}_Q[d(X,\phi(Y,U))] < \Delta \\ \quad [R - I_Q(X;U|Y)]^+ & \text{and } I(X;U|Y) \geq R \\ \infty & \text{otherwise.} \end{cases}$$

Note now that the only difference between Q_{XYU} and $P_{XY}Q_{U|X}$ occurs in Q_X, so it follows that the quantity above can be written as

$$= \inf_{Q_X} \sup_{Q_{U|X}:Y=\nu(U),\phi} \begin{cases} D(Q_X||P_X) & \text{if } \mathbb{E}_Q[d(X,\phi(Y,U))] \geq \Delta \text{ or } I(X;U|Y) \geq R \\ \infty & \text{otherwise.} \end{cases}$$

$$= \inf_{Q_X} \sup_{Q_{U|X},\phi} \begin{cases} D(Q_X||P_X) & \text{if } \mathbb{E}_Q[d(X,\phi(Y,U))] \geq \Delta \text{ or } I(X;U|Y) \geq R \\ \infty & \text{otherwise} \end{cases}$$

To argue the final equality, let Q_X and R be fixed. The direction $\leq$ is clear since we maximize over a larger set. For $\geq$, it suffices to show that if the optimization on the left side yields $D(Q_X||P_X)$ then so does the optimization on the right. On account of the fact that the objective is piecewise constant (over $Q_{U|X}$ and ϕ), when the left side is finite, there exists a $Q^*_{U|X} : Y = \nu(U)$ and ϕ causing evaluation to $D(Q_X||P_X)$. Suppose by way of contradiction there exists a non-deterministic $Q_{U|X}$ which yields an infinite exponent. This means that

$$I(X;U|Y) < R \text{ and } \mathbb{E}_Q[d(X,\phi(Y,U))] < \Delta$$

but then by Lemma 28 (which follows) we can find a deterministic $Q_{\tilde{U}|X}$ and corresponding $\tilde{\phi}$ with the property that

$$I(X;\tilde{U}|Y) < R \text{ and } \mathbb{E}_Q[d(X,\tilde{\phi}(Y,\tilde{U}))] < \Delta$$

implying that $Q_{\tilde{U}|X}$ would yield an infinite exponent, contradicting the optimality of $Q^*_{U|X}$.

Lemma 28. *Let Q_X be given and let $Y = f(X)$ with $P_{Y|X}$ denoting the induced conditional distribution. Then for any $Q_{U|X}, \phi$, there exists a $Q_{\tilde{U}|X}$ and $\tilde{\phi}$ so that when $Q_{XYU} = Q_X Q_{U|X} P_{Y|X}$,*

$$1)\ \mathbb{E}_{Q_{XYU}}[d(X,\phi(Y,U))] = \mathbb{E}_{Q_{XY\tilde{U}}}[d(X,\tilde{\phi}(Y,\tilde{U}))],$$

$$2)\ I(X;U|Y) = I(X;\tilde{U}|Y)$$

and 3) $Y = \nu(\tilde{U})$ for some deterministic function ν.

Proof. Define $\tilde{U} = (U, Y)$ and $\tilde{\phi}(Y, \tilde{U}) = \phi(Y, U)$. Then clearly conditions 1 and 3 hold. To see condition 2 note by the chain rule

$$I(X; \tilde{U}|Y) = I(X; U, Y|Y) = I(X; U|Y) + I(X; Y|Y, U) = I(X; U|Y).$$

Finally we point out that since $Y = f(X)$ we also have $\tilde{U} \leftrightarrow X \leftrightarrow Y$. □

Rewriting this final optimization problem as

$$\begin{aligned}
&\inf_{Q_X} \sup_{Q_{U|X}, \phi} \begin{cases} D(Q_X||P_X) & \text{if } \mathbb{E}_Q[d(X, \phi(Y, U))] \geq \Delta \text{ or } I(X; U|Y) \geq R \\ \infty & \text{otherwise} \end{cases} \\
&= \inf_{Q_X : R_{WZ}(\Delta, Q_X) \geq R} D(Q_X||P_X) \\
&\leq \pi(R, \Delta, P_{XY}, d)
\end{aligned}$$

where $R_{WZ}(\Delta, Q_X)$ denotes the Wyner-Ziv rate distortion function for the source with $X \sim Q_X$ and $Y = f(X)$ with distortion measure d. But according to the change-of-measure argument of 18,

$$\pi(R, \Delta, P_{XY}, d) \leq \inf_{Q_X : R_{WZ}(\Delta, Q_X) \geq R} D(Q_X||P_X).$$

Thus our scheme is optimal in the sense that it meets the change-of-measure upper bound.

5.6 Connection to Channel Coding

In this section we briefly mention that κ has applications in zero-error channel coding problems. Let $G = G(W)$ be the characteristic graph of the channel W,

and $c(G)$ denote the zero error capacity (see [61, Section III] for definitions). It is known that [81]

$$c(G) = \max_{Q_X} [H(Q_X) - \bar{H}(G_X, Q_X)]$$

and therefore Lemma 24 implies

$$c(G) \geq \max_{Q_X} [H(Q_X) - \kappa(G_X, Q_X)].$$

Thus κ can be used to provide a lower bound on zero-error channel capacity.

APPENDIX A

CHAPTER 2 - PROOFS

A.1 Proofs: Section 2.2

This appendix is dedicated to the proof of Lemma 2, showing that the variance of F, the test random variable used for α-sources tends to zero when $0 < \alpha < 2$. (Note, such a result does not follow via other means, say Efron-Stein or bounded differences conditions.) We start by reproducing the moments of the binomial distribution.

Lemma 29 (Higher Moments of the Binomial)**.** *Suppose* $N \sim$ *Binomial*(n, p)

$$\begin{aligned}
\mathbb{E}[N^2] &= n^2p^2 + np(1-p) \\
\mathbb{E}[N^3] &= n^3p^3 + 3n^2p^2 - 3n^2p^3 + np - 3np^2 + 2np^3 \\
\mathbb{E}[N^4] &= n^4p^4 + 6n^3p^3 - 6n^3p^4 + 7n^2p^2 - 18n^2p^3 + 11n^2p^4 + np \\
&\quad - 7np^2 + 12np^3 - 6np^4
\end{aligned}$$

Proof. Direct calculation. □

Computing the variance of F will require that the following results on covariance of multinomial vectors.

Lemma 30. *Suppose* $X^n \sim p^n$*. For* $a \neq b$

$$\begin{aligned}
\mathbb{E}[N^2(a|X^n)N^2(b|X^n)] &= n(n-1)(n-2)(n-3)p^2(a)p^2(b) \\
&\quad + n(n-1)(n-2)[p(a)p^2(b) + p^2(a)p(b)] \\
&\quad + n(n-1)p(a)p(b)
\end{aligned}$$

Proof. Start by writing

$$\mathbb{E}[N^2(a|X^n)]N^2(b|X^n)] = \mathbb{E}\Big[\Big(\sum_{i=1}^{n}\mathbf{1}\{X_i = a\}\Big)^2\Big(\sum_{i=1}^{n}\mathbf{1}\{X_i = b\}\Big)\Big]$$
$$= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\mathbb{E}[\mathbf{1}\{X_i = a\}\mathbf{1}\{X_j = a\}\mathbf{1}\{X_k = b\}\mathbf{1}\{X_l = b\}]$$

Now observe that only certain cases have positive expectation these are

1. $i \neq j \neq k \neq l$ which occurs $n(n-1)(n-2)(n-3)$ times.
2. $i = j$ and $k \neq l$ with $i \neq k$ and $i \neq l$, which occurs $n(n-1)(n-2)$ times
3. $k = l$ and $i \neq j$ with $k \neq i$ and $k \neq j$, which occurs $n(n-1)(n-2)$ times.
4. $i = j$ and $k = l$ with $i \neq k$ which occurs $n(n-1)$ times.

□

Lemma 31. *Suppose $X^n \sim p^n$. For $a \neq b$*

$$\mathbb{E}[N^2(a|X^n)N(b|X^n)] = n(n-1)(n-2)p^2(a)p(b) + n(n-1)p(a)p(b)$$
$$= (n^3 - 3n^2 + 2n)p^2(a)p(b) + (n^2 - n)p(a)p(b)$$

Proof. We have

$$\mathbb{E}[N^2(a|X^n)N(b|X^n)] = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\mathbb{E}[\mathbf{1}\{X_i = a\}\mathbf{1}\{X_j = a\}\mathbf{1}\{X_k = b\}]$$

As in the proof of Lemma 30 only cases $i = j \neq k$ and $i \neq j \neq k$ yield a positive expectation. □

To simplify the analysis we will use the following lemma to discard terms that vanish in the limit.

Lemma 32 (Discarding Rule)**.** *Suppose* $0 < \alpha < 2$ *and for all* $a, b \in \mathcal{A}_n$ *that* $p(a) = O(n^{-\alpha})$ *and* $q(b) = O(n^{-\alpha})$ *. For integers* i, j *such that* $4 \geq j \geq i \geq 2$

$$1.\ n^{2\alpha-4} \sum_{a \in \mathcal{A}_n} n^i p^j(a) \to 0 \text{ as } n \to \infty.$$

For positive integers i, j, k *such that* $4 \geq j + k > i \geq 2$ *or* $j + k = 4$ *and* $i = 1$

$$2.\ n^{2\alpha-4} \sum_{a,b \in \mathcal{A}_n} n^i p^j(a) q^k(b) \to 0 \text{ as } n \to \infty.$$

Proof. For the first property

$$\begin{aligned} n^{2\alpha-4} \sum_{a \in \mathcal{A}_n} n^i p^j &\leq n^{2\alpha-4} \frac{3n^\alpha}{\check{c}} n^i \Big(\frac{\hat{c}}{n^\alpha}\Big)^j \\ &= n^{\alpha(3-j)-4+i} \frac{3\hat{c}^j}{\check{c}} \end{aligned}$$

Since $\alpha < 2$, examining the exponent alone we have for $3 \geq j$

$$\begin{aligned} \alpha(3-j) - 4 + i &< 2 - 2j + i \\ &\leq 2 - 2i + i \\ &\leq 2 - i \\ &\leq 0 \end{aligned}$$

i.e $\alpha(3-j) - 4 + i < 0$. When $j = 4$ we have

$$-\alpha - 4 + i,$$

so for $i = 3$ the exponent is $-\alpha - 1 < 0$ and for $i = 4$ it is $-\alpha < 0$.

For the second property, argue with cases:

$$n^{2\alpha-4} \sum_{a,b \in \mathcal{A}_n} n^i p^j(a) q^k(b) \leq n^{4\alpha-4-(j+k)\alpha+i} \frac{9\hat{c}^{j+k}}{\check{c}^2}$$

when $i = 2, j + k = 3$ the sum behaves like $n^{\alpha-2}$, for $i = 2, (j + k) = 4$ it behaves like n^{-2} and for $i = 3, (j + k) = 4$ it behaves like n^{-1}, thus in all three cases the

sum goes to zero when $0 < \alpha < 2$ as $n \to \infty$. For $j + k = 4$ and $i = 1$ the sum behaves like n^{-3}, which again goes to zero as $n \to \infty$. □

Lemma (2). *For* $i = 0, 1$

$$\mathrm{Var}_i[n^\alpha F] \to 0$$

for all $0 < \alpha < 2$.

Proof. Throughout we suppose hypothesis $\mathcal{H}_1$ is in effect and simply write $\mathbb{E}$ for $\mathbb{E}_1$, the other case is handled analogously.

$$\begin{aligned}
\mathbb{E}[F^2] = \mathbb{E}[\|\Lambda_{X^n}\|_2^4] &- 2\mathbb{E}[\|\Lambda_{X^n}\|_2^2]\mathbb{E}[\|\Lambda_{Y^n}\|_2^2] \\
&- 4\mathbb{E}[\|\Lambda_{X^n}\|_2^2\langle\Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle] + \mathbb{E}[\|\Lambda_{Y^n}\|_2^4] \\
&+ 4\mathbb{E}[\|\Lambda_{Y^n}\|_2^2\langle\Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle] \\
&+ 4\mathbb{E}[\langle\Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle^2].
\end{aligned}$$

Notice that every term in the expansion above has a common factor n^{-4} and therefore we will be dealing with terms such as

$$\begin{aligned}
n^{2\alpha}\mathbb{E}[\|\Lambda_{X^n}\|_2^4] &= n^{2\alpha-4}\mathbb{E}\Big[\Big(\sum_{a\in\mathcal{A}_n} N^2(a|X^n)\Big)^2\Big] \\
&= n^{2\alpha-4}\mathbb{E}\Big[\sum_{a,b\in\mathcal{A}_n} N^2(a|X^n)N^2(b|X^n)\Big] \\
&= n^{2\alpha-4}\Big[\sum_{a\in\mathcal{A}_n} \mathbb{E}[N^4(a|X^n)] + \sum_{a\neq b\in\mathcal{A}_n} \mathbb{E}[N^2(a|X^n)N^2(b|X^n)]\Big].
\end{aligned} \tag{A.1}$$

Using the discarding rule (Lemma 32) we can safely ignore terms that vanish in

the limit. For example since $N(a|X^n)$ is binomial, recalling Fact 29 we see

$$\begin{aligned} n^{2\alpha-4}\sum_{a\in\mathcal{A}_n}\mathbb{E}[N^4(a|X^n)] &= n^{2\alpha-4}\sum_{a\in\mathcal{A}_n} n^4p_n^4(a) + 6n^3p_n^3(a) - 6n^3p_n^4(a) + 7n^2p_n^2(a) \\ &\quad - 18n^2p_n^3(a) + 11n^2p_n^4(a) + np_n(a) - 7np_n^2(a) \\ &\quad + 12np_n^3(a) - 6np_n^4(a) \\ &\simeq n^{2\alpha-4}\sum_{a\in\mathcal{A}_n} np_n(a) = n^{2\alpha-4}n, \end{aligned} \tag{A.2}$$

where the notation

$$a_n \simeq b_n \text{ means } \lim_{n\to\infty} a_n - b_n = 0.$$

For the "cross-terms", by Lemma 30 we have

$$\begin{aligned} \sum_{a\neq b}\mathbb{E}[N^2(a|X^n)N^2(b|X^n)] &= \sum_{a\neq b} n^4p_n^2(a)p_n^2(b) - 6n^3p_n^2(a)p_n^2(b) + 11n^2p_n^2(a)p_n^2(b) \\ &\quad - 6np_n^2(a)p_n^2(b) \\ &\quad + (n^3-3n^2+2n)p_n^2(a)p_n(b) + (n^3-3n^2+2n)p_n^2(b)p_n(a) \\ &\quad + (n^2-n)p_n(a)p_n(b) \end{aligned}$$

Note that

$$\sum_{a\neq b} p(a)q^i(b) = \sum_b q^i(b)(1-p(b)) = \sum_a q^i(a) - q^i(a)p(a)$$

therefore

$$\begin{aligned} \sum_{a\neq b}\mathbb{E}[N^2(a|X^n)N^2(b|X^n)] &= \sum_{a\neq b} n^4p_n^2(a)p_n^2(b) - 6n^3p_n^2(a)p_n^2(b) + 11n^2p_n^2(a)p_n^2(b) \\ &\quad - 6np_n^2(a)p_n^2(b) \\ &\quad + 2(n^3-3n^2+2n)\Big[\sum_a p_n^2(a) - p_n^3(a)\Big] \\ &\quad + (n^2-n) - (n^2-n)\sum_a p_n^2(a). \end{aligned}$$

Applying the discarding rule we see the terms

$$\sum_{a\neq b} n^4 p_n^2(a)p_n^2(b) + 2n^3 \sum_a p_n^2(a) + n^2 - n$$

are significant in the limit. Therefore combining the previous display and (A.2) calculations gives

$$n^{2\alpha}\mathbb{E}[\|\Lambda_{X^n}\|_2^4] \simeq n^{2\alpha-4}\Big(\sum_{a\neq b} n^4 p_n^2(a)p_n^2(b) + 2n^3\sum_a p_n^2(a) + n^2\Big).$$

An analogous argument tells us that

$$n^{2\alpha}\mathbb{E}[\|\Lambda_{Y^n}\|_2^4] \simeq n^{2\alpha-4}\Big(\sum_{a\neq b} n^4 q_n^2(a)q_n^2(b) + 2n^3\sum_a q_n^2(a) + n^2\Big).$$

We now turn our attention to

$$\begin{aligned}
&n^{2\alpha}\mathbb{E}[\|\Lambda_{X^n}\|_2^2]\mathbb{E}[\|\Lambda_{Y^n}\|_2^2] \\
&= n^{2\alpha-4}\Big(\sum_a n^2 p_n^2(a) + np_n(a) - np_n^2(a)\Big)\Big(\sum_a n^2 q_n^2(a) + nq_n(a) - nq_n^2(a)\Big) \\
&= n^{2\alpha-4}\Big(n + \sum_a n^2 p_n^2(a) - np_n^2(a)\Big)\Big(n + \sum_a n^2 q_n^2(a) - nq_n^2(a)\Big) \\
&= n^{2\alpha-4}\Big(n^2 + \sum_a (n^3 q_n^2(a) - n^2 q_n^2(a)) + \sum_a (n^3 p_n^2(a) - n^2 p_n^2(a)) \\
&\quad + \sum_{a,b}(n^2 p_n^2(a) - np_n^2(a))(n^2 q_n^2(b) - nq_n^2(b))\Big)
\end{aligned}$$

In the final sum, the expansion starts with $n^4 p_n^2(a)q_n^2(b)$ plus terms of lower order in n (still with a product of 4 probabilities), therefore applying our discarding rule we see

$$\begin{aligned}
&- 2n^{2\alpha}\mathbb{E}[\|\Lambda_{X^n}\|_2^2]\mathbb{E}[\|\Lambda_{Y^n}\|_2^2] \\
&\simeq -2n^{2\alpha-4}\Big(n^2 + \sum_a n^3 q_n^2(a) + \sum_a n^3 p_n^2(a) + \sum_{a,b} n^4 p_n^2(a)q_n^2(b)\Big).
\end{aligned}$$

Now we turn to

$$\mathbb{E}[\|\Lambda_{X^n}\|_2^2\langle \Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n}\rangle] = \mathbb{E}[\|\Lambda_{X^n}\|_2^2\langle \Lambda_{Z^n}, \Lambda_{X^n}\rangle] - \mathbb{E}[\|\Lambda_{X^n}\|_2^2]\mathbb{E}[\langle \Lambda_{Z^n}, \Lambda_{Y^n}\rangle] \tag{A.3}$$

The first term on the right is

$$
\begin{aligned}
&n^{-4}\mathbb{E}\Big[\Big(\sum_a N^2(a|X^n)\Big)\Big(\sum_a N(a|Z^n)N(a|X^n)\Big)\Big] \\
&= n^{-4}\sum_{a,b}\mathbb{E}[N^2(a|X^n)N(b|X^n)]\mathbb{E}[N(b|Z^n)] \\
&= n^{-4}\sum_a \mathbb{E}[N^3(a|X^n)]\mathbb{E}[N(a|Z^n)] + n^{-4}\sum_{a\neq b}\mathbb{E}[N^2(a|X^n)N(b|X^n)]\mathbb{E}[N(b|Z^n)].
\end{aligned}
\tag{A.4}
$$

Applying Fact 1 and the discarding rule to the first sum in (A.4) gives

$$
\begin{aligned}
&n^{2\alpha-4}\sum_a \mathbb{E}[N^3(a|X^n)]\mathbb{E}[N(a|Z^n)] = \\
&n^{2\alpha-4}\Big[\sum_a (n^3p_n^3(a) + 3n^2p_n^2(a) - 3n^2p_n^3(a) + np_n(a) \\
&\qquad\qquad - 3np_n^2(a) + 2np_n^3(a))nq_n(a)\Big] \\
&\simeq 0.
\end{aligned}
$$

For the second sum of (A.4), applying Lemma 31 gives

$$
n^{-4}\Big[\sum_{a\neq b}(n^4 - 3n^3 + 2n^2)p_n^2(a)p_n(b)q_n(b) + (n^3 - n^2)p_n(a)p_n(b)q_n(b)\Big]
$$

and we see only terms

$$
n^{-4}\Big[\sum_{a\neq b} n^4p_n^2(a)p_n(b)q_n(b) + n^3p_n(a)p_n(b)q_n(b)\Big]
$$

are significant. Turning to the second term of the right of (A.3) we have

$$
\mathbb{E}[\|\Lambda_{X^n}\|_2^2]\mathbb{E}[\langle\Lambda_{Z^n},\Lambda_{Y^n}\rangle] = n^{-4}\sum_{a,b}[np_n(a) + n^2p_n^2(a) - np_n^2(a)]n^2q_n^2(b)
$$

and it follows that the significant terms are

$$
n^{-4}\sum_{a\neq b} n^3p_n(a)q_n^2(b) + n^4p_n^2(a)q_n^2(b).
$$

Therefore

$$
\begin{aligned}
&-4n^{2\alpha}\mathbb{E}[\|\Lambda_{X^n}\|_2^2\langle\Lambda_{Z^n},\Lambda_{X^n}-\Lambda_{Y^n}\rangle]\\
&\simeq -4n^{2\alpha-4}\Big(\sum_{a\neq b} n^4p_n^2(a)p_n(b)q_n(b)+n^3p_n(a)p_n(b)q_n(b)-n^3p_n(a)q_n^2(b)-n^4p_n^2(a)q_n^2(b)\Big).
\end{aligned}
$$

The term

$$
\mathbb{E}[\|\Lambda_{Y^n}\|_2^2\langle\Lambda_{Z^n},\Lambda_{X^n}-\Lambda_{Y^n}\rangle]
$$

can be handled as above and we see that

$$
\begin{aligned}
&4n^{2\alpha}\mathbb{E}[\|\Lambda_{Y^n}\|_2^2\langle\Lambda_{Z^n},\Lambda_{X^n}-\Lambda_{Y^n}\rangle]\\
&\simeq 4n^{2\alpha-4}\Big(\sum_{a\neq b} n^4q_n^2(a)q_n(b)p_n(b)+n^3q_n(a)q_n(b)p_n(b)-n^3q_n(a)q_n^2(b)-n^4q_n^2(a)q_n^2(b)\Big).
\end{aligned}
$$

The final term is

$$
\begin{aligned}
&\mathbb{E}[\langle\Lambda_{Z^n},\Lambda_{X^n}-\Lambda_{Y^n}\rangle^2]\\
&=n^{-4}\mathbb{E}\Big[\Big(\sum_a N(a|Z^n)\big(N(a|X^n)-N(a|Y^n)\big)\Big)^2\Big]\\
&=n^{-4}\sum_a[nq_n(a)+(n^2-n)q_n^2(a)][np_n(a)+(n^2-n)p_n^2(a)]\\
&\quad-[nq_n(a)+(n^2-n)q_n^2(a)]n^2p_n(a)q_n(a)\\
&\quad-[nq_n(a)+(n^2-n)q_n^2(a)]n^2q_n(a)p_n(a)\\
&\quad+[nq_n(a)+(n^2-n)q_n^2(a)][nq_n(a)+(n^2-n)q_n^2(a)]\\
&\quad+\sum_{a\neq b}(n^2-n)^2q_n(a)q_n(b)p_n(a)q_n(b)-(n^4-n^3)q_n(a)q_n(b)p_n(a)q_n(b)\\
&\quad-(n^4-n^3)q_n(a)q_n(b)q_n(a)p_n(b)+(n^2-n)^2q_n(a)q_n(b)q_n(a)q_n(b).
\end{aligned}
$$

In the last line of the previous display, terms in the summation over a are such that every probability is accompanied by an n of the same or lesser power and therefore these terms vanish in the limit. In the summation over $a \neq b$ every

term involves four probabilities so we only keep the n^4 terms. Hence

$$\mathbb{E}[\langle \Lambda_{Z^n}, \Lambda_{X^n} - \Lambda_{Y^n} \rangle^2]$$
$$\sim 4n^{2\alpha-4}\Big(n^4 \sum_{a\neq b} q_n(a)q_n(b)p_n(a)q_n(b) - q_n(a)q_n^2(b)p_n(a) - q_n^2(a)q_n(b)p_n(b) + q_n^2(a)q_n^2(b)\Big).$$

Combining all the above we have shown that

$$\begin{aligned}
E[n^{2\alpha}F^2] \simeq n^{2\alpha-4}\Bigg[&\Big(n^2 + 2\sum_a n^3 p_n^2(a) + \sum_{a\neq b} n^4 p_n^2(a)p_n^2(b)\Big) \\
&- 2\Big(n^2 + \sum_a n^3 q_n^2(a) + \sum_a n^3 p_n^2(a) + \sum_{a\neq b} n^4 p_n^2(a)q_n^2(b)\Big) \\
&- 4\Big(\sum_{a\neq b} n^4 p_n^2(a)p_n(b)q_n(b) + n^3 p_n(a)p_n(b)q_n(b) - n^3 p_n(a)q_n^2(b) - n^4 p_n^2(a)q_n^2(b)\Big) \\
&+ \Big(\sum_{a\neq b} n^4 q_n^2(a)q_n^2(b) + 2\sum_a n^3 q_n^2(a) + n^2\Big) \\
&+ 4\Big(\sum_{a\neq b} n^4 q_n^2(a)q_n(b)p_n(b) + n^3 q_n(a)q_n(b)p_n(b) - n^3 q_n(a)q_n^2(b) - n^4 q_n^2(a)q_n^2(b)\Big) \\
&+ 4\Big(\sum_{a\neq b} n^4 q_n(a)q_n(b)p_n(a)q_n(b) - n^4 q_n(a)q_n^2(b)p_n(a) \\
&\quad - n^4 q_n^2(a)q_n(b)p_n(b) + n^4 q_n^2(a)q_n^2(b)\Big)\Bigg].
\end{aligned}$$

In the above there are several simplifications, for example all of the n^3 terms self-cancel (note

$$n^3 \sum_{a\neq b} p_n(a)q_n^2(b) = n^3 \sum_a q_n^2(a) - q_n^3(a) \sim n^3 \sum_a q_n^2(a)).$$

After performing the cancellations we have

$$\begin{aligned}
E[n^{2\alpha}F^2] \simeq n^{2\alpha}\Big(&\sum_{a\neq b} p_n^2(a)p_n^2(b) - 4p_n^2(a)p_n(b)q_n(b) + 2p_n^2(a)q_n^2(b) \\
&+ q_n^2(a)q_n^2(b) + 4q_n(a)q_n(b)p_n(a)q_n(b) - 4q_n(a)q_n^2(b)p_n(a)\Big).
\end{aligned}$$

We now compute

$$n^{2\alpha}\mathbb{E}[F]^2 = n^{2\alpha}\Big(\sum_{a\in\mathcal{A}_n}(p_n(a)-q_n(a))^2 + n^{-1}(q_n^2(a)-p_n^2(a))\Big)^2$$

Since every term in the above sum involves a quartic product of probabilities it follows that

$$\begin{aligned}
n^{2\alpha}\mathbb{E}[F]^2 &\simeq n^{2\alpha}\sum_{a\neq b}((p_n(a)-q_n(a))^2((p_n(b)-q_n(b))^2 \\
&= n^{2\alpha}\sum_{a\neq b}(p_n^2(a)-2p_n(a)q_n(a)+q_n^2(a))(p_n^2(b)-2p_n(b)q_n(b)+q_n^2(b)) \\
&= n^{2\alpha}\sum_{a\neq b}p_n^2(a)p_n^2(b)-2p_n^2(a)p_n(b)q_n(b)+p_n^2(a)q_n^2(b) \\
&\quad -2p_n(a)q_n(a)p_n^2(b)+4p_n(a)q_n(a)p_n(b)q_n(b)-2p_n(a)q_n(a)q_n^2(b) \\
&\quad +q_n^2(a)p_n^2(b)-2q_n^2(a)p_n(b)q_n(b)+q_n^2(a)q_n^2(b) \\
&= n^{2\alpha}\sum_{a\neq b}p_n^2(a)p_n^2(b)-4p_n^2(a)p_n(b)q_n(b)+2p_n^2(a)q_n^2(b) \\
&\quad +4p_n(a)q_n(a)p_n(b)q_n(b)-4p_n(a)q_n(a)q_n^2(b)+q_n^2(a)q_n^2(b).
\end{aligned}$$

Therefore we have shown for $0<\alpha<2$

$$n^{2\alpha}\mathbb{E}[F^2] \simeq n^{2\alpha}\mathbb{E}[F]^2$$

giving the result. □

We note that concentration results sharper than those obtained with Chebyshev's inequality and the variance calculation can be obtained in some cases using Martingale techniques. For $\alpha=1$ one such result is as follows.

Theorem 25. *For $\alpha = 1$ and any $\gamma > 0$*

$$\Pr\Big(|F - \mathbb{E}[F]| > \gamma\Big) \leq 2\exp\Big(-\frac{\epsilon^2\gamma^2 n}{96(n^{1/3}+\Theta(1))^2}\Big) \tag{A.5}$$

$$+\Big(1+\frac{\Theta(1)}{\gamma(1-\epsilon)}\Big)3n\exp\Big(-\frac{(n^{1/3}-\Theta(1))^2}{2(\hat{c}+(n^{1/3}-\Theta(1))/3)}\Big). \tag{A.6}$$

Proof.

$$t_y(j) = \begin{cases} j-n & \text{if } j \in \{n+1,\ldots,2n\} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{and } t_z(j) = \begin{cases} j-2n & \text{if } j \in \{2n+1,\ldots,3n\} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\{\mathcal{F}_j\}_{j=1}^{3n}$ be a filtration defined as

$$\mathcal{F}_j = \sigma(X_1^j, Y_1^{t_y(j)}, Z_1^{t_z(j)})$$

and define a Doob Martingale $\{W_j\}_{j=0}^{3n}$ as follows

$$W_j = \begin{cases} \mathbb{E}[F(X^n,Y^n,Z^n)] & \text{if } j=0 \\ \mathbb{E}[F(X^n,Y^n,Z^n)|\mathcal{F}_j] & j \in \{1,\ldots,3n\}. \end{cases}$$

Let $D_j = W_j - W_{j-1}$ be the resulting martingale difference sequence (MDS) and $a^* \in \mathcal{A}_n$ be a most likely symbol over the measures p_n, q_n. Using the bounds established in Lemma 33 we have for $j \in \{1,\ldots,n\}$

$$\begin{aligned}\Pr(|D_j| > \alpha) &\leq \Pr\Big(\frac{2}{n}(N(X_j|X_1^{j-1}) + \Theta(1)) > \alpha\Big) \\ &\leq \Pr(N(a^*|X^n) + \Theta(1) > (n/2)\alpha).\end{aligned}$$

Taking $\alpha = \frac{2}{n}(n^{1/3}+\Theta(1))$ gives

$$\Pr\Big(|D_j| > \frac{2}{n}(n^{1/3}+\Theta(1))\Big) \leq \Pr(N(a^*|X^n) > n^{1/3})$$

We now make use of the following 'Chernoff Inequality' [87], which states that if X is binomial, then

$$\Pr(X \geq \mathbb{E}[X] + \lambda) \leq \exp\Big(- \frac{\lambda^2}{2(\mathbb{E}[X] + \lambda/3)}\Big).$$

Now using $\lambda = n^{1/3} - \hat{c}$ in Chernoff Inequality, we have[1] for $j \in \{1...n\}$

$$\Pr(|D_j| > \frac{2}{n}(n^{1/3} + \Theta(1))) \leq \exp\Big(- \frac{(n^{1/3} - \hat{c})^2}{2(\hat{c} + (n^{1/3} - \hat{c})/3)}\Big),$$

similar bounds apply for $j \in \{n+1, \ldots, 3n\}$. A result of [88, 89] states for any MDS (D_j), for every $\gamma > 0$ and each sequence of positive numbers (w_j) and any $0 < \epsilon < 1$,

$$\Pr\Big(|\sum_j D_j| > \gamma\Big) \leq 2\exp\Big(\frac{-\epsilon^2\gamma^2}{8\sum_{j=1}^n w_j^2}\Big) + \Big(1 + \frac{\|D^*\|_\infty}{\gamma(1-\epsilon)}\Big)\sum_{j=1}^n \Pr(|D_j| > w_j),$$

where $\|D^*\|_\infty = \sup_i \|D_i\|_\infty$. For our particular set of D_i, it follows from Lemma 33 that the worst case jump is only $\Theta(1)$, therefore $\|D_i\|_\infty \leq \Theta(1)$. Choosing $w_j = \frac{2}{n}(n^{1/3} + \Theta(1))$, $j = 1, \ldots, 3n$ gives

$$\Pr\Big(|\sum_j D_j| > \gamma\Big) \leq 2\exp\Big(- \frac{\epsilon^2\gamma^2 n}{96(n^{1/3} + \Theta(1))^2}\Big) + \Big(1 + \frac{\Theta(1)}{\gamma(1-\epsilon)}\Big)3n\exp\Big(- \frac{(n^{1/3} - \Theta(1))^2}{2(\hat{c} + (n^{1/3} - \Theta(1))/3)}\Big).$$

□

Lemma 33. *Let $\{D_j\}$ be the martingale difference sequence appearing in the proof of*

[1]Recall $\Pr(\mathcal{B}(n,p) > x)$ is monotonic increasing in p and for rare events sources $p_n(a) \leq \frac{\hat{c}}{n}$

Theorem 1 and t be the function defined there, then

$$|D_j| \leq \begin{cases} \frac{2}{n}(N(X_j|X_1^{j-1}) + \Theta(1)) & j \in \{1, \ldots, n\} \\ \frac{2}{n}(N(Y_{t_y(j)}|Y_1^{t_y(j)-1}) & \\ \quad + \Theta(1)) & j \in \{n+1, \ldots, 2n\} \\ \frac{2}{n}(N(Z_{t_z(j)}|Y^n) & \\ \quad + N(Z_{t_z(i)}|X^n) + \Theta(1)) & j \in \{2n+1, \ldots, 3n\}. \end{cases}$$

Proof. We will only do the third case, the others are similar. Let $j \in \{2n+1, \ldots, 3n\}$, let $\tilde{Z}_{t_z(j)}$ be an independent copy of $Z_{t_z(j)}$ define $\tilde{Z}^n = (Z_1, \ldots, \tilde{Z}_{t_z(j)}, \ldots, Z_n)$, then

$$\begin{aligned} |D_j| = \frac{1}{n}\Big| \sum_{a \in \mathcal{A}_n} & \mathbb{E}[(N(a|X^n) - N(a|Z^n))^2 \\ & - (N(a|Y^n) - N(a|Z^n))^2 - N((a|X^n) - N(a|\tilde{Z}^n))^2 \\ & + (N(a|Y^n) - N(a|\tilde{Z}^n))^2 | \mathcal{F}_j] \Big|. \end{aligned}$$

Expanding the squares and cancelling gives

$$\begin{aligned} |D_j| &= \frac{1}{n}\Big| \sum_{a \in \mathcal{A}_n} \mathbb{E}[2N(a|X^n)(N(a|\tilde{Z}^n) - N(a|Z^n)) \\ &\quad + 2N(a|Y^n)(N(a|Z^n) - N(a|\tilde{Z}^n))|\mathcal{F}_j] \Big| \\ &= \frac{2}{n}\Big| \sum_{a \in \mathcal{A}_n} \mathbb{E}[N(a|X^n)(\mathbf{1}\{\tilde{Z}_{t_z(j)} = a\} - \mathbf{1}\{Z_{t_z(j)} = a\}) \\ &\quad + N(a|Y^n)(\mathbf{1}\{Z_{t_z(j)} = a\} - \mathbf{1}\{\tilde{Z}_{t_z(j)} = a\})|\mathcal{F}_j] \Big| \\ &= \Big| \frac{2}{n} \sum_{a \in \mathcal{A}_n} (N(a|Y^n) - N(a|X^n))\mathbf{1}\{Z_{t_z(j)} = a\}) \\ &\quad + (N(a|X^n) - N(a|Y^n))\mathbb{E}[\mathbf{1}\{\tilde{Z}_{t_z(j)} = a\})] \end{aligned}$$

where on the previous line we used the fact that $X^n, Y^n, Z_{t_z(j)}$ are measurable with respect to $\mathcal{F}_j$. Applying the triangle inequality and the bound $p_n(a) \leq \hat{c}/n$

for all $a \in \mathcal{A}_n$ gives

$$\begin{aligned}|D_j| &\le \frac{2}{n}\Big(N(Z_{t_z(j)}|X^n) + N(Z_{t_z(j)}|Y^n) \\ &\quad + \sum_{a\in\mathcal{A}_n}(N(a|X^n) + N(a|Y^n))\frac{\hat{c}}{n}\Big) \\ &= \frac{2}{n}(N(Z_{t_z(j)}|X^n) + N(Z_{t_z(j)}|Y^n) + 2\hat{c}).\end{aligned}$$

□

A.2 Proofs: Section 2.3

Lemma (8). *Suppose p and q are distributions on an alphabet $\mathcal{A}$, then*

$$G(p,q,\mathcal{A}) = \sum_{a\in\mathcal{A}}\sum_{i:\text{even}} \frac{1}{i(i-1)}\frac{(q(a)-p(a))^i}{(p(a)+q(a))^{i-1}}.$$

Further,

$$p(a)\log\frac{2p(a)}{p(a)+q(a)} + q(a)\log\frac{2q(a)}{p(a)+q(a)} \ge 0.$$

For another proof along the same lines see [70, Th. 1].

Proof. Suppose first that $\operatorname{supp} p = \operatorname{supp} q = \mathcal{A}$, then

$$\begin{aligned}D\Big(p\Big|\Big|\frac{p+q}{2}\Big) &= \sum_a p(a)\log\Big(\frac{2p(a)}{p(a)+q(a)}\Big) \\ &= \sum_a p(a)\log\Big(1+\frac{p(a)-q(a)}{p(a)+q(a)}\Big) \\ &= \sum_a \Big[\frac{p(a)+q(a)}{2} + \frac{p(a)-q(a)}{2}\Big]\log\Big(1+\frac{p(a)-q(a)}{p(a)+q(a)}\Big) \\ &= \sum_a \Big[\frac{p(a)+q(a)}{2} + \frac{p(a)-q(a)}{2}\Big]\sum_{i=1}^{\infty}(-1)^{i+1}\Big(\frac{p(a)-q(a)}{p(a)+q(a)}\Big)^i\frac{1}{i} \\ &= \sum_a\sum_{i=1}^{\infty}(-1)^{i+1}\frac{1}{2i}\left(\frac{(p(a)-q(a))^i}{(p(a)+q(a))^{i-1}} + \frac{(p(a)-q(a))^{i+1}}{(p(a)+q(a))^i}\right).\end{aligned}$$

Similarly

$$D\Big(q\Big|\Big|\frac{p+q}{2}\Big) = \sum_a \sum_{i=1}^{\infty} (-1)^{i+1} \frac{1}{2i} \left(\frac{(q(a)-p(a))^i}{(p(a)+q(a))^{i-1}} + \frac{(q(a)-p(a))^{i+1}}{(p(a)+q(a))^i} \right).$$

Combining the terms and using the fact that for i odd $(x-y)^i + (y-x)^i = 0$, we get

$$\begin{aligned} D\Big(p\Big|\Big|\frac{p+q}{2}\Big) + D\Big(q\Big|\Big|\frac{p+q}{2}\Big) &= \sum_a \sum_{i:\text{odd}} \frac{1}{i} \frac{(q(a)-p(a))^{i+1}}{(p(a)+q(a))^i} - \sum_{i:\text{even}} \frac{1}{i} \frac{(q(a)-p(a))^i}{(p(a)+q(a))^{i-1}} \\ &= \sum_a \sum_{i:\text{even}} \frac{1}{i(i-1)} \frac{(q(a)-p(a))^i}{(p(a)+q(a))^{i-1}}. \end{aligned}$$

Turning to mismatched supports. Firstly whenever $p(a) > 0$ and $q(a) = 0$, by continuity conventions

$$\begin{aligned} D(p(a)||(p(a)+q(a))/2) + D(q(a)||(p(a)+q(a))/2) &= D(p(a)||p(a)/2) \\ &= p(a)\log(2) \end{aligned}$$

where $D(p(a)||q(a)) = p(a)\log(p(a)/q(a))$, but since in this case

$$\begin{aligned} \sum_{i:\text{even}} \frac{1}{i(i-1)} \frac{(q(a)-p(a))^i}{(p(a)+q(a))^{i-1}} &= p(a) \sum_{i:\text{even}} \frac{(-1)^i}{i(i-1)} \\ &= p(a)\log(2) \end{aligned}$$

the expansion is valid. An analogous argument holds for $q(a) > 0$ and $p(a) = 0$ concluding the proof. □

Lemma (10). *Let $X_{n,m}$, $1 \le m \le n$ be i.i.d. with distribution p_n on alphabet $\mathcal{A}_n$. If $|\mathcal{A}_n| = o(n)$ then for any $\epsilon > 0$*

$$p_n^n(D(\Lambda_{X^n}||p_n) > \epsilon) \le e^{-n(\epsilon - \delta_n)},$$

where $\delta_n(|\mathcal{A}_n|) \to 0$ as $n \to \infty$.

Proof.

$$\begin{aligned}
p_n^n(D(\Lambda_{X^n}||p_n) > \epsilon) &= \sum_{\substack{Q\in\mathcal{P}^n(\mathcal{A}_n):\\ D(Q||p_n)>\epsilon}} \sum_{\mathbf{x}\in T(Q)} p_n^n(\mathbf{x}) \\
&= \sum_{\substack{Q\in\mathcal{P}^n(\mathcal{A}_n):\\ D(Q||p_n)>\epsilon}} |T(Q)| e^{-n[D(Q||p_n)+H(Q)]} \\
&\leq \sum_{\substack{Q\in\mathcal{P}^n(\mathcal{A}_n):\\ D(Q||p_n)>\epsilon}} e^{-n\epsilon} \\
&\leq |\mathcal{P}^n(\mathcal{A}_n)| e^{-n\epsilon}.
\end{aligned}$$

Applying Lemma 9 gives the result. □

Lemma 34.

$$\begin{aligned}
&\sup_{j\in[0,n),k\in[0,n]} \Big| \frac{j+1}{n} \log \frac{2(j+1)}{j+1+k} - \frac{j}{n} \log \frac{2j}{j+k} \Big| \\
&\qquad \leq \frac{1}{n}(1 + \log 2 + \log(1+n)) \qquad \text{(A.7)}
\end{aligned}$$

$$\textit{and} \sup_{j\in[0,n),k\in[0,n]} \Big| \frac{k}{n} \log \frac{2k}{k+j+1} - \frac{k}{n} \log \frac{2k}{k+j} \Big| \leq \frac{1}{n}. \qquad \text{(A.8)}$$

Proof. First we prove (A.7). Suppose $j \neq 0$, then

$$\begin{aligned}
&\Big| \frac{j+1}{n} \log \frac{2(j+1)}{j+1+k} - \frac{j}{n} \log \frac{2j}{j+k} \Big| \\
&= \frac{1}{n} \Big| j \log \frac{2(j+1)}{j+1+k} \frac{j+k}{2j} + \log \frac{2(j+1)}{j+1+k} \Big| \\
&\leq \frac{1}{n} \Big(\Big| j \log \frac{j^2+jk+j+k}{(j+1+k)j} \Big| + \log 2 + \Big| \log \frac{j+1}{j+1+k} \Big| \Big) \\
&\leq \frac{1}{n} \Big(\frac{k}{(j+1+k)} + \log 2 + \log \Big(1 + \frac{k}{j+1} \Big) \Big).
\end{aligned}$$

Using the monotonicity of $\log(1+x)$ gives the bound of the lemma. For $j = 0$,

continuity gives

$$\begin{aligned}
&\Big|\frac{j+1}{n}\log\frac{2(j+1)}{j+1+k} - \frac{j}{n}\log\frac{2j}{j+k}\Big| \\
&= \frac{1}{n}\Big|\log\frac{2}{1+k}\Big| \\
&\leq \frac{1}{n}(\log 2 + \log(1+k)) \\
&\leq \frac{1}{n}(\log 2 + \log(1+n)),
\end{aligned}$$

but since the bound of the lemma is larger, we have the result. To show (A.8), observe for $k \neq 0$ we have

$$\begin{aligned}
\Big|\frac{k}{n}\log\frac{2k}{k+j+1} - \frac{k}{n}\log\frac{2k}{k+j}\Big| &= \frac{1}{n}\Big|k\log\frac{2k}{k+j+1}\frac{k+j}{2k}\Big| \\
&= \frac{1}{n}\Big|k\log\frac{k+j+1}{k+j}\Big| \\
&\leq \frac{1}{n}\frac{k}{k+j} \\
&\leq \frac{1}{n},
\end{aligned}$$

where the previous line follows from $k \leq k+j$. The case $k = 0$ is handled by continuity. □

Lemma (12)**.** *The quantity*

$$D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}} + \Lambda_{\mathbf{z}})/2)$$

viewed as a real-valued function of the vector $(\mathbf{x}, \mathbf{z}) = (x_1, \ldots, x_n, z_1, \ldots, z_n)$ *has the bounded differences property with constant*

$$\frac{2}{n}(1 + \log 2 + \log(1+n)).$$

Proof. Consider the difference

$$|D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}} + \Lambda_{\mathbf{z}})/2) - D(\Lambda_{\mathbf{x}'}||(\Lambda_{\mathbf{x}'} + \Lambda_{\mathbf{z}})/2)|$$

where $\mathbf{x}'$ is identical to $\mathbf{x}$ except for one position. Without loss of generality suppose the change from $\mathbf{x}$ to $\mathbf{x}'$ replaced an occurrence of $a \in \mathcal{A}_n$ with $b \in \mathcal{A}_n$ where $a \neq b$. It follows from the definition of relative entropy that

$$\begin{aligned}
&|D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}} + \Lambda_{\mathbf{z}})/2) - D(\Lambda_{\mathbf{x}'}||(\Lambda_{\mathbf{x}'} + \Lambda_{\mathbf{z}})/2)| \\
&\leq \Big| \frac{N(a|\mathbf{x})}{n} \log \frac{2N(a|\mathbf{x})}{N(a|\mathbf{x}) + N(a|\mathbf{z})} \\
&\qquad - \frac{N(a|\mathbf{x}')}{n} \log \frac{2N(a|\mathbf{x}')}{N(a|\mathbf{x}') + N(a|\mathbf{z})} \Big| \\
&\quad + \Big| \frac{N(b|\mathbf{x})}{n} \log \frac{2N(b|\mathbf{x})}{N(b|\mathbf{x}) + N(b|\mathbf{z})} \\
&\qquad - \frac{N(b|\mathbf{x}')}{n} \log \frac{2N(b|\mathbf{x}')}{N(b|\mathbf{x}') + N(b|\mathbf{z})} \Big|. \qquad \text{(A.9)}
\end{aligned}$$

Let

$$j + 1 = N(a|\mathbf{x}) \text{ and } k = N(a|\mathbf{z}), \text{ then } j = N(a|\mathbf{x}'),$$

then the first absolute value in the righthand side of (A.9) is of the form

$$\Big| \frac{j+1}{n} \log \frac{2(j+1)}{(j+1)+k} - \frac{j}{n} \log \frac{2j}{j+k} \Big|$$

which is bounded by $\frac{1}{n}(1 + \log 2 + \log(1+n))$ from Lemma 34. For the second summand, suppose

$$j = N(b|\mathbf{x}) \text{ and } k = N(b|\mathbf{z}), \text{then } (j+1) = N(b|\mathbf{x}'),$$

and it follows the same bound holds. Now instead consider the difference

$$|D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}} + \Lambda_{\mathbf{z}})/2) - D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}} + \Lambda_{\mathbf{z}'})/2)|$$

where $\mathbf{z}'$ is identical to $\mathbf{z}$ except for one position. Again, without loss of generality suppose that the change replaced an occurrence of $a \in \mathcal{A}_n$ with $b \in \mathcal{A}_n$ where

$a \neq b$. It follows that

$$
\begin{aligned}
&|D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}}+\Lambda_{\mathbf{z}})/2) - D(\Lambda_{\mathbf{x}}||(\Lambda_{\mathbf{x}}+\Lambda_{\mathbf{z}'})/2)| \\
&\leq \Big| N(a|\mathbf{x}) \log \frac{2N(a|\mathbf{x})}{N(a|\mathbf{x})+N(a|\mathbf{z})} \\
&\qquad - N(a|\mathbf{x}) \log \frac{2N(a|\mathbf{x})}{N(a|\mathbf{x})+N(a|\mathbf{z}')} \Big| \\
&\quad + \Big| N(b|\mathbf{x}) \log \frac{2N(b|\mathbf{x})}{N(b|\mathbf{x})+N(b|\mathbf{z})} \\
&\qquad - N(b|\mathbf{x}) \log \frac{2N(b|\mathbf{x})}{N(b|\mathbf{x})+N(b|\mathbf{z}')} \Big|. \qquad \text{(A.10)}
\end{aligned}
$$

Let

$$
j+1 = N(a|\mathbf{z}) \text{ and } k = N(a|\mathbf{x}), \text{ then } j = N(a|\mathbf{z}'),
$$

then by way of Lemma 34 the first absolute value of (A.10) is bounded by $\frac{1}{n}$. The second term is handled analogously. Since $\frac{2}{n} < \frac{2}{n}(1+\log 2+\log(1+n))$, the bounded differences property is established. □

Lemma (13). *Let $\{p_n, q_n\}$ be a sequence of pairs of distributions and denote by $\mu_n^2(x,y)$ the shadow (see [20]), i.e. distribution of the random vector $\big(np_n(X_n), nq_n(X_n)\big)$ when $X_n \sim p_n$. If $\mu_n^2(x,y)$ converges weakly to $\mu^2(x,y)$, then under hypothesis $\mathcal{H}_0$ (i.e. $Z^n \sim p_n^n$)*

$$
\begin{aligned}
\mathbb{E}[D(\Lambda_{Z^n}||\hat{p}_n)] \to \int_{C^2} \Big[&\sum_{j=1}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!} \log(2j) \\
&- \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-x)x^{k}}{k!} \log(j+k) \Big]\, d\mu^2(x,y)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[D(\Lambda_{Z^n}||\hat{q}_n)] \to \int_{C^2} \Big[&\sum_{j=1}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!} \log(2j) \\
&- \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-y)y^{k}}{k!} \log(j+k) \Big]\, d\mu^2(x,y).
\end{aligned}
$$

Proof. For notational convenience let

$$g_k^n(x) = \binom{n}{k}\Big(\frac{x}{n}\Big)^k\Big(1-\frac{x}{n}\Big)^{n-k}$$
$$\text{and } g_k(x) = \frac{x^k\exp(-x)}{k!},$$

and note for all sequences $x_n \to x$, $g_k^n(x_n) \to g_k(x)$. Now we compute

$$\begin{aligned}\mathbb{E}[D(\Lambda_{Z^n}||\hat{p}_n)] &= n^{-1}\sum_{a\in\mathcal{A}_n}\mathbb{E}[N(a|Z^n)\log 2N(a|Z^n)] \\ &\quad - \mathbb{E}[N(a|Z^n)\log(N(a|X^n)+N(a|Z^n))]. \end{aligned} \tag{A.11}$$

Starting with the second term on the righthand side (recalling the convention that $0\log 0 = 0$)

$$\begin{aligned} & n^{-1}\sum_{a\in\mathcal{A}_n}\mathbb{E}[N(a|Z^n)\log(N(a|X^n)+N(a|Z^n))] \\ &= \sum_{a\in\mathcal{A}_n}\sum_{j=1}^{n}\frac{j}{n}\binom{n}{j}p_n(a)^j(1-p_n(a))^{n-j} \\ &\quad \times\sum_{k=0}^{n}\binom{n}{k}p_n(a)^k(1-p_n(a))^{n-k}\log(j+k) \\ &= \sum_{j=1}^{n}\sum_{k=0}^{n}\left[\sum_{a\in\mathcal{A}_n}p_n(a)g_{j-1}^{n-1}((n-1)p_n(a))\times g_k^n(np_n(a))\right]\log(j+k). \end{aligned}$$

Using $\mathcal{B}(n,p)$ to denote a Binomial(n,p) random variable we have for all $n \geq \check{c}$

$$\begin{aligned} 1 &= \sum_{a\in\mathcal{A}_n} n^{-1}\mathbb{E}[\mathcal{B}(n,p_n(a))] \\ &= \sum_{a\in\mathcal{A}_n}\sum_{j=0}^{n}\frac{j}{n}\binom{n}{j}p_n(a)^j(1-p_n(a))^{n-j} \\ &= \sum_{j=1}^{n}\sum_{k=0}^{n}\sum_{a\in\mathcal{A}_n}p_n(a)g_{j-1}^{n-1}((n-1)p_n(a))g_k^n(np_n(a)) \\ &= \sum_{j=1}^{n}\sum_{k=0}^{n}\int_C g_{j-1}^{n-1}\Big(\frac{n-1}{n}x\Big)g_k^n(x)d\mu_n(x), \end{aligned}$$

where $\mu_n(\cdot) = \int_C \mu(\cdot, y)dy$. Thus it follows there exist a pair of random variables (J_n, K_n) taking values in $\{1, \ldots, n\} \times \{0, \ldots, n\}$,

$$\Pr(J_n = j, K_n = k) =$$
$$\begin{cases} \int_C g_{j-1}^{n-1}\Big(\frac{n-1}{n}x\Big) g_k^n(x) d\mu_n(x) & j \in \{1, \ldots, n\}, \\ & k \in \{0, \ldots, n\}. \\ 0 & \text{otherwise.} \end{cases}$$

Since $np_n(X_n)$ converges in distribution to W with distribution $\mu(\cdot) = \int_C \mu^2(\cdot, y)dy$, we can create a sequence of random variables $\{W_n\}$ such that $W_n =^d np_n(X_n)$ and converges to W almost surely. Then since $g_k^n(W_n) \to g_k(W)$ almost surely and g_k^n is bounded,

$$\lim_{n\to\infty} \mathbb{E}[g_{j-1}^{n-1}\Big(\frac{n-1}{n}W_n\Big) g_k^n(W_n)] = \mathbb{E}[g_{j-1}(W)g_k(W)],$$

and there are random variables (J, K) taking values in $\{1, \ldots\} \times \{0, \ldots\}$ with joint distribution so that

$$\Pr(J = j, K = k) =$$
$$\begin{cases} \mathbb{E}[g_{j-1}(W)g_k(W)] & j, k \in \{1, \ldots\} \times \{0, \ldots\} \\ 0 & \text{otherwise,} \end{cases}$$

and (J_n, K_n) converge in distribution to the pair (J, K). Now,

$$\begin{aligned} \mathbb{E}[(J_n + K_n)] &= \sum_{j=1}^{n}\sum_{k=0}^{n}(j+k)\int_C g_{j-1}^{n-1}\Big(\frac{n-1}{n}x\Big) g_k^n(x) d\mu_n(x) \\ &= \int_C \sum_{j=1}^{n}\sum_{k=0}^{n}(j+k) g_{j-1}^{n-1}\Big(\frac{n-1}{n}x\Big) g_k^n(x) d\mu_n(x) \\ &= \int_C (1 + 2x - \frac{x}{n})\, d\mu_n(x) \\ &\to 1 + \int_C 2x\, d\mu(x) \end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[J+K] &= \sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\int_C (j+k)\frac{e^{-x}x^{j-1}}{(j-1)!}\frac{e^{-x}x^{k}}{k!}\,d\mu(x)\\ &= \int_C \sum_{j=1}^{\infty}\sum_{k=0}^{\infty}(j+k)\frac{e^{-x}x^{j-1}}{(j-1)!}\frac{e^{-x}x^{k}}{k!}\,d\mu(x)\\ &= \int_C (1+2x)\,d\mu(x).\end{aligned}$$

Hence $\mathbb{E}[(J_n+K_n)] \to \mathbb{E}[J+K]$ implying that J_n+K_n is uniformly integrable. It follows that $\log(J_n+K_n)$ is uniformly integrable and by way of monotone convergence

$$\mathbb{E}[\log(J_n+K_n)] \to \mathbb{E}[\log(J+K)].$$

which gives the convergence of the second term on the right of (A.11). A similar argument applies to the first term. Therefore

$$\begin{aligned}\mathbb{E}[D(\Lambda_{Z^n}||\hat{p}_n)] \to \int_C \Big[\sum_{j=1}^{\infty}\frac{\exp(-x)x^{j-1}}{(j-1)!}\log(2j)\\ -\sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-x)x^{k}}{k!}\log(j+k)\Big]d\mu(x).\end{aligned}$$

An analogous argument establishes the second claim of the lemma. □

Lemma (14). *Let $\{p_n, q_n\}$ be a sequence of pairs of distributions and denote by $\mu_n^2(x,y)$ the shadow (see [20]), i.e. distribution of the random vector $\big(np_n(X_n), nq_n(X_n)\big)$ when $X_n \sim p_n$. If $\mu_n^2(x,y)$ converges weakly to $\mu^2(x,y)$, then under hypothesis $\mathcal{H}_0$ (i.e. $Z^n \sim p_n^n$)*

$$\mathbb{E}[\chi_2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n)] \to 2\int_{C^2}\sum_{j=1}^{\infty}\sum_{k=0}^{\infty}\frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-x)x^{k}}{k!}\frac{(j-k)}{j+k}\,d\mu^2(x,y)$$

and

$$\mathbb{E}[\chi_2(\Lambda_{Y^n}, \Lambda_{Z^n}, \mathcal{A}_n)] \to \int_{C^2} \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-y)y^{j-1}}{(j-1)!}\frac{\exp(-x)x^k}{k!}\frac{(j-k)}{j+k}\frac{y}{x}\, d\mu^2(x,y)$$
$$+ \int_{C^2} \sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \frac{\exp(-x)x^{j-1}}{(j-1)!}\frac{\exp(-y)y^{k}}{k!}\frac{(j-k)}{j+k}\, d\mu^2(x,y)$$

Proof. The proof immediately follows that of Lemma 13, once one notices that

$$\begin{aligned}
\mathbb{E}[\chi^2(\Lambda_{X^n}, \Lambda_{Z^n}, \mathcal{A}_n)] &= n^{-1}\sum_{a\in\mathcal{A}_n} \mathbb{E}\Big[\frac{(N(a|X^n) - N(a|Z^n))^2}{N(a|X^n)+N(a|Z^n)}\Big] \\
&= n^{-1}\sum_{a\in\mathcal{A}_n} \mathbb{E}\Big[\frac{N(a|X^n)(N(a|X^n) - N(a|Z^n))}{N(a|X^n)-N(a|Z^n)}\Big] \\
&\quad + \mathbb{E}\Big[\frac{N(a|Z^n)(N(a|Z^n) - N(a|X^n))}{N(a|X^n)+N(a|Z^n)}\Big] \\
&= 2\sum_{a\in\mathcal{A}_n}\sum_{j=1}^{n} \frac{j}{n}\binom{n}{j} p_n(a)^j(1-p_n(a))^{n-j} \\
&\quad \times \sum_{k=0}^{n}\binom{n}{k} p_n(a)^k(1-p_n(a))^{n-k}\frac{(j-k)}{j+k}.
\end{aligned}$$

□

Lemma 35. *For all $k \in [0,n]$ and $j \in [0,n]$*

$$\left|\frac{\left(\frac{j+1}{n}-\frac{k}{n}\right)^2}{\frac{j+1}{n}+\frac{k}{n}} - \frac{\left(\frac{j}{n}-\frac{k}{n}\right)^2}{\frac{j}{n}+\frac{k}{n}}\right| \le \frac{4}{n}$$

Proof.

$$\begin{aligned}
\left|\frac{\left(\frac{j+1}{n}-\frac{k}{n}\right)^2}{\frac{j+1}{n}+\frac{k}{n}} - \frac{\left(\frac{j}{n}-\frac{k}{n}\right)^2}{\frac{j}{n}+\frac{k}{n}}\right| &= \frac{1}{n}\Big|\frac{(j+1-k)^2}{j+1+k} - \frac{(j-k)^2}{j+k}\Big| \\
&= \frac{1}{n}\Big|\frac{((j-k)^2+2(j-k)+1)(j+k)}{(j+1+k)(j+k)} - \frac{(j-k)^2(j+1+k)}{(j+k)(j+1+k)}\Big| \\
&= \frac{1}{n}\Big|\frac{-(j-k)^2+(2(j-k)+1)(j+k)}{(j+1+k)(j+k)}\Big| \\
&\le \frac{1}{n}\Big|\frac{(j-k)^2}{(j+1+k)(j+k)} + \frac{(2j+2k+1)}{j+k+1}\Big| \\
&\le \frac{1+2+1}{n}
\end{aligned}$$

Where the final inequality uses the triangle inequality and the fact that $(j-k)^2 \leq (j+k)^2$. □

Lemma 36. *Define the sets*

$$\mathcal{Z}'_n(p,q,i) = \left\{a : p(a) = 0 \text{ and } q(a) = \frac{i}{n}\right\}$$
$$\text{and } \mathcal{Z}_n(p,q,j) = \bigcup_{i=1}^{j} \mathcal{Z}'_n(p,q,i) \cup \mathcal{Z}'_n(q,p,i).$$

For all $j \geq 1$

$$G(p,q,\mathcal{A}) \geq \log(2)\chi^2(p,q,\mathcal{Z}_n(p,q,j))$$

Proof. Note that from the proof of Lemma 8 we know that the summand in the definition of $G(p,q,\mathcal{A})$ is non-negative, therefore

$$G(p,q,\mathcal{A}) \geq G(p,q,\mathcal{Z}_n(p,q)).$$

On the set $\mathcal{Z}_n(p,q,j)$ either $q(a) = 0$ or $p(a) = 0$ and when $q(a) = 0$ we have that

$$p(a)\log\left(\frac{2p(a)}{p(a)+q(a)}\right) + q(a)\log\left(\frac{2q(a)}{p(a)+q(a)}\right) = p(a)\log(2)$$

and analogously the summand is $q(a)\log(2)$ when $p(a) = 0$. Therefore

$$\begin{aligned} G(p,q,\mathcal{Z}_n(p,q,j)) &= \sum_{a\in\mathcal{Z}_n(p,q,j)} p(a)\log\left(\frac{2p(a)}{p(a)+q(a)}\right) + q(a)\log\left(\frac{2q(a)}{p(a)+q(a)}\right) \\ &= \log(2)\sum_{a\in\mathcal{Z}_n(p,q,j)} \frac{(p(a)-q(a))^2}{p(a)+q(a)} \\ &= \log(2)\chi^2(p,q,\mathcal{Z}_n(p,q,j)) \end{aligned}$$

□

A.3 Proofs: Section 2.5

In this appendix we prove the following result.

Lemma (15). *Let $\tilde{p}_n$ and $\tilde{q}_n$ be a sequence of $\alpha = 1$ large alphabet sources, defined on alphabet $\tilde{\mathcal{A}}_n$ such that $n\|\tilde{p}_n - \tilde{q}_n\|_2^2 = \epsilon$ for every n. Denote by ω a special symbol that does not occur in any of $\tilde{\mathcal{A}}_n$ and define*

$$\mathcal{A}_n \ = \tilde{\mathcal{A}}_n \cup \{\omega\}.$$

Let δ_a denote a point-mass at a and define $p_n = \frac{1}{2}\tilde{p}_n + \frac{1}{2}\delta_\omega$ and $q_n = \frac{1}{2}\tilde{q}_n + \frac{1}{2}\delta_\omega$. Then the test

$$\|\Lambda_{X^n} - \Lambda_{Z^n}\|_2^2 \lesseqgtr \|\Lambda_{Y^n} - \Lambda_{Z^n}\|_2^2 \tag{A.12}$$

is inconsistent.

Throughout this appendix we assume the setup of Lemma 15, i.e. $X^n \sim p_n^n$, $Y^n \sim q_n^n$ and we will see it suffices to consider the case $Z^n \sim p_n^n$, i.e. hypothesis $\mathcal{H}_0$ is in effect.

We use the notation $X^{n/i}$ to mean X^n without the ith component, i.e.

$$X^{n/i} = X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n.$$

Lemma 37. *For any $i \in \{1, \ldots, n\}$*

$$N^2(a|X^n) = \mathbf{1}\{X_i = a\}(1 + 2N(a|X^{n/i})) + N^2(a|X^{n/i}).$$

Proof.

$$\begin{aligned} N^2(a|X^n) &= \Big(\mathbf{1}\{X_i = a\} + N(a|X^{n/i})\Big)^2 \\ &= \mathbf{1}\{X_i = a\} + 2N(a|X^{n/i})\mathbf{1}\{X_i = a\} + N^2(a|X^{n/i}) \\ &= \mathbf{1}\{X_i = a\}(1 + 2N(a|X^{n/i})) + N^2(a|X^{n/i}). \end{aligned}$$

□

Lemma 38.

$$\mathbb{E}[N(a|X^n)N(b|X^n)] = \begin{cases} (n^2-n)p_n(a)p_n(b) & \text{if } a \neq b \\ np_n(a) + (n^2-n)p_n^2(a) & \text{if } a = b. \end{cases}$$

Proof. The proof is similar to that of Lemma 30 and so is omitted. □

Let T denote the restriction of the L_2-norm test (A.12) to $\tilde{\mathcal{A}}_n$, i.e.

$$T(X^n, Y^n, Z^n) = \frac{1}{n^2} \sum_{a \in \tilde{\mathcal{A}}_n} N^2(a|X^n) - N^2(a|Y^n) - 2N(a|Z^n)[N(a|X^n) - N(a|Y^n)].$$

Lemma 39. *Under distribution* $P_n = p_n^n \times q_n^n \times p_n^n$

$$\mathrm{Var}[nT(X^n, Y^n, Z^n)] \to 0.$$

Proof. Recall the Efron-Stein inequality, which states that

$$\mathrm{Var}(nT) = n^2 \mathrm{Var}(T) \leq \frac{1}{2} n^2 \sum_{i=1}^{3n} \mathbb{E}[(T - \tilde{T}_i)^2]$$

where $\tilde{T}_i$ is identical to T except that the ith argument of T is replaced with an independent copy having the same distribution. Thus we now investigate what happens when we replace one of the X_i, Y_i or Z_i.

Denote by $\tilde{X}_i^n = X_1, X_2, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n$, where $\tilde{X}_i =^d X_i$. Then for $i \in \{1, \ldots, n\}$

$$\begin{aligned} T - \tilde{T}_i &= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} N^2(a|X^n) - N^2(a|\tilde{X}_i^n) - 2N(a|Z^n)(N(a|X^n) - N(a|\tilde{X}_i^n)) \\ &= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} (\mathbf{1}\{X_i = a\} - \mathbf{1}\{\tilde{X}_i = a\})(1 + 2N(a|X^{n/i}) - 2N(a|Z^n)) \end{aligned}$$

where on the previous line we used Lemma 37.

Hence for $i \in \{1, \ldots, n\}$ we have

$$\begin{aligned} n^2\mathbb{E}[(T - \tilde{T}_i)^2] &= n^{-2}\mathbb{E}\Big[\Big(\sum_{a\in\tilde{\mathcal{A}}_n}(\mathbf{1}\{X_i = a\} - \mathbf{1}\{\tilde{X}_i = a\})(1 + 2N(a|X^{n/i}) - 2N(a|Z^n))\Big)^2\Big] \\ &= n^{-2}\mathbb{E}\Big[\sum_{a\in\tilde{\mathcal{A}}_n}\sum_{b\in\tilde{\mathcal{A}}_n}(\mathbf{1}\{X_i = a\} - \mathbf{1}\{\tilde{X}_i = a\})(\mathbf{1}\{X_i = b\} - \mathbf{1}\{\tilde{X}_i = b\}) \\ &\qquad \times (1 + 2N(a|X^{n/i}) - 2N(a|Z^n))(1 + 2N(b|X^{n/i}) - 2N(b|Z^n))\Big] \end{aligned}$$

Let $S(a, b) = (1 + 2N(a|X^{n/i}) - 2N(a|Z^n))(1 + 2N(b|X^{n/i}) - 2N(b|Z^n))$, so that

$$\begin{aligned} n^2\mathbb{E}[(T - \tilde{T}_i)^2] &= n^{-2}\sum_{a\in\tilde{\mathcal{A}}_n}\sum_{b\in\tilde{\mathcal{A}}_n}\mathbb{E}[\mathbf{1}\{X_i = a\}\mathbf{1}\{X_i = b\}S(a, b)] \\ &\quad - \mathbb{E}[\mathbf{1}\{X_i = a\}\mathbf{1}\{\tilde{X}_i = b\}S(a, b)] \\ &\quad - \mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{X_i = b\}S(a, b)] + \mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{\tilde{X}_i = b\}S(a, b)]. \end{aligned}$$

Because the indicators act like selectors the above display may be written as

$$\begin{aligned} n^2\mathbb{E}[(T - \tilde{T}_i)^2] &= n^{-2}\sum_{a\in\tilde{\mathcal{A}}_n}\mathbb{E}[\mathbf{1}\{X_i = a\}S(a, a)] + \mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}S(a, a)] \\ &\quad - \sum_{a\in\tilde{\mathcal{A}}_n}\sum_{b\in\tilde{\mathcal{A}}_n}\mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{X_i = b\}S(a, b)] \\ &\quad + \mathbb{E}[\mathbf{1}\{X_i = a\}\mathbf{1}\{\tilde{X}_i = b\}S(a, b)]. \end{aligned}$$

Now because $X_i =^d \tilde{X}_i$, we may write

$$\begin{aligned} n^2\mathbb{E}[(T - \tilde{T}_i)^2] &= n^{-2}\Big[\sum_{a\in\tilde{\mathcal{A}}_n}2\mathbb{E}[\mathbf{1}\{X_i = a\}S(a, a)] - 2\sum_{a\in\tilde{\mathcal{A}}_n}\sum_{b\in\tilde{\mathcal{A}}_n}\mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{X_i = b\}S(a, b)]\Big] \\ &= n^{-2}\Big[\sum_{a\in\tilde{\mathcal{A}}_n}2\mathbb{E}[\mathbf{1}\{X_i = a\}S(a, a)] - 2\sum_{a\in\tilde{\mathcal{A}}_n}\mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{X_i = a\}S(a, a)] \\ &\quad - 2\sum_{a\neq b\in\tilde{\mathcal{A}}_n}\mathbb{E}[\mathbf{1}\{\tilde{X}_i = a\}\mathbf{1}\{X_i = b\}S(a, b)].\Big]. \end{aligned}$$

Since $S \perp\!\!\!\perp (X_i, \tilde{X}_i)$ it remains to compute $\mathbb{E}[S(a, b)]$.

Expanding S gives

$$S(a,b) = 1 + 2N(b|X^{n/i}) - 2N(b|Z^n) + 2N(a|X^{n/i}) + 4N(a|X^{n/i})N(b|X^{n/i}) \\ - 4N(a|X^{n/i})N(b|Z^n) - 2N(a|Z^n) - 4N(a|Z^n)N(b|X^{n/i}) + 4N(a|Z^n)N(b|Z^n)$$

For $a = b$ applying Lemma 38 gives

$$\begin{aligned}\mathbb{E}[S(a,b)] &= 1 + 2(n-1)p_n(a) - 2np_n(a) + 2(n-1)p_n(a) + 4(n^2-3n+2)p_n^2(a) \\ &\quad + 4(n-1)p_n(a) - 4(n-1)p_n(a)np_n(a) - 2np_n(a) \\ &\quad - 4np_n(a)(n-1)p_n(a) + 4(n^2-n)p_n^2(a) + 4np_n(a) \\ &= 1 + 8np_n(a) - 8p_n(a) - 8np_n^2(a) + 8p_n^2(a).\end{aligned}$$

Similarly for $a \neq b$ we get

$$\begin{aligned}\mathbb{E}[S(a,b)] &= 1 + 2(n-1)p_n(b) - 2np_n(b) + 2(n-1)p_n(a) + 4(n^2-3n+2)p_n(a)p_n(b) \\ &\quad - 4(n-1)p_n(a)np_n(b) - 2np_n(a) - 4np_n(a)(n-1)p_n(b) \\ &\quad + 4(n^2-n)p_n(a)p_n(b) \\ &= 1 - 2p_n(b) - 2p_n(a) - 8np_n(a)p_n(b) + 8p_n(a)p_n(b).\end{aligned}$$

Putting things together we can now evaluate to give

$$\begin{aligned}n^2\mathbb{E}[(T-\tilde{T}_i)^2] = n^{-2}\Big[&\sum_{a\in\tilde{\mathcal{A}}_n} 2p_n(a)(1 + 8np_n(a) - 8p_n(a) - 8np_n^2(a) + 8p_n^2(a)) \\ &- 2\sum_{a\in\tilde{\mathcal{A}}_n} p_n(a)p_n(a)(1 + 8np_n(a) - 8p_n(a) - 8np_n^2(a) + 8p_n^2(a)) \\ &- 2\sum_{a\neq b\in\tilde{\mathcal{A}}_n} p_n(a)p_n(b)(1 - 2p_n(b) - 2p_n(a) - 8np_n(a)p_n(b) + 8p_n(a)p_n(b))\Big].\end{aligned}$$

We can get a valid upper bound by keeping only those terms which are pos-

itive, i.e.

$$n^2\mathbb{E}[(T-\tilde{T}_i)^2] \leq n^{-2}\Big[\sum_{a\in\tilde{\mathcal{A}}_n} 2p_n(a)(1+8np_n(a)+8p_n^2(a)) \\ -2\sum_{a\in\tilde{\mathcal{A}}_n} p_n(a)p_n(a)(-8p_n(a)-8np_n^2(a)) \\ -2\sum_{a\neq b\in\tilde{\mathcal{A}}_n} p_n(a)p_n(b)(-2p_n(b)-2p_n(a)-8np_n(a)p_n(b))\Big].$$

Now summing each factor in the squares braces, and just writing the order of the resulting sum we have

$$n^2\mathbb{E}[(T-\tilde{T}_i)^2] \leq n^{-2}[O(1)+O(1)+O(n^{-2})+O(n^{-2})+O(n^{-2})+O(n^{-1})+O(n^{-1})+O(n^{-1})] \\ = O(n^{-2})$$

and therefore

$$\sum_{i=1}^{n} n^2\mathbb{E}[(T-\tilde{T}_i)^2] \leq O(n^{-1}).$$

When changing a Y_i, proceeding as before we get

$$T-\tilde{T}_{i+n} = T(X^n,Y^n,Z^n)-T(X^n,\tilde{Y}_i^n,Z^n) \\ = n^{-2}\sum_{a\in\tilde{\mathcal{A}}_n} N^2(a|\tilde{Y}_i^n)-N^2(a|Y^n)+2N(a|Z^n)[N(a|Y^n)-N(a|\tilde{Y}_i^n)] \\ = n^{-2}\sum_{a\in\tilde{\mathcal{A}}_n} [\mathbf{1}\{Y_i=a\}-\mathbf{1}\{\tilde{Y}_i=a\}](2N(a|Z^n)-1-2N(a|Y^{n/i})).$$

Now define $U(a,b) = (2N(a|Z^n)-1-2N(a|Y^{n/i}))(2N(b|Z^n)-1-2N(b|Y^{n/i}))$,

then

$$
\begin{aligned}
n^2\mathbb{E}[(T-\tilde{T}_i)^2] &= n^{-2}\sum_{a\in\tilde{\mathcal{A}}_n}\sum_{b\in\tilde{\mathcal{A}}_n}\mathbb{E}\Big[(\mathbf{1}\{Y_i=a\}-\mathbf{1}\{\tilde{Y}_i=a\})(\mathbf{1}\{Y_i=b\}-\mathbf{1}\{\tilde{Y}_i=b\})U(a,b)\Big] \\
&= n^{-2}\sum_{a\in\tilde{\mathcal{A}}_n}\sum_{b\in\tilde{\mathcal{A}}_n}\mathbb{E}[\mathbf{1}\{Y_i=a\}\mathbf{1}\{Y_i=b\}U(a,b)]-\mathbb{E}[\mathbf{1}\{Y_i=a\}\mathbf{1}\{\tilde{Y}_i=b\}U(a,b)] \\
&\quad -\mathbb{E}[\mathbf{1}\{\tilde{Y}_i=a\}\mathbf{1}\{Y_i=b\}U(a,b)]+\mathbb{E}[\mathbf{1}\{\tilde{Y}_i=a\}\mathbf{1}\{\tilde{Y}_i=b\}U(a,b)] \\
&= n^{-2}\Bigg[\sum_{a\in\tilde{\mathcal{A}}_n}2\mathbb{E}[\mathbf{1}\{Y_i=a\}U(a,a)]-\sum_{a\in\tilde{\mathcal{A}}_n}\sum_{b\in\tilde{\mathcal{A}}_n}2\mathbb{E}[\mathbf{1}\{\tilde{Y}_i=a\}\mathbf{1}\{Y_i=b\}U(a,b)]\Bigg] \\
&= n^{-2}\Bigg[\sum_{a\in\tilde{\mathcal{A}}_n}2\mathbb{E}[\mathbf{1}\{Y_i=a\}U(a,a)]-\sum_{a\in\tilde{\mathcal{A}}_n}2\mathbb{E}[\mathbf{1}\{\tilde{Y}_i=a\}\mathbf{1}\{Y_i=a\}U(a,a)] \\
&\quad -\sum_{a\neq b\in\tilde{\mathcal{A}}_n}2\mathbb{E}[\mathbf{1}\{\tilde{Y}_i=a\}\mathbf{1}\{Y_i=b\}U(a,b)]\Bigg].
\end{aligned}
$$

Computing $\mathbb{E}[U(a,b)]$ yields

$$
\begin{aligned}
\mathbb{E}[U(a,a)] &= 4np_n(a)+4(n^2-n)p_n^2(a)-2np_n(a)-4np_n(a)(n-1)q_n(a) \\
&\quad -2np_n(a)+1+2(n-1)q_n(a)-4(n-1)q_n(a)np_n(a) \\
&\quad +2(n-1)q_n(a)+4(n-1)q_n(a)+4(n-1)(n-2)q_n^2(a)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[U(a,b)] &= 4(n^2-n)p_n(a)p_n(b)-2np_n(a)-4np_n(a)(n-1)q_n(b) \\
&\quad -2np_n(b)+1+2(n-1)q_n(b)-4(n-1)q_n(a)np_n(b) \\
&\quad +2(n-1)q_n(a)+4(n-1)(n-2)q_n(a)q_n(b).
\end{aligned}
$$

For any $a,b\in\tilde{\mathcal{A}}_n$ the absolute value of every term appearing in $U(\cdot,\cdot)$ is $O(1)$, and since $U(a,b)\perp\!\!\!\perp(Y_i,\tilde{Y}_i)$ it follows that

$$
\sum_{i=1}^{n}n^2\mathbb{E}[(T-\tilde{T}_{i+n})^2]=O(n^{-1}).
$$

When replacing a Z_i, we have

$$T - \tilde{T}_{i+2n} = n^{-2} 2 \sum_{a \in \tilde{\mathcal{A}}_n} (\mathbf{1}\{\tilde{Z}_i = a\} - \mathbf{1}\{Z_i = a\})(N(a|X^n) - N(a|Y^n))$$

Thus for $i \in \{1, \dots, n\}$ we have

$$\begin{aligned} n^2 \mathbb{E}[(T - \tilde{T}_{i+2n})^2] &= n^{-2} \mathbb{E}\Big[\Big(\sum_{a \in \tilde{\mathcal{A}}_n} (\mathbf{1}\{\tilde{Z}_i = a\} - \mathbf{1}\{Z_i = a\})(N(a|X^n) - N(a|Y^n))\Big)^2\Big] \\ &= n^{-2} \sum_{a \in \tilde{\mathcal{A}}_n} \sum_{b \in \tilde{\mathcal{A}}_n} \mathbb{E}\Big[(\mathbf{1}\{Z_i = a\} - \mathbf{1}\{\tilde{Z}_i = a\})(\mathbf{1}\{Z_i = b\} - \mathbf{1}\{\tilde{Z}_i = b\})V(a,b)\Big] \end{aligned}$$

where we defined

$$\begin{aligned} V(a,b) &= (N(a|X^n) - N(a|Y^n))(N(b|X^n) - N(b|Y^n)) \\ &= N(a|X^n)N(b|X^n) - N(a|X^n)N(b|Y^n) - N(a|Y^n)N(b|X^n) + N(a|Y^n)N(b|Y^n). \end{aligned}$$

Expanding the terms and using the selection property we get

$$\begin{aligned} n^2 \mathbb{E}[(T - \tilde{T}_i)^2] &= n^{-2} \Bigg[\sum_{a,b \in \tilde{\mathcal{A}}_n} \mathbb{E}[\mathbf{1}\{Z_i = a\}\mathbf{1}\{Z_i = b\}V(a,b)] - \mathbb{E}[\mathbf{1}\{Z_i = a\}\mathbf{1}\{\tilde{Z}_i = b\}V(a,b)] \\ &\qquad - \mathbb{E}[\mathbf{1}\{\tilde{Z}_i = a\}\mathbf{1}\{Z_i = b\}V(a,b)] + \mathbb{E}[\mathbf{1}\{\tilde{Z}_i = a\}\mathbf{1}\{\tilde{Z}_i = b\}V(a,b)] \Bigg] \\ &= n^{-2} \Bigg[2 \sum_{a \in \tilde{\mathcal{A}}_n} \mathbb{E}[\mathbf{1}\{Z_i = a\}V(a,a)] - 2 \sum_{a,b \in \tilde{\mathcal{A}}_n} \mathbb{E}[\mathbf{1}\{Z_i = a\}\mathbf{1}\{\tilde{Z}_i = b\}V(a,b)] \Bigg] \end{aligned}$$

On account of the independence of $(Z_i, \tilde{Z}_i)$ and $V(\cdot,\cdot)$ it remains to compute $\mathbb{E}[V(a,b)]$, yielding

$$\begin{aligned} \mathbb{E}[V(a,a)] &= np_n(a) + (n^2 - n)p_n^2(a) - n^2 p_n(a)q_n(a) \\ &\qquad - n^2 p_n(a)q_n(a) + nq_n(a) + (n^2 - n)q_n^2(a) \end{aligned}$$

and for $a \neq b$

$$\begin{aligned} \mathbb{E}[V(a,b)] &= (n^2 - n)p_n(a)p_n(b) - n^2 p_n(a)q_n(b) \\ &\qquad - n^2 p_n(b)q_n(a) + (n^2 - n)q_n(a)q_n(b). \end{aligned}$$

Each term appearing in $V(\cdot,\cdot)$ has absolute value $O(1)$ and so it follows that

$$\sum_{i=1}^{n} n^2 \mathbb{E}[(T - \tilde{T}_{i+2n})^2] = O(n^{-1}).$$

Therefore we have shown

$$\mathrm{Var}(nT) \leq O(n^{-1}) \to 0.$$

□

Proof of Lemma 15. Suppose hypothesis $\mathcal{H}_0$ is in effect. Chebyshev's inequality combined with Lemma 39 imply that

$$n\Big[\sum_{a\in\tilde{\mathcal{A}}_n} (\Lambda_{X^n}(a) - \Lambda_{Z^n}(a))^2 - (\Lambda_{Y^n}(a) - \Lambda_{Z^n}(a))^2\Big]$$

is close to its mean with high probability. Thus, using $\to_{P_n}$ to denote convergence in probability, we have

$$n\sum_{a\in\tilde{\mathcal{A}}_n} (\Lambda_{X^n}(a) - \Lambda_{Z^n}(a))^2 - (\Lambda_{Y^n}(a) - \Lambda_{Z^n}(a))^2 \to_{P_n} -\epsilon/4.$$

Next we note that by the Central Limit Theorem,

$$2\sqrt{n}\Big(\Lambda_{X^n}(\omega) - \frac{1}{2}\Big) = 2\sqrt{n}\Big(\sum_{i=1}^{n} \frac{\mathbf{1}(X_i = \omega)}{n} - \frac{1}{2}\Big) \Rightarrow \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ denotes a standard Normal random variable. Similarly $2\sqrt{n}(\Lambda_{Y^n}(\omega) - 1/2) \Rightarrow \mathcal{N}(0,1)$ and $2\sqrt{n}(\Lambda_{Z^n}(\omega) - 1/2) \Rightarrow \mathcal{N}(0,1)$. Furthermore the independence of the X^n, Y^n, Z^n sequences implies the independence of the limiting distributions. Let $\tilde{X}, \tilde{Y}, \tilde{Z}$ be independent $\mathcal{N}(0,1)$. Now by the continuous mapping theorem [90, Ch.1 §7] it follows that

$$\begin{aligned} 4n\Big[(\Lambda_{X^n}(\omega) - \Lambda_{Z^n}(\omega))^2 - (\Lambda_{Y^n}(\omega) - \Lambda_{Z^n}(\omega))^2\Big] &\Rightarrow \tilde{X}^2 + \tilde{Z}^2 - 2\tilde{X}\tilde{Z} - \tilde{Y}^2 - \tilde{Z}^2 + 2\tilde{Y}\tilde{Z} \\ &= \tilde{X}^2 - \tilde{Y}^2 - 2\tilde{Z}(\tilde{X} - \tilde{Y}). \end{aligned}$$

Finally, Slutsky's theorem [90, Ch.1 §5.4] tells us that if $X_n \Rightarrow X$ and $Y_n \to_P c$ then $X_n + Y_n \Rightarrow X + c$, therefore

$$4n\left[\|\Lambda_{X^n} - \Lambda_{Z^n}\|_2^2 - \|\Lambda_{Y^n} - \Lambda_{Z^n}\|_2^2\right] \Rightarrow \tilde{X}^2 - \tilde{Y}^2 - 2\tilde{Z}(\tilde{X} - \tilde{Y}) - \epsilon.$$

This random variable has positive probability of being positive, and thus the test is inconsistent. □

APPENDIX B

CHAPTER 4 - PROOFS

B.1 Proof of Theorem 15

If $R_1 \geq \log|\mathcal{X}_1|$, then clearly $\eta(P_{XY}, R_1, R_2) = \infty$ and the result trivially holds, so suppose that $R_1 < \log|\mathcal{X}_1|$.

B.1.1 Scheme

We start by describing a scheme and then show the scheme has the performance specified in the theorem. Let $\epsilon > 0$ be given. For a given blocklength n, we operate on a type-by-type basis and define the encoding and decoding functions as follows.

Encoder 1: For each type-class $T^n_{Q_X}$, the encoder and decoder agree on a random binning scheme. In particular, for every sequence in $T^n_{Q_X}$, a bin index is assigned uniformly at random from $\{1, 2, \ldots, \exp(nR_1)\}$. To encode a sequence $\mathbf{x}$, the encoder sends the type $Q_{\mathbf{x}}$ and its bin index, $U_1(\cdot)$. Mathematically $f_1^n : \mathcal{X}^n \to \mathcal{M}_1$ is

$$f_1^n(\mathbf{x}) = (k(Q_{\mathbf{x}}), U_1(\mathbf{x})),$$

where

$$\begin{aligned}
\mathcal{M}_1 &= \mathcal{M}_1' \times \mathcal{M}_1'', \\
\mathcal{M}_1' &= \{1, 2, \ldots, M_1 \triangleq \exp(nR_1)\}, \\
\mathcal{M}_1'' &= \{1, 2, \ldots, (n+1)^{|\mathcal{X}|}\}.
\end{aligned}$$

Encoder 2: For each type Q_Y, fix a conditional type $Q^*_{S|Y}(Q_Y) \in \mathcal{C}^n(Q_Y, \mathcal{S})$ so that $I(Q_Y; Q^*_{S|Y}(Q_Y)) \leq R_2$ and randomly choose a set of codewords $B^n(Q_Y)$ in the following way. The size of $B^n(Q_Y)$ is an integer satisfying

$$\begin{aligned} &\exp(nI(Q_Y; Q^*_{S|Y}(Q_Y)) + (|\mathcal{Y}||\mathcal{S}| + 2)\log(n+1)) \\ &\leq |B^n(Q_Y)| \\ &\leq \exp(nI(Q_Y; Q^*_{S|Y}(Q_Y)) + (|\mathcal{Y}||\mathcal{S}| + 4)\log(n+1)) \end{aligned} \tag{B.1}$$

and the codewords are drawn uniformly, with replacement, from the marginal type class $T^n_{Q^*_S}$ induced by Q_Y and $Q^*_{S|Y}(Q_Y)$. Define $S : T^n_{Q_\mathbf{y}} \to B^n(Q_\mathbf{y})$ as follows. Let $\mathcal{G}(\mathbf{y}) = B^n(Q_\mathbf{y}) \cap T^n_{Q^*_{S|Y}(Q_\mathbf{y})}(\mathbf{y})$, if $\mathcal{G}(\mathbf{y})$ is non-empty, then the output of $S(\mathbf{y})$ is drawn uniformly at random from $\mathcal{G}(\mathbf{y})$. If $\mathcal{G}(\mathbf{y})$ is empty the output of $S(\mathbf{y})$ is drawn uniformly at random from $B^n(Q_\mathbf{y})$. The function $S(\cdot)$ determines the codeword sent by the helper encoder to the decoder. We define $S^n = S(Y^n)$. To encode a sequence $\mathbf{y} \in T^n_{Q_X}$, the encoder sends the type of $\mathbf{y}$ and the index, $U_2(S(\mathbf{y}))$, of the codeword $S(\mathbf{y})$. Mathematically the second encoder, $f^n_2 : \mathcal{Y}^n \to \mathcal{M}_2$ operates as follows

$$f^n_2(\mathbf{y}) = (k(Q_\mathbf{y}), U_2(S(\mathbf{y})))$$

where

$$\begin{aligned} \mathcal{M}_2 &= \mathcal{M}'_2 \times \mathcal{M}''_2, \\ \mathcal{M}'_2 &= \{1, 2, \ldots, M_2 \triangleq \exp(n(R_2 + \epsilon/2))\}, \\ \mathcal{M}''_2 &= \{1, 2, \ldots, (n+1)^{|\mathcal{Y}|}\}. \end{aligned}$$

Decoder:

1. If $\log |T^n_{Q_\mathbf{x}}| \leq nR_1$ then $\mathbf{x}$ can be decoded without error;

2. If $\log |T^n_{Q_{\mathbf{x}}}| > nR_1$ the decoder receives a bin index from encoder one and uses the coded side information from encoder two to pick the "best" $\mathbf{x}$ from the bin in the minimum conditional entropy sense: it searches for a $\hat{\mathbf{x}}$ in the received bin so that among all $\tilde{\mathbf{x}}$ in the bin, $H(\tilde{\mathbf{x}}|\mathbf{s}) > H(\hat{\mathbf{x}}|\mathbf{s})$. If there is no such $\hat{\mathbf{x}}$ it picks uniformly at random from the bin.

$$g^n(k(Q_X), i, k(Q_Y), \mathbf{s}) = \begin{cases} \hat{\mathbf{x}} & \text{if } U_1(\hat{\mathbf{x}}) = i \text{ and } \forall \tilde{\mathbf{x}} \neq \hat{\mathbf{x}}, U_1(\tilde{\mathbf{x}}) = i : \\ & \quad H(\tilde{\mathbf{x}}|\mathbf{s}) > H(\hat{\mathbf{x}}|\mathbf{s}) \\ \text{any } \tilde{\mathbf{x}} \text{ with } U_1(\tilde{\mathbf{x}}) = i & \text{if no such } \hat{\mathbf{x}} \end{cases}$$

B.1.2 Error Probability Calculation

We define the following sets

$$\mathcal{E}_r = \{(\mathbf{x}, \mathbf{y}, \mathbf{s}) : H(Q_{\mathbf{x}}) > R_1)\}, \mathcal{D}_r = \{Q_{XYS} : H(Q_X) > R_1\},$$
$$\mathcal{E}_c = \{(\mathbf{x}, \mathbf{y}, \mathbf{s}) : \mathbf{s} \notin T^n_{Q^*_{S|Y}(Q_{\mathbf{y}})}(\mathbf{y}))\}, \mathcal{D}_c = \{Q_{XYS} : Q_{S|Y} \neq Q^*_{S|Y}(Q_Y)\},$$

and let F denote the event that there exists some $\tilde{\mathbf{s}} \in B^n(Q_Y)$ with $\tilde{\mathbf{s}} \in T^n_{Q_{S|Y}(Q_{\mathbf{y}})^*}(\mathbf{y})$.

The following lemmas will be useful.

Lemma 40. *Let $X^n, Y^n, S^n = S(Y^n)$ be generated according to our scheme and suppose that $(\mathbf{x}, \mathbf{y}, \mathbf{s})$ is in $(\mathcal{E}_c)^c$, i.e. that $\mathbf{s} \in T^n_{Q^*_{S|Y}(Q_{\mathbf{y}})}(\mathbf{y})$. Then*

$$\Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \tag{B.2}$$
$$\leq P^n_{XY}(\mathbf{x}, \mathbf{y}) \frac{1}{|T^n_{Q^*_{S|Y}(Q_{\mathbf{y}})}(\mathbf{y})|}. \tag{B.3}$$

Proof. For the $\mathbf{x}, \mathbf{y}, \mathbf{s}$ in this lemma, $\{X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}\}$ implies that the event F has occurred. Thus

$$\begin{aligned}
&\Pr(X^n = \mathbf{x},\ Y^n = \mathbf{y}, S^n = \mathbf{s}) \\
&\quad = \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}, F) \\
&\quad = P_{XY}^n(\mathbf{x}, \mathbf{y}) \Pr(F | X^n = \mathbf{x}, Y^n = \mathbf{y}) \\
&\qquad \times \Pr(S^n = \mathbf{s} | X^n = \mathbf{x}, Y^n = \mathbf{y}, F) \\
&\quad \leq P_{XY}^n(\mathbf{x}, \mathbf{y}) \Pr(S^n = \mathbf{s} | X^n = \mathbf{x}, Y^n = \mathbf{y}, F) \\
&\quad = P_{XY}^n(\mathbf{x}, \mathbf{y}) \frac{1}{|T^n_{Q^*_{S|Y}(Q_\mathbf{y})}(\mathbf{y})|}
\end{aligned}$$

where in the final line we used that conditional on F, S^n is uniformly distributed over $T^n_{Q_{S|Y}(Q_\mathbf{y})}(\mathbf{y})$. □

Lemma 41. *Let $X^n, Y^n, S^n = S(Y^n)$ be generated according to our scheme and suppose that $(\mathbf{x}, \mathbf{y}, \mathbf{s}) \in \mathcal{E}_c$. Then*

$$\begin{aligned}
&\Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \\
&\qquad \leq \exp(-(n+1)^2). \qquad \text{(B.4)}
\end{aligned}$$

Proof. For the $\mathbf{x}, \mathbf{y}, \mathbf{s}$ in this lemma, $\{X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}\}$ implies that event F^c has occurred. Thus

$$\begin{aligned}
&\Pr(X^n = \mathbf{x},\ Y^n = \mathbf{y}, S^n = \mathbf{s}) \\
&\quad = \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}, F^c) \\
&\quad = P_Y^n(\mathbf{y}) \Pr(F^c | Y^n = \mathbf{y}) \Pr(X^n = \mathbf{x} | Y^n = \mathbf{y}, F^c) \\
&\qquad \times \Pr(S^n = \mathbf{s} | X^n = \mathbf{x}, Y^n = \mathbf{y}, F^c) \\
&\quad \leq \Pr(F^c | Y^n = \mathbf{y}).
\end{aligned}$$

$\Pr(F^c | Y^n = \mathbf{y})$ is the probability that there is no $\tilde{\mathbf{s}} \in B^n(Q_\mathbf{y})$ so that $\tilde{\mathbf{s}} \in T^n_{Q^*_{S|Y}(Q_\mathbf{y})}(\mathbf{y})$. We will now give an upper bound on this probability using the

properties of the codeword set. Let $m = |B^n(Q_{\mathbf{y}})|$ and $B^n(Q_{\mathbf{y}})[i]$ be the ith codeword in the set $B^n(Q_{\mathbf{y}})$. Then

$$\begin{aligned}
&\Pr(F^c|Y^n = \mathbf{y}) \\
&\qquad = \prod_{i=1}^{m} \Pr(B^n(Q_{\mathbf{y}})[i] \notin T^n_{Q^*_{S|Y}(Q_{\mathbf{y}})}(\mathbf{y})) \\
&\qquad = \prod_{i=1}^{m} [1 - \Pr(B^n(Q_{\mathbf{y}})[i] \in T^n_{Q^*_{S|Y}(Q_{\mathbf{y}})}(\mathbf{y}))] \\
&\qquad = \left(1 - \frac{|T^n_{Q^*_{S|Y}(Q_{\mathbf{y}})}(\mathbf{y})|}{|T_{Q^*_S}|}\right)^m \\
&\qquad \leq \exp\left(-\frac{|T^n_{Q^*_{S|Y}(Q_{\mathbf{y}})}(\mathbf{y})|}{|T_{Q^*_S}|} m\right)
\end{aligned}$$

where the last line followed by applying the inequality $(1-t)^m \leq \exp(-tm)$. Next, using the following bounds on the cardinality of type classes [27, lemmas 2.3 and 2.6],

$$\begin{aligned}
|T^n_{Q_S}| &\leq \exp(nH(Q_S)) \\
|T^n_{Q_{S|Y}}(\mathbf{y})| &\geq (n+1)^{-|\mathcal{Y}||\mathcal{S}|} \exp(nH(Q_{S|Y}|Q_Y))
\end{aligned}$$

and that $I(Q^*_{S|Y}(Q_{\mathbf{y}}); Q_{\mathbf{y}}) = H(Q^*_S) - H(Q^*_{S|Y}(Q_{\mathbf{y}})|Q_{\mathbf{y}})$ we have

$$-\frac{|T^n_{Q^*_{S|Y}(Q_{\mathbf{y}})}(\mathbf{y})|}{|T_{Q^*_S}|} \leq -(n+1)^{-|\mathcal{Y}||\mathcal{S}|} \exp(-nI(Q_{\mathbf{y}}; Q^*_{S|Y}(Q_{\mathbf{y}}))).$$

Thus,

$$\begin{aligned}
&\Pr(F^c|Y^n = \mathbf{y}) \\
&\qquad \leq \exp\left(-(n+1)^{-|\mathcal{Y}||\mathcal{S}|} \exp(-nI(Q_{\mathbf{y}}; Q^*_{S|Y}(Q_{\mathbf{y}})))m\right) \\
&\qquad \leq \exp(-(n+1)^2)
\end{aligned}$$

where the final line followed by substitution our choice of m from (B.1). □

Lemma 42. *For all strings* $\mathbf{x}, \mathbf{y}$*, let*

$$S(\mathbf{x}|\mathbf{y}) = \{\tilde{\mathbf{x}} | H(\tilde{\mathbf{x}}|\mathbf{y}) \leq H(\mathbf{x}|\mathbf{y}), Q_{\tilde{\mathbf{x}}} = Q_{\mathbf{x}}\}.$$

Then

$$|S(\mathbf{x}|\mathbf{y})| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(nH(\mathbf{x}|\mathbf{y})).$$

Proof.

$$\begin{aligned}
|S(\mathbf{x}|\mathbf{y})| &\leq |\{\tilde{\mathbf{x}}|H(\tilde{\mathbf{x}}|\mathbf{y}) \leq H(\mathbf{x}|\mathbf{y})\}| \\
&= \sum_{V:V\in\mathcal{C}^n(Q_{\mathbf{y}},\mathcal{X})} \sum_{\tilde{\mathbf{x}}\in T_V(\mathbf{y}):H(\tilde{\mathbf{x}}|\mathbf{y})\leq H(\mathbf{x}|\mathbf{y})} 1 \\
&= \sum_{\substack{V:V\in\mathcal{C}^n(Q_{\mathbf{y}},\mathcal{X}) \\ H(V|Q_{\mathbf{y}})\leq H(\mathbf{x}|\mathbf{y})}} |T_V(\mathbf{y})| \\
&\leq \sum_{\substack{V:V\in\mathcal{C}^n(Q_{\mathbf{y}},\mathcal{X}) \\ H(V|Q_{\mathbf{y}})\leq H(\mathbf{x}|\mathbf{y})}} \exp(nH(\mathbf{x}|\mathbf{y})) \\
&\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(nH(\mathbf{x}|\mathbf{y})).
\end{aligned}$$

□

Lemma 43. *Let* $(\mathbf{x},\mathbf{y},\mathbf{s}) \in \mathcal{E}_r \cap (\mathcal{E}_c)^c$. *Then*

$$\begin{aligned}
&\Pr(X^n \neq \hat{X}^n | X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \\
&\quad \leq \exp\left(-n\left(R - H(Q_{\mathbf{x}|\mathbf{s}}|Q_{\mathbf{s}}) - \delta_n\right)^+\right) \qquad \text{(B.5)}
\end{aligned}$$

where

$$\delta_n = \frac{1}{n}\log(n+1)^{|\mathcal{S}||\mathcal{X}|}.$$

Proof. The decoder makes an error if it selects the wrong source sequence from the bin. We note that the set $S(\mathbf{x}|\mathbf{s})$ of Lemma 42 contains all the sequences with lower empirical entropy, but having the same type as $\mathbf{x}$. Therefore we can

bound the decoding error probability as

$$\begin{aligned}
&\Pr(X^n \neq \hat{X}^n | X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \\
\leq &\sum_{\tilde{\mathbf{x}} \in S(\mathbf{x}|\mathbf{s})} \Pr(U(\tilde{\mathbf{x}}) = U(\mathbf{x})) \\
\leq &|S(\mathbf{x}|\mathbf{s})| \exp(-nR_1) \\
\leq &\exp(-n(R_1 - H(Q_{\mathbf{x}|\mathbf{s}}|Q_{\mathbf{s}}) - \delta_n))
\end{aligned}$$

where the final line used the result from Lemma 42. Further bounding the probability by one gives the result. □

Lemma 44. *Let*

$$F^n(P_{XY}, R_1, R_2) = \min_{Q_Y} \max_{\substack{Q_{S|Y} \in \mathcal{C}^n(Q_Y, \mathcal{S}): \\ I(Q_Y; Q_{S|Y}) \leq R_2}} \min_{\substack{Q_{XYS} \\ H(Q_X) \geq R_1}} D(Q_{XYS} || P_{XY} Q_{Y|S}) + [R_1 - H(Q_{X|S}|Q_S) - \delta_n]^+ \tag{B.6}$$

and

$$F^\infty(P_{XY}, R_1, R_2) = \inf_{Q_Y} \sup_{\substack{Q_{S|Y} \in \mathcal{C}(\mathcal{Y} \to \mathcal{S}): \\ I(Q_Y; Q_{S|Y}) \leq R_2}} \inf_{\substack{Q_{XYS} \\ H(Q_X) \geq R_1}} D(Q_{XYS} || P_{XY} Q_{Y|S}) + [R_1 - H(Q_{X|S}|Q_S)]^+. \tag{B.7}$$

Then

$$\liminf_{n \to \infty} F^n(P_{XY}, R_1, R_2) \geq F^\infty(P_{XY}, R_1, R_2).$$

Proof. A more intricate result is proved in Lemma 51 for the discrete memoryless Wyner-Ziv problem. The approach taken there is applicable here. □

Proof of Theorem 15. To prove the theorem we will upper bound $P_e = \Pr(X^n \neq \hat{X}^n)$, the probability of error for our scheme. For any $\epsilon > 0$, we note that for n sufficiently large the constraints in (4.2) are met. On $(\mathcal{E}_r)^c$ our scheme makes no

error, thus

$$
\begin{aligned}
P_e &= \sum_{\mathcal{E}_r} \Pr(X^n \neq \hat{X}^n | X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \\
&= \sum_{\mathcal{E}_r \cap (\mathcal{E}_c)^c} \Pr(X^n \neq \hat{X}^n | X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \\
&\quad + \sum_{\mathcal{E}_r \cap \mathcal{E}_c} \Pr(X^n \neq \hat{X}^n | X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \\
&\leq \sum_{\mathcal{E}_r \cap (\mathcal{E}_c)^c} \Pr(X^n \neq \hat{X}^n | X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s}) \\
&\quad + \sum_{\mathcal{E}_c} \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, S^n = \mathbf{s})
\end{aligned}
$$

where the final inequality follows by bounding the conditional error probability by 1 on $\mathcal{E}_c$. Applying Lemmas 40 and 43 to the summation over $\mathcal{E}_r \cap (\mathcal{E}_c)^c$, and Lemma 41 to summation over $\mathcal{E}_c$ we get

$$
\begin{aligned}
P_e \leq & \sum_{\mathcal{E}_r \cap (\mathcal{E}_c)^c} \exp(-n[R_1 - H(Q_{\mathbf{x}|\mathbf{s}}|Q_{\mathbf{s}}) - \delta_n]^+) \frac{P_{XY}^n(\mathbf{x}, \mathbf{y})}{|T^n_{Q^*_{S|Y}(Q_{\mathbf{y}})}(\mathbf{y})|} \\
& + \sum_{\mathcal{E}_c} \exp(-(n+1)^2).
\end{aligned}
$$

Now summing first over types and then over sequences within the type class, we get

$$
\begin{aligned}
P_e \leq \sum_{Q_Y} \Big[& \sum_{Q_{XYS} \in \mathcal{D}_r \cap (\mathcal{D}_c)^c} \sum_{(\mathbf{x},\mathbf{y},\mathbf{s}) \in T^n_{Q_{XYS}}} \exp(-n[R_1 - H(Q_{\mathbf{x}|\mathbf{s}}|Q_{\mathbf{s}}) - \delta_n]^+) \frac{P_{XY}^n(\mathbf{x}, \mathbf{y})}{|T^n_{Q^*_{S|Y}(Q_{\mathbf{y}})}(\mathbf{y})|} \\
& + \sum_{Q_{XYS} \in \mathcal{D}_c} \sum_{(\mathbf{x},\mathbf{y},\mathbf{s}) \in T^n_{Q_{XYS}}} \exp(-(n+1)^2) \Big], \qquad \text{(B.8)}
\end{aligned}
$$

where in the summation over joint types Q_{XYS}, the marginal type of Y is fixed to be that set by the earlier summation. Using the following facts

$$
\begin{aligned}
P_{XY}^n(\mathbf{x}, \mathbf{y}) &= \exp(-n(D(Q_{\mathbf{xy}}||P_{XY}) + H(Q_{\mathbf{xy}}))) \\
|T^n_{Q_{XYS}}| &\leq \exp(n(H(Q_{XYS}))) \leq \exp(n \log(|\mathcal{X}||\mathcal{Y}||\mathcal{S}|)) \qquad \text{(B.9)} \\
|T^n_{Q_{S|Y}}| &\geq (n+1)^{-|\mathcal{Y}||\mathcal{S}|} \exp(n(H(Q_{S|Y}|Q_Y))) \qquad \text{(B.10)}
\end{aligned}
$$

and continuing from (B.8), we can further bound P_e as follows

$$P_e \leq \sum_{Q_Y} \Bigg[\sum_{Q_{XYS} \in \mathcal{D}_r \cap (\mathcal{D}_c)^c} \exp\Big(-n\Big([R_1 - H(Q_{X|S}|Q_S) - \delta_n]^+ + D(Q_{XY}||P_{XY}) + H(Q_{XY}) + H(Q_{S|Y}|Q_Y) - H(Q_{XYS}) \Big) \Big) + \sum_{Q_{XYS} \in \mathcal{D}_c} \exp\big(-(n+1)^2 + n \log(|\mathcal{X}||\mathcal{Y}||\mathcal{S}|) \big) \Bigg]. \tag{B.11}$$

Next we note that

$$\begin{aligned} & D(Q_{XY}||P_{XY}) + H(Q_{XY}) + H(Q_{S|Y}|Q_Y) - H(Q_{XYS}) \\ =& D(Q_{XY}||P_{XY}) + H(Q_{S|Y}|Q_Y) - H(Q_{S|XY}|Q_{XY}) \\ =& D(Q_{XYS}|P_{XY}Q_{S|Y}), \end{aligned}$$

and substituting this identity into (B.11) gives

$$P_e \leq \sum_{Q_Y} \Bigg[\sum_{Q_{XYS} \in \mathcal{D}_r \cap (\mathcal{D}_c)^c} \exp\Big(-n\Big([R_1 - H(Q_{X|S}|Q_S) - \delta_n]^+ + D(Q_{XYS}||P_{XY}Q_{S|Y}) \Big) \Big) + \sum_{Q_{XYS} \in \mathcal{D}_c} \exp\big(-(n+1)^2 + n \log(|\mathcal{X}||\mathcal{Y}||\mathcal{S}|) \big) \Bigg].$$

We now upper bound the summations by maximizing over the types and optimizing over the choice of test channel $Q_{S|Y}$. This gives

$$\begin{aligned} P_e \leq |\mathcal{P}^n(\mathcal{X} \times \mathcal{Y} \times \mathcal{S})||\mathcal{P}^n(\mathcal{Y})| \max_{Q_Y} \min_{\substack{Q_{S|Y} \in \mathcal{C}^n(Q_Y, \mathcal{S}): \\ I(Q_Y; Q_{S|Y}) \leq R_2}} \max_{Q_{XYS} \in \mathcal{D}_r \cap (\mathcal{D}_c)^c} & \\ \exp\Big(-n\Big([R_1 - H(Q_{X|S}|Q_S) - \delta_n]^+ & \\ + D(Q_{XYS}||P_{XY}Q_{S|Y}) \Big) \Big) & \\ + \exp\big(-(n+1)^2 + n \log(|\mathcal{X}||\mathcal{Y}||\mathcal{S}|) \big). & \end{aligned} \tag{B.12}$$

Let F^n be as defined in (B.6). We may move the optimizations appearing in

(B.12) into the exponent and this yields

$$P_e \leq |\mathcal{P}^n(\mathcal{X}\times\mathcal{Y}\times\mathcal{S})||\mathcal{P}^n(\mathcal{Y})|\Big[\exp(-n(F^n(P_{XY},R_1,R_2)))$$
$$+\exp\big(-(n+1)^2+n\log(|\mathcal{X}||\mathcal{Y}||\mathcal{S}|)\big)\Big].$$

Then we have

$$\begin{aligned}\liminf_{n\to\infty}-\frac{1}{n}\log P_e &\geq \liminf_{n\to\infty}-\frac{1}{n}\log\Big(|\mathcal{P}^n(\mathcal{X}\times\mathcal{Y}\times\mathcal{S})||\mathcal{P}^n(\mathcal{Y})|\Big[\exp(-n(F^n(P_{XY},R_1,R_2)))\\ &\qquad+\exp\big(-(n+1)^2+n\log(|\mathcal{X}||\mathcal{Y}||\mathcal{S}|)\big)\Big]\Big)\\ &\geq \liminf_{n\to\infty}F^n(P_{XY},R_1,R_2)\\ &\geq F^\infty(P_{XY},R_1,R_2)\end{aligned}$$

where the final line followed by an application of Lemma 44. □

B.2 Proof of Theorem 16

Before proving Theorem 16, we prove two technical lemmas. We first prove the cardinality bound on S given in (4.6). This argument differs from conventional cardinality-bound proofs in that it uses the KKT conditions in addition to Carathéodory's theorem. We then prove a continuity lemma that is similar to Lemma 44. For the purposes of these lemmas define two new quantities

$$\tilde{\eta}_U(P_{XY},R_1,R_2)\triangleq\inf_{Q_Y}\sup_{\substack{Q_{S|Y}:|\mathcal{S}|\leq|\mathcal{X}|\cdot|\mathcal{Y}|+|\mathcal{Y}|+2\\ I(Q_Y;Q_{S|Y})\leq R_2}}\inf_{\substack{Q_{X|Y}:\\ H(Q_{X|S}|Q_S)\geq R_1}}D(Q_{XY}||P_{XY})$$

$$\text{and }\overline{\eta}_U(P_{XY},R_1,R_2)\triangleq\inf_{Q_Y}\sup_{\substack{Q_{S|Y}:\\ I(Q_Y;Q_{S|Y})\leq R_2}}\inf_{\substack{Q_{X|Y}:\\ H(Q_{X|S}|Q_S)\geq R_1}}D(Q_{XY}||P_{XY}).$$

Note that $\tilde{\eta}_U$ differs from η_U only in that the inequality in the inner-most infimum is no longer strict, and $\overline{\eta}_U$ differs from $\tilde{\eta}_U$ only in the omission of the

cardinality bound on S. Since for $R_1 \geq \log|\mathcal{X}_1|$, $\eta_U(P_{XY}, R_1, R_2) = \infty$ and Theorem 16 is trivial, we assume throughout this appendix that $R_1 < \log|\mathcal{X}_1|$.

Lemma 45. *If $R_1 < \log|\mathcal{X}_1|$ and $P_{XY}(x,y) > 0$ for all x and y, then $\tilde{\eta}_U = \overline{\eta}_U$.*

Proof. Clearly $\overline{\eta}_U \geq \tilde{\eta}_U$. To show the reverse inequality, it suffices to show that for all Q_Y and all $Q_{S|Y}$ such that $I(Q_Y; Q_{S|Y}) \leq R_2$, there exists $\tilde{Q}_{S|Y}$ such that

1. $I(Q_Y, \tilde{Q}_{S|Y}) \leq R_2$
2. $|\mathcal{S}| \leq |\mathcal{X}| \cdot |\mathcal{Y}| + |\mathcal{Y}| + 2$
3. $\gamma(Q_Y, Q_{S|Y}) \leq \gamma(Q_Y, \tilde{Q}_{S|Y})$,

where

$$\gamma(Q_Y, Q_{S|Y}) = \inf_{\substack{Q_{X|Y}: \\ H(Q_{X|S}|Q_S) \geq R_1}} D(Q_{XY}||P_{XY}).$$

Fix Q_Y and $Q_{S|Y}$. For the P_{XY} of the hypothesis, $\gamma(Q_Y, \cdot)$ has a continuous objective and a compact feasible set, so there exists $Q^*_{X|Y}$ such that

$$\gamma(Q_Y, Q_{S|Y}) = D(Q_Y Q^*_{X|Y}||P_{XY})$$

and $H(Q^*_{X|S}|Q_S) \geq R_1$. Since $\gamma(\cdot,\cdot)$ is convex in $Q_{X|Y}$ and strictly feasible, $Q^*_{X|Y}$ must satisfy the KKT conditions for optimality [91, p.g. 243]: there exists[1]

$$\mu_{x,y} \geq 0 \quad \text{for all } x, y$$

$$\lambda \geq 0$$

$$\nu_y \geq 0 \quad \text{for all } y$$

[1]The assumption that $P_{XY}(x,y) > 0$ for all x and y guarantees that $D(Q_Y Q^*_{X|Y}||P_{XY})$ is finite. If this quantity is infinite, then the KKT conditions may not hold at $Q^*_{X|Y}$.

such that

$$Q(y)\Big(\log\frac{Q^*(x|y)Q(y)}{P(x,y)}+1+\lambda\Big)-\mu_{x,y}+\nu_y$$
$$+\lambda\Big(\sum_s Q(s)\Big(Q(y|s)\log\Big(\sum_{y'}Q^*(x|y')Q(y'|s)\Big)\Big)\Big)=0 \quad \text{for all } x,y$$

$$\mu_{x,y}Q(x|y)=0 \quad \text{for all } x,y$$

$$\lambda(H(Q^*_{X|S}|Q_S)-R_1)=0$$

$$\nu_y(\sum_x Q^*(x|y)-1)=0 \quad \text{for all } y.$$

By Carathéodory's theorem, there exists $\tilde{Q}(s)$ such that

$$|s:\tilde{Q}(s)>0|\le|\mathcal{X}|\cdot|\mathcal{Y}|+|\mathcal{Y}|+2$$

and

$$\sum_s\tilde{Q}(s)Q(y|s)=Q(y) \quad \text{for all } y$$

$$Q(y)\Big(\log\frac{Q^*(x|y)Q(y)}{P(x,y)}+1+\lambda\Big)-\mu_{x,y}+\nu_y$$
$$\lambda\Big(\sum_s\tilde{Q}(s)\Big(Q(y|s)\log\Big(\sum_{y'}Q^*(x|y')Q(y'|s)\Big)\Big)\Big)=0 \quad \text{for all } x,y$$

$$I(Q_S;Q_{Y|S})=I(\tilde{Q}_S;Q_{Y|S})$$

$$H(Q^*_{X|S}|Q_S)=H(Q^*_{X|S}|\tilde{Q}_S).$$

Define $\tilde{Q}_{S|Y}$ via $Q_{Y|S}\tilde{Q}_S/Q_Y$. Then $Q^*_{X|Y}$ satisfies

$$H(Q^*_{X|S}|\tilde{Q}_S)\ge R_1$$

and

$$Q(y)\Big(\log\frac{Q^*(x|y)Q(y)}{P(x,y)}+1+\lambda\Big)-\mu_{x,y}+\nu_y$$
$$+\lambda\Big(\sum_s\tilde{Q}(s)\Big(Q(y|s)\log\Big(\sum_{y'}Q^*(x|y')Q(y'|s)\Big)\Big)\Big)=0 \quad \text{for all } x,y$$

$$\mu_{x,y}Q(x|y)=0$$
$$\lambda(H(Q^*_{X|S}|\tilde{Q}_S)-R_1)=0$$
$$\nu_y(\sum_x Q^*(x|y)-1)=0 \quad \text{for all } y.$$

Since $\gamma(Q_Y,\cdot)$ is convex, the KKT conditions are also sufficient for optimality, and we have

$$\gamma(Q_Y,\tilde{Q}_{S|Y})=D(Q_{XY}||P_{XY})=\gamma(Q_Y,Q_{S|Y}).$$

□

Lemma 46. *For $R_1<\log|\mathcal{X}_1|$, we have*

$$\lim_{\epsilon\to 0}\tilde{\eta}_U(P_{XY},R_1+\epsilon,R_2+\epsilon)=\eta_U(P_{XY},R_1,R_2).$$

Proof. Clearly $\tilde{\eta}_U(P_{XY},R_1+\epsilon,R_2+\epsilon)\geq\eta_U(P_{XY},R_1,R_2)$ for all $\epsilon>0$. To show the reverse inequality, fix a sequence $\epsilon_n\downarrow 0$. Note that there exists Q^*_Y such that[2]

$$\sup_{\substack{Q_{S|Y}:\\ I(Q^*_Y;Q_{S|Y})\leq R_2}}\ \inf_{\substack{Q_{X|Y}:\\ H(Q_{X|S}|Q_S)>R_1}} D(Q^*_YQ_{X|Y}||P_{XY})\leq$$

$$\inf_{Q_Y}\ \sup_{\substack{Q_{S|Y}:\\ I(Q_Y;Q_{S|Y})\leq R_2}}\ \inf_{\substack{Q_{X|Y}:\\ H(Q_{X|S}|Q_S)>R_1}} D(Q_{XY}||P_{XY})+\delta.$$

For each n, there exists $Q^{(n)}_{S|Y}$ such that

$$\inf_{\substack{Q_{X|Y}:\\ H(Q_{X|S}|Q^{(n)}_S)\geq R_1+\epsilon_n}} D(Q_{X|Y}Q^*_Y||P_{XY})\geq \sup_{\substack{Q_{S|Y}:\\ I(Q^*_Y;Q_{S|Y})\leq R_2+\epsilon_n}}\ \inf_{\substack{Q_{X|Y}:\\ H(Q_{X|S}|Q_S)\geq R_1+\epsilon_n}} D(Q_{X|Y}Q^*_Y||P_{XY})-\delta.$$

[2]Throughout this proof, $Q_{S|Y}$ is assumed to satisfy the cardinality bound (4.6).

By considering subsequences, we may assume that

$$Q_{S|Y}^{(n)} \to Q_{S|Y}^{\infty}.$$

Then there exists $Q_{X|Y}^{\infty}$ such that

$$H(Q_{X|S}^{\infty}|Q_S^{\infty}) > R_1.$$

and

$$D(Q_{X|Y}^{\infty}Q_Y^*||P_{XY}) \leq \inf_{\substack{Q_{X|Y}: \\ H(Q_{X|S}|Q_S^{\infty})>R_1}} D(Q_{X|Y}Q_Y^*||P_{XY}) + \delta.$$

Note that for all sufficiently large n, we have

$$H(Q_{X|S}^{\infty}|Q_S^{(n)}) \geq R_1 + \epsilon_n.$$

Then for all sufficiently large n,

$$\begin{aligned}
\tilde{\eta}_U(P_{XY}, R_1+\epsilon_n, R_2+\epsilon_n) &\leq \sup_{\substack{Q_{S|Y}: \\ I(Q_Y^*;Q_{S|Y})\leq R_2+\epsilon_n}} \inf_{\substack{Q_{X|Y}: \\ H(Q_{X|S}|Q_S)\geq R_1+\epsilon_n}} D(Q_{X|Y}Q_Y^*||P_{XY}) \\
&\leq \inf_{\substack{Q_{X|Y}: \\ H(Q_{X|S}|Q_S^{(n)})\geq R_1+\epsilon_n}} D(Q_{X|Y}Q_Y^*||P_{XY}) + \delta \\
&\leq D(Q_{X|Y}^{\infty}Q_Y^*||P_{XY}) + \delta.
\end{aligned}$$

Thus

$$\limsup_{n\to\infty} \tilde{\eta}_U(P_{XY}, R_1+\epsilon_n, R_2+\epsilon_n) \leq D(Q_{X|Y}^{\infty}Q_Y^*||P_{XY}) + \delta. \tag{B.13}$$

On the other hand, we have

$$\begin{aligned}
D(Q_{X|Y}^{\infty}Q_Y^*||P_{XY}) &\leq \inf_{\substack{Q_{X|Y}: \\ H(Q_{X|S}|Q_S^{\infty})>R_1}} D(Q_{X|Y}Q_Y^*||P_{XY}) + \delta \\
&\leq \sup_{\substack{Q_{S|Y}: \\ I(Q_Y^*;Q_{S|Y})\leq R_2}} \inf_{\substack{Q_{X|Y}: \\ H(Q_{X|S}|Q_S)>R_1}} D(Q_{X|Y}Q_Y^*||P_{XY}) + \delta \\
&\leq \eta_U(P_{XY}, R_1, R_2) + 2\delta.
\end{aligned}$$

Combining this with (B.13) yields

$$\limsup_{n\to\infty} \tilde{\eta}_U(P_{XY}, R_1+\epsilon_n, R_2+\epsilon_n) \le \eta_U(P_{XY}, R_1, R_2) + 3\delta,$$

but $\delta > 0$ and $\epsilon_n \to 0$ were arbitrary. □

Proof of Theorem 16. Recall that we may assume $R_1 < \log|\mathcal{X}_1|$. As we are eventually considering small ϵ, we may assume that $R_1 + 2\epsilon \le \log|\mathcal{X}_1|$. Take n sufficiently large so that $\frac{1}{n} \le \frac{\epsilon}{2}$.

Let (f_1^n, f_2^n, g^n) be any code satisfying (4.2) and let

$$\mathcal{E}^n(f_1^n, f_2^n, g^n) = \{(\mathbf{x}, \mathbf{y}) : g^n(f_1^n(\mathbf{x}), f_2^n(\mathbf{y})) \neq \mathbf{x}\}.$$

denote its erroneous sequences. Take any Q_{XY} such that

$$H_{Q_{XY}}(X^n | f_2^n(Y^n)) \ge n(R_1 + 2\epsilon). \tag{B.14}$$

We first show that for this choice of Q_{XY} the following inequality holds

$$Q_{XY}^n(\mathcal{E}^n(f_1^n, f_2^n, g^n)) \ge \frac{\epsilon}{2\log|\mathcal{X}|} > 0. \tag{B.15}$$

Fano's inequality gives

$$Q_{XY}^n(\mathcal{E}^n(f_1^n, f_2^n, g^n)) \ge \frac{H(X^n | f_1^n(X^n), f_2^n(Y^n)) - 1}{\log|\mathcal{X}^n|}. \tag{B.16}$$

But

$$\begin{aligned} H(X^n, f_1^n(X^n) | f_2^n(Y^n)) &= H(X^n | f_2^n(Y^n)) + H(f_1^n(X^n) | X^n, f_2^n(Y^n)) = H(X^n | f_2^n(Y^n)) \\ &= H(f_1^n(X^n) | f_2^n(Y^n)) + H(X^n | f_1^n(X^n), f_2^n(Y^n)). \end{aligned}$$

Therefore

$$\begin{aligned} H(X^n | f_1^n(X^n), f_2^n(Y^n)) &= H(X^n | f_2^n(Y^n)) - H(f_1^n(X^n) | f_2^n(Y^n)) \\ &\ge H(X^n | f_2^n(Y^n)) - H(f_1^n(X^n)) \\ &\ge H(X^n | f_2^n(Y^n)) - n(R_1 + \epsilon) \\ &\ge n\epsilon. \end{aligned} \tag{B.17}$$

The fact that $\frac{1}{n} \leq \frac{\epsilon}{2}$ along with equations (B.16) and (B.17) gives (B.15). For $\delta > 0$ define the set

$$\mathcal{D}^n = \left\{ (\mathbf{x}, \mathbf{y}) : \left| \frac{1}{n} \log \frac{Q_{XY}^n(\mathbf{x}, \mathbf{y})}{P_{XY}^n(\mathbf{x}, \mathbf{y})} - D(Q_{XY}||P_{XY}) \right| \leq \delta \right\}.$$

Fix $0 < \alpha < \infty$ such that for all distributions Q_{XY},

$$\mathbb{E}_Q \left[\log^2 \frac{Q(X,Y)}{P(X,Y)} \right] \leq \alpha.$$

Such an α exists because the alphabet is finite. By Chebyshev's inequality we have

$$\begin{aligned}
Q_{XY}^n(\mathcal{D}^n) &= 1 - Q_{XY}^n((\mathcal{D}^n)^c) \\
&\geq 1 - (\delta^{-2}) \mathbb{E}_Q \left[\left(\frac{1}{n} \sum_i \log \frac{Q(X_i, Y_i)}{P(X_i, Y_i)} - D(Q_{XY}||P_{XY}) \right)^2 \right] \\
&\geq 1 - \frac{\mathbb{E}_Q \left[\log^2 \frac{Q(X,Y)}{P(X,Y)} \right]}{n\delta^2} \\
&\geq 1 - \frac{\alpha}{\delta^2 n}
\end{aligned}$$

We may bound the error probability as follows

$$\begin{aligned}
P_{XY}^n(\mathcal{E}^n(f_1^n, f_2^n, g^n)) &\geq P_{XY}^n(\mathcal{E}^n(f_1^n, f_2^n, g^n) \cap \mathcal{D}^n) \\
&= \sum_{\mathcal{E}^n(f_1^n, f_2^n, g^n) \cap \mathcal{D}^n} Q_{XY}^n(\mathbf{x}, \mathbf{y}) \exp\left(- \log \frac{Q_{XY}^n(\mathbf{x}, \mathbf{y})}{P_{XY}^n(\mathbf{x}, \mathbf{y})} \right) \\
&\geq Q_{XY}^n(\mathcal{E}^n(f_1^n, f_2^n, g^n) \cap \mathcal{D}^n) \exp\left(-n(D(Q_{XY}||P_{XY}) + \delta) \right) \\
&\geq \left(\frac{\epsilon}{2 \log |\mathcal{X}|} - \frac{\alpha}{\delta^2 n} \right) \exp\left(-n(D(Q_{XY}||P_{XY}) + \delta) \right). \quad \text{(B.18)}
\end{aligned}$$

However, for n large enough

$$\frac{\epsilon}{2 \log |\mathcal{X}|} - \frac{\alpha}{\delta^2 n} \geq \frac{\epsilon}{4 \log |\mathcal{X}|} \triangleq \beta > 0,$$

thus, observing that the argument above holds for every Q_{XY} satisfying (B.14) we see that

$$P_{XY}^n(\mathcal{E}^n(f_1^n, f_2^n, g^n)) \geq \sup_{Q_{XY} : H_Q(X^n | f_2^n(Y^n)) \geq n(R_1 + 2\epsilon)} \beta \exp\left(-n(D(Q_{XY}||P_{XY}) + \delta) \right).$$

Now we note that the above holds for every code satisfying (4.2), thus, observing that the right hand side does not depend on f_1^n, g^n, we conclude that

$$\min_{f_1^n, f_2^n, g^n} P_{XY}^n(\mathcal{E}^n(f_1^n, f_2^n, g^n)) \geq \min_{f_2^n} \sup_{Q_{XY}: H_Q(X^n|f_2^n(Y^n)) \geq n(R_1+2\epsilon)} \beta \exp\left(-n(D(Q_{XY}||P_{XY}) + \delta)\right).$$

We now move the optimizations into the exponent and focus our attention there.

$$\begin{aligned}
&\max_{\substack{f_2^n: \\ \log|f_2^n| \leq n(R_2+\epsilon)}} \inf_{Q_{XY}: H_Q(X^n|f_2^n(Y^n)) \geq n(R_1+2\epsilon)} D(Q_{XY}||P_{XY}) \\
&= \max_{\substack{f_2^n: \\ \log|f_2^n| \leq n(R_2+\epsilon)}} \inf_{Q_Y} \inf_{\substack{Q_{X|Y}: \\ H_Q(X^n|f_2^n(Y^n)) \geq n(R_1+2\epsilon)}} D(Q_{XY}||P_{XY}) \\
&\leq \inf_{Q_Y} \max_{\substack{f_2^n: \\ \log|f_2^n| \leq n(R_2+\epsilon)}} \inf_{\substack{Q_{X|Y}: \\ H_Q(X^n|f_2^n(Y^n)) \geq n(R_1+2\epsilon)}} D(Q_{XY}||P_{XY}) \\
&\leq \inf_{Q_Y} \max_{\substack{f_2^n: \\ I(Y^n; f_2^n(Y^n)) \leq n(R_2+\epsilon)}} \inf_{\substack{Q_{X|Y}: \\ H_Q(X^n|f_2^n(Y^n)) \geq n(R_1+2\epsilon)}} D(Q_{XY}||P_{XY}) \\
&\leq \inf_{Q_Y} \sup_{\substack{Q_{U|Y^n}: \\ I(Y^n;U) \leq n(R_2+\epsilon)}} \inf_{\substack{Q_{X|Y}: \\ H_Q(X^n|U) \geq n(R_1+2\epsilon)}} D(Q_{XY}||P_{XY})
\end{aligned} \tag{B.19}$$

In the previous line, we note that the deterministic functions are still feasible and on deterministic functions the previous two bounds agree. Henceforth the joint distribution of X, Y, U is $Q_Y Q_{U|Y} Q_{X|Y}$, so that X, Y and U form a Markov chain. To continue we use of the following, obtained via the chain rule

$$\begin{aligned}
H(X^n|U) &= \sum_{i=1}^n H(X_i|U, X_1^{i-1}) \\
&\geq \sum_{i=1}^n H(X_i|U, X_1^{i-1}, Y_1^{i-1}) \\
&= \sum_{i=1}^n H(X_i|U, Y_1^{i-1})
\end{aligned} \tag{B.20}$$

where on the final line we used the fact that $X_i - (U, Y_1^{i-1}) - X_1^{i-1}$. The following

identity also holds

$$\begin{aligned} I(Y^n;U) &= \sum_{i=1}^{n} I(Y_i;U|Y_1^{i-1}) \\ &= \sum_{i=1}^{n} H(Y_i|Y_1^{i-1}) - H(Y_i|Y_1^{i-1},U) \\ &= \sum_{i=1}^{n} I(Y_i;Y_1^{i-1},U). \end{aligned} \tag{B.21}$$

Substituting (B.20) into (B.19) makes the feasible set smaller because of the inequality. After substituting (B.21), we can continue to bound the exponent by

$$\begin{aligned} &\leq \inf_{Q_Y} \sup_{\substack{Q_{U|Y^n}: \\ \frac{1}{n}\sum_{i=1}^n I(Y_i;Y_1^{i-1},U)\leq R_2+\epsilon}} \inf_{\substack{Q_{X|Y}: \\ \frac{1}{n}\sum_{i=1}^n H(X_i|U,Y_1^{i-1})\geq R_1+2\epsilon}} D(Q_{XY}||P_{XY}) \\ &= \inf_{Q_Y} \sup_{\substack{Q_{U|Y^n}: \\ \frac{1}{n}\sum_{i=1}^n I(Y_i;V_i)\leq R_2+\epsilon}} \inf_{\substack{Q_{X|Y}: \\ \frac{1}{n}\sum_{i=1}^n H(X_i|V_i)\geq R_1+2\epsilon}} D(Q_{XY}||P_{XY}) \end{aligned}$$

where on the previous line, we let $V_i = (Y_1^{i-1}, U)$. Let T denote a time sharing random variable, uniformly distributed on $\{1,\ldots,n\}$ and independent of everything else. Then the quantity above can be written

$$\begin{aligned} &\inf_{Q_Y} \sup_{\substack{Q_{U|Y^n}: \\ I(Y_T;V_T,T)\leq R_2+\epsilon}} \inf_{\substack{Q_{X|Y}: \\ H(X_T|V_T,T)\geq R_1+2\epsilon}} D(Q_{XY}||P_{XY}) \\ &= \inf_{Q_Y} \sup_{\substack{Q_{U|Y^n}: \\ I(Y_T;W)\leq R_2+\epsilon}} \inf_{\substack{Q_{X|Y}: \\ H(X_T|W)\geq R_1+2\epsilon}} D(Q_{XY}||P_{XY}). \end{aligned} \tag{*}$$

where we set $W = (V_T, T) = (Y_1^{T-1}, U, T)$. Since $(X_T, Y_T) \stackrel{d}{=} (X, Y)$, the above quantity is upper bounded by

$$\inf_{Q_Y} \sup_{\substack{Q_{S|Y}: \\ I(Y;S)\leq R_2+2\epsilon}} \inf_{\substack{Q_{X|Y}: \\ H(X|S)\geq R_1+2\epsilon}} D(Q_{XY}||P_{XY}) = \overline{\eta}_U(P_{XY}, R_1+2\epsilon, R_2+2\epsilon).$$

To see this, we note that every choice in $(*)$ is a feasible choice in F. In particular for a given Q_Y, let U^* denote a choice for $Q_{U|Y^n}$ in $(*)$, then choosing S

so that $(Y,S) \stackrel{d}{=} (Y, Y_1^{T-1}, U^*, T)$, is feasible. By Lemma 45, this quantity equals $\tilde{\eta}_U(P_{XY}, R_1+2\epsilon, R_2+2\epsilon)$. Thus we have shown that

$$\min_{f_1^n, f_2^n, g^n} P_{XY}^n(\mathcal{E}^n(f_1^n, f_2^n, g^n)) \geq \beta \exp\left(-n(\tilde{\eta}_U(P_{XY}, R_1+2\epsilon, R_2+2\epsilon)+\delta)\right).$$

Taking logs and the $\limsup$ as $n \to \infty$, and letting $\delta \downarrow 0$ and $\epsilon \downarrow 0$ (and invoking Lemma 46) gives the result. □

B.3 Proof of Theorem 17

B.3.1 Scheme

For a given blocklength n, we operate on a type-by-type basis and define the encoder and decoder functions as follows. For each type Q_X, fix a conditional type $Q^*_{Z|X}(Q_X) \in \mathcal{C}^n(Q_X, \mathcal{Y})$, a decoding function $f(Q_X, Q_Y) \in \mathcal{F}$, and randomly choose a set of codewords $B^n(Q_X)$ in the following way. The size of $B^n(Q_X)$ is an integer satisfying

$$\begin{aligned} &\exp(nI(Q_X; Q^*_{Z|X}(Q_X)) + (|\mathcal{X}||\mathcal{Z}|+2)\log(n+1)) \\ &\leq |B^n(Q_X)| \\ &\leq \exp(nI(Q_X; Q^*_{Z|X}(Q_X)) + (|\mathcal{X}||\mathcal{Z}|+4)\log(n+1)) \end{aligned} \tag{B.22}$$

and the codewords are drawn uniformly, with replacement, from the marginal type class $T^n_{Q^*_Z}$ induced by Q_X and $Q^*_{Z|X}(Q_X)$.

Define $Z : T^n_{Q_{\mathbf{x}}} \to B^n(Q_{\mathbf{x}})$ as follows. Let $\mathcal{G}(\mathbf{x}) \triangleq B^n(Q_{\mathbf{x}}) \cap T^n_{Q^*_{Z|X}(Q_{\mathbf{x}})}(\mathbf{x})$, if $\mathcal{G}(\mathbf{x})$ is non-empty, then the output of $Z(\mathbf{x})$ is drawn uniformly at random from $\mathcal{G}(\mathbf{x})$. If $\mathcal{G}(\mathbf{x})$ is empty the output of $Z(\mathbf{x})$ is drawn uniformly at random from

$B^n(Q_{\mathbf{x}})$. The function $Z(\cdot)$ determines the codeword sent by the encoder to the decoder. We define $Z^n = Z(X^n)$ and define the encoder's message set as follows

$$\begin{aligned}
\mathcal{M} &= \mathcal{M}_1 \times \mathcal{M}_2, \\
\mathcal{M}_1 &= \{1, 2, \ldots, M_1 \triangleq \exp(nR)\}, \\
\mathcal{M}_2 &= \{1, 2, \ldots, (n+1)^{|\mathcal{X}|}\}.
\end{aligned}$$

Operation of the encoder: To encode a sequence $\mathbf{x} \in T^n_{Q_X}$, the encoder sends the type of $\mathbf{x}$ and an index, $U(Z(\mathbf{x}))$, of the codeword $Z(\mathbf{x})$. There are two cases to consider:

1. $\log |B^n(Q_X)| < nR$, in which case we can map each member of $B^n(Q_X)$ to an element of $\mathcal{M}_1$ in a one-to-one manner.
2. $\log |B^n(Q_X)| \geq nR$, in which case we assign each member of $B^n(Q_X)$ to $\mathcal{M}_1$ uniformly at random.

Let $U(Z(\mathbf{x}))$ denote the element to which $Z(\mathbf{x})$ is mapped. The encoder can be expressed mathematically as

$$\psi(\mathbf{x}) = (U(Z(\mathbf{x})), k(Q_X)) \text{ for } \mathbf{x} \in T^n_{Q_X} \tag{B.23}$$

Operation of the Decoder: The decoder operates in a two-step manner. First it attempts to recover the codeword Z^n:

1. If $|B^n(Q_X)| < nR$ then Z^n can be decoded without error,
2. If $|B^n(Q_X)| \geq nR$ the decoder receives a bin index and uses the side information to pick the "best" $\mathbf{z}$ from the bin in the minimum conditional

entropy sense: it searches for a $\hat{\mathbf{z}}$ in the received bin so that among all $\tilde{\mathbf{z}}$ in the bin, $H(\tilde{\mathbf{z}}|\mathbf{y}) > H(\hat{\mathbf{z}}|\mathbf{y})$. If there is no such $\hat{\mathbf{z}}$ it picks uniformly at random from the bin.

Let

$$\varphi_1(i, k(Q_X), \mathbf{y}) = \begin{cases} \hat{\mathbf{z}} & \hat{\mathbf{z}} \in \mathrm{Bin}(i) \text{ and } \forall \tilde{\mathbf{z}} \in \mathrm{Bin}(i), \\ & \tilde{\mathbf{z}} \neq \hat{\mathbf{z}} : H(\tilde{\mathbf{z}}|\mathbf{y}) > H(\hat{\mathbf{z}}|\mathbf{y}) \\ \text{any } \tilde{\mathbf{z}} \in \mathrm{Bin}(i) & \text{if no such } \hat{\mathbf{z}} \in \mathrm{Bin}(i) \end{cases} \tag{B.24}$$

where $\mathrm{Bin}(i) = \{\mathbf{z} : \mathbf{z} \in B^n(Q_X) \text{ and } U(\mathbf{z}) = i\}$ denotes the set of codewords that are assigned to index i. Second, the decoder uses the estimation function, f, to combine the side information $\mathbf{y}$ with codeword $\mathbf{z}$ to give the reproduction $\hat{\mathbf{x}}$. This is expressed mathematically as

$$\varphi(i, k(Q_X), \mathbf{y}) = \hat{\mathbf{x}} \text{ s.t. } \hat{\mathbf{x}}_j = f(\varphi_1(i, k(Q_X), \mathbf{y})_j, \mathbf{y}_j). \tag{B.25}$$

B.3.2 Error probability calculation

It will be convenient to consider the following subsets of the sequence space

$$\begin{aligned} \mathcal{E}_b &= \Big\{(\mathbf{x}, \mathbf{y}, \mathbf{z}) : \mathbf{z} \in T^n_{Q^*_{Z|X}(Q_\mathbf{x})}(\mathbf{x}), d(\mathbf{x}, f(\mathbf{y}, \mathbf{z})) < \Delta, \\ &\qquad \log |B^n(Q_\mathbf{x})| \geq nR\} \\ \mathcal{E}_c &= \Big\{(\mathbf{x}, \mathbf{y}, \mathbf{z}) : \mathbf{z} \notin T^n_{Q^*_{Z|X}(Q_\mathbf{x})}(\mathbf{x})\Big\} \\ \mathcal{E}_d &= \Big\{(\mathbf{x}, \mathbf{y}, \mathbf{z}) : \mathbf{z} \in T^n_{Q^*_{Z|X}(Q_\mathbf{x})}(\mathbf{x}), d(\mathbf{x}, f(\mathbf{y}, \mathbf{z})) \geq \Delta\Big\} \end{aligned}$$

$\mathcal{E}_b$ corresponds to a potential binning error, $\mathcal{E}_c$ to a covering error and $\mathcal{E}_d$ to a distortion error. We will consider the errors on these sets separately. Equivalently

we can view these error events as properties of the joint type, so we define

$$\begin{aligned}
\mathcal{D}_b &= \{Q_{XYZ} : \mathbb{E}[d(X, f(Y,Z))] \leq \Delta, Q_{Z|X} = Q^*_{Z|X}(Q_X) \\
&\qquad \log |B^n(Q_X)| \geq nR\} \\
\mathcal{D}_c &= \{Q_{XYZ} : Q_{Z|X} \neq Q^*_{Z|X}(Q_X)\} \\
\mathcal{D}_d &= \{Q_{XYZ} : \mathbb{E}[d(X, f(Y,Z))] > \Delta, \\
&\qquad Q_{Z|X} = Q^*_{Z|X}(Q_X)\}.
\end{aligned}$$

Before we proceed with the proof of Theorem 1, we establish the following useful facts.

Lemma 47. *Let $X^n, Y^n, Z^n = \hat{Z}(X^n)$ be generated according to our scheme and suppose that $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ is in $(\mathcal{E}_c)^c$, i.e. that $\mathbf{z} \in T^n_{Q^*_{Z|X}(Q_{\mathbf{x}})}(\mathbf{x})$. Then*

$$\begin{aligned}
&\Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) && \text{(B.26)} \\
&\qquad \leq P^n_{XY}(\mathbf{x}, \mathbf{y}) \frac{1}{|T^n_{Q^*_{Z|X}(Q_{\mathbf{x}})}(\mathbf{x})|}. && \text{(B.27)}
\end{aligned}$$

Proof. The proof mirrors that of Lemma 40 and is omitted. □

Lemma 48. *Let $X^n, Y^n, Z^n = Z(X^n)$ be generated according to our scheme and suppose that $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{E}_c$. Then*

$$\begin{aligned}
&\Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \\
&\qquad \leq \exp(-(n+1)^2). && \text{(B.28)}
\end{aligned}$$

Proof. The proof mirrors that of Lemma 41 and is omitted. □

Lemma 49. *For all strings* $\mathbf{x}, \mathbf{z}$ *such that* $\mathbf{z} \in T^n_{Q^*_Z}$,

$$\begin{aligned}\Pr(\mathbf{z} \in B^n(Q_\mathbf{x})) \leq \\ (n+1)^{|\mathcal{Z}|(1+|\mathcal{X}|)+4} \\ \times \exp(n(I(Q_\mathbf{x}; Q^*_{Z|X}(Q_\mathbf{x})) - H(Q_\mathbf{z}))).\end{aligned}$$

Proof. By the construction of $B^n(Q_\mathbf{x})$, each of the codewords is chosen with replacement from the set $T^n_{Q^*_\mathbf{z}}$. Thus each string has probability $|T^n_{Q^*_\mathbf{z}}|^{-1}$ and we make $|B^n(Q_\mathbf{x})|$ such choices (bounded by (B.22)). From [27, lemma 2.3] we have

$$|T_{Q_\mathbf{z}}| \geq (n+1)^{-|\mathcal{Z}|} \exp(nH(Q_\mathbf{z})).$$

Invoking the union bound gives the result. □

Lemma 50. *Let* $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (\mathcal{E}_c \cup \mathcal{E}_d)^c$. *Then*

$$\begin{aligned}&\Pr(d(X^n, \hat{X}^n) > \Delta | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \\ &\leq \exp\left(-n\left((R - J(Q_{\mathbf{xyz}}) - \delta^n_b)^+\right)\right) \qquad \text{(B.29)}\end{aligned}$$

where

$$\begin{aligned}J(Q_{\mathbf{xyz}}) &= I(Q_\mathbf{x}; Q^*_{Z|X}(Q_\mathbf{x})) - I(Q_\mathbf{y}; Q_{\mathbf{z}|\mathbf{y}}) \\ \text{and } \delta^n_b &= \frac{1}{n} \log(n+1)^{|\mathcal{Z}|(|\mathcal{Y}|+1+|\mathcal{X}|)+4}.\end{aligned}$$

Moreover, if $\log |B^n(Q_\mathbf{x})| < nR$ *then*

$$\Pr(d(X^n, \hat{X}^n) > \Delta | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) = 0.$$

Proof. For the given sequence $\mathbf{x}, \mathbf{y}, \mathbf{z}$ let L be the event that $\mathbf{z} \neq \varphi_1(\psi(\mathbf{x}), \mathbf{y})$. (Observe that L occurs when the decoder decodes the wrong codeword and that $\Pr(d(X^n, \hat{X}^n) > \Delta | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z})$ is upper bounded by $\Pr(L | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z})$.)

If $Q_{\mathbf{x}}$ is such that $\log|B^n(Q_{\mathbf{x}})| < nR$, then

$$\Pr(L|X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) = 0.$$

For the case where $\log|B^n(Q_{\mathbf{x}})| \geq nR$ (i.e. $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{E}_b$), we note that the set $S(\mathbf{z}|\mathbf{y})$ contains all strings $\tilde{\mathbf{z}}$ having the property that $\tilde{\mathbf{z}}$ has the same type as $\mathbf{z}$ and lower conditional empirical entropy.

$$\begin{aligned}
&\Pr(L|X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \\
&\quad \leq \sum_{\tilde{\mathbf{z}} \in S(\mathbf{z}|\mathbf{y})} \Pr(\tilde{\mathbf{z}} \in B^n(Q_{\mathbf{x}}), U(\tilde{\mathbf{z}}) = U(\mathbf{z})) \\
&\quad = \sum_{\tilde{\mathbf{z}} \in S(\mathbf{z}|\mathbf{y})} \Pr(\tilde{\mathbf{z}} \in B^n(Q_{\mathbf{x}})) \\
&\qquad \times \Pr(U(\tilde{\mathbf{z}}) = U(\mathbf{z}) | \tilde{\mathbf{z}} \in B^n(Q_{\mathbf{x}})) \\
&\quad \leq \sum_{\tilde{\mathbf{z}} \in S(\mathbf{z}|\mathbf{y})} (n+1)^{|\mathcal{Z}|(1+|\mathcal{X}|)+4} \frac{1}{M_1} \\
&\qquad \times \exp(n(I(Q_{\mathbf{x}}; Q^*_{Z|X}(Q_{\mathbf{x}})) - H(Q_{\mathbf{z}}))) \qquad \text{(B.30)}
\end{aligned}$$

where the last line follows from Lemma 49. Next,

$$\begin{aligned}
&\Pr(L|X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \\
&\quad \leq (n+1)^{|\mathcal{Z}|(|\mathcal{Y}|+1+|\mathcal{X}|)+4} \exp(nH(Q_{\mathbf{z}|\mathbf{y}}|Q_{\mathbf{y}})) \\
&\qquad \times \exp(n(I(Q_{\mathbf{x}}; Q^*_{Z|X}(Q_{\mathbf{x}})) - H(Q_{\mathbf{z}}))) \frac{1}{M_1} \\
&\quad = (n+1)^{|\mathcal{Z}|(|\mathcal{Y}|+1+|\mathcal{X}|)+4} \exp\left(-n\left(R - J(Q_{\mathbf{xyz}})\right)\right)
\end{aligned}$$

where the first line follows from Lemma 42. Also, since $\Pr(L|X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \leq 1$ we get

$$\begin{aligned}
&\Pr(L|X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \\
&\qquad = \exp(-n\left(R - J(Q_{\mathbf{xyz}}) - \delta_b^n\right)^+).
\end{aligned}$$

□

Lemma 51. *Let $\delta_b^n \to 0$ with n,*

$$G^n[Q_{XYZ}, P_{XY}, f, d, \Delta, R] =$$
$$\begin{cases} D(Q_{XYZ}||P_{XY}Q_{Z|X}) & \mathbb{E}_Q[d(X, f(Y,Z))] \geq \Delta \\ D(Q_{XYZ}||P_{XY}Q_{Z|X}) & \\ \quad +\big(R - I(Q_X; Q^*_{Z|X}(Q_X)) & \\ \quad +I(Q_Y; Q_{Z|Y}) - \delta_b^n\big)^+ & \mathbb{E}_Q[d(X, f(Y,Z))] < \Delta \\ & \quad I(Q_X; Q_{Z|X}) \geq R \\ 0 & \textit{otherwise} \end{cases}$$

and

$$\theta^n(P_{XY}, d, \Delta, R) = \min_{Q_X} \max_{Q_{Z|X} \in \mathcal{C}^n(\mathcal{X}\to\mathcal{Z})} \min_{Q_Y} \max_{f\in\mathcal{F}} \min_{Q_{XYZ}} G^n(Q_{XYZ}, P_{XY}, f, d, \Delta, R),$$

$$\theta^\infty(P_{XY}, d, \Delta, R) = \inf_{Q_X} \sup_{Q_{Z|X}} \inf_{Q_Y} \sup_{f\in\mathcal{F}} \inf_{Q_{XYZ}} G(Q_{XYZ}, P_{XY}, f, d, \Delta, R).$$

In θ^n the minimizations and maximizations on $Q_X, Q_{Z|X}, Q_Y$ and Q_{XYZ} are over types/conditional types, and in θ^∞ they are over distributions. And, in the optimization of Q_{XYZ} the marginal type/distribution of X and Y and conditional type/distribution of Z given X are taken to be those specified earlier in the optimization. Then

$$\liminf_{n\to\infty} \theta^n(P_{XY}, d, \Delta, R) \geq \theta^\infty(P_{XY}, d, \Delta, R) \tag{B.31}$$

Proof. Choose $\delta > 0$ and n sufficiently large so that $G - G^n < \frac{\delta}{2}$ (i.e. $\delta_b^n < \frac{\delta}{2}$). Let $Q_X^{(n)}, Q_{Z|X}^{(n)}, Q_Y^{(n)}, Q_{XYZ}^{(n)}$ and $f^{(n)}$ be such that

$$\theta^n(P_{XY}, d, \Delta, R) = G^n(Q_{XYZ}^{(n)}, P_{XY}, f^{(n)}, d, \Delta, R).$$

For convenience, henceforth we omit writing the arguments P_{XY}, d, Δ and R in $G(\cdot)$ and $G^n(\cdot)$. Also, when necessary for clarity, we expand $Q_{XYZ} = Q_X, Q_{Z|X}, Q_Y, Q_{Y|XZ}$ in the argument to G and $G^n(\cdot)$.

By boundedness there exists a subsequence of $(Q_X^{(n)}, Q_{Z|X}^{(n)}, Q_Y^{(n)}, Q_{XYZ}^{(n)})$ with index n' such that the sequence $(Q_X^{(n')}, Q_{Z|X}^{(n')}, Q_Y^{(n')}, Q_{XYZ}^{(n')}, f^{(n')})$ converges to a limit $(Q_X^\infty, Q_{Z|X}^\infty, Q_Y^\infty, Q_{XYZ}^\infty, f^\infty)$. There exists $\tilde{Q}_{Z|X}^\infty$ so that

$$\inf_{Q_Y} \sup_f \inf_{Q_{Y|XZ}} G(Q_X^\infty, \tilde{Q}_{Z|X}^\infty, Q_Y, Q_{XYZ}, f) \geq \sup_{Q_{Z|X}} \inf_{Q_Y} \sup_f \inf_{Q_{Y|XZ}} G(Q_X^\infty, Q_{Z|X}, Q_Y, Q_{XYZ}, f) - \frac{\delta}{2}$$

and there is a sequence $\tilde{Q}_{Z|X}^{(n')}$ converging to $\tilde{Q}_{Z|X}^\infty$. Let

$$\tilde{Q}_Y^{(n')} = \arg\min_{\bar{Q}_Y} \max_f \min_{\substack{Q_{XYZ}: \\ Q_X = Q_X^{(n')} \\ Q_{Z|X} = \tilde{Q}_{Z|X}^{(n')} \\ Q_Y = \bar{Q}_Y}} G^{n'}(Q_{XYZ}, f)$$

and by considering a further subsequence we may assume that $\tilde{Q}_Y^{(n')} \to \tilde{Q}_Y^\infty$. Then there exists $\tilde{f}^\infty$ so that

$$\inf_{Q_{Y|XZ}} G(Q_X^\infty, \tilde{Q}_{Z|X}^\infty, \tilde{Q}_Y^\infty, Q_{Y|XZ}, \tilde{f}^\infty) \geq \max_f \inf_{Q_{Y|XZ}} G(Q_X^\infty, \tilde{Q}_{Z|X}^\infty, \tilde{Q}_Y^\infty, Q_{Y|XZ}, f)$$

and we set $\tilde{f}^{(n')} = \tilde{f}^\infty$. Let

$$Q_{XYZ}^{(n')} = \arg\min_{\substack{Q_{XYZ}: \\ Q_X = Q_X^{(n')} \\ Q_{Z|X} = \tilde{Q}_{Z|X}^{(n')} \\ Q_Y = Q_Y^{(n')}}} G^{n'}(Q_{XYZ}, \tilde{f}^{(n')})$$

and by considering a further subsequence we may assume that $\tilde{Q}_{XYZ}^{(n')} \to \tilde{Q}_{XYZ}^\infty$. Observe that

$$\begin{aligned}
\theta^{n'}(P_{XY}, d, \Delta, R) &= \max_{Q_{Z|X} \in \mathcal{C}^{n'}(\mathcal{X} \to \mathcal{Z})} \min_{Q_Y} \max_{f \in \mathcal{F}} \min_{Q_{Y|XZ}} G^{n'}(Q_X^{(n')}, Q_{Z|X}, Q_Y, Q_{Y|XZ}, f) \\
&\geq \min_{Q_Y} \max_{f \in \mathcal{F}} \min_{Q_{Y|XZ}} G^{n'}(Q_X^{(n')}, \tilde{Q}_{Z|X}^{(n')}, Q_Y, Q_{Y|XZ}, f) \\
&= \max_{f \in \mathcal{F}} \min_{Q_{Y|XZ}} G^{n'}(Q_X^{(n')}, \tilde{Q}_{Z|X}^{(n')}, \tilde{Q}_Y^{(n')}, Q_{Y|XZ}, f) \\
&\geq \min_{Q_{Y|XZ}} G(Q_X^{(n')}, Q_{Z|X}^{(n')}, Q_Y^{(n')}, Q_{Y|XZ}, \tilde{f}^{(n')}) \\
&= G^{n'}(\tilde{Q}_{XYZ}^{(n')}, \tilde{f}^{(n')})
\end{aligned}$$

But $G^{n'}(\tilde{Q}^{(n')}_{XYZ}, \tilde{f}^{(n')}) \geq G(\tilde{Q}^{(n')}_{XYZ}, \tilde{f}^{(n')})) - \frac{\delta}{2}$ and since G is lower-semicontinuous

$$
\begin{aligned}
&\liminf_{n'\to\infty} G^{n'}(Q^{(n')}_X, \tilde{Q}^{(n')}_{Z|X}, \tilde{Q}^{(n')}_Y, \tilde{Q}^{(n')}_{XYZ}, \tilde{f}^{\infty}) \\
&\geq G(Q^{\infty}_X, \tilde{Q}^{\infty}_{Z|X}, \tilde{Q}^{\infty}_Y, \tilde{Q}^{\infty}_{XYZ}, \tilde{f}^{\infty}) - \frac{\delta}{2} \\
&\geq \inf_{Q_{XYZ}} G(Q^{\infty}_X, \tilde{Q}^{\infty}_{Z|X}, \tilde{Q}^{\infty}_Y, Q_{XYZ}, \tilde{f}^{\infty}) - \frac{\delta}{2} \\
&= \sup_f \inf_{Q_{XYZ}} G(Q^{\infty}_X, \tilde{Q}^{\infty}_{Z|X}, Q^{\infty}_Y, Q_{XYZ}, f) - \frac{\delta}{2} \\
&\geq \inf_{Q_Y} \sup_f \inf_{Q_{XYZ}} G(Q^{\infty}_X, \tilde{Q}^{\infty}_{Z|X}, Q_Y, Q_{XYZ}, f) - \frac{\delta}{2} \\
&\geq \sup_{Q_{Z|X}} \inf_{Q_Y} \sup_f \inf_{Q_{XYZ}} G(Q^{\infty}_X, Q_{Z|X}, Q_Y, Q_{XYZ}, f) - \delta \\
&\geq \inf_{Q_X} \sup_{Q_{Z|X}} \inf_{Q_Y} \sup_f \inf_{Q_{XYZ}} G(Q_X, Q_{Z|X}, Q_Y, Q_{XYZ}, f) - \delta \\
&= \theta^{\infty}(P_{XY}, d, \Delta, R) - \delta
\end{aligned}
$$

Hence $\liminf_{n'\to\infty} \theta^{n'}(P_{XY}, d, \Delta, R) \geq \liminf_{n'\to\infty} G(\tilde{Q}^{(n')}_{XYZ}, \tilde{f}^{(n')}) \geq \theta(P_{XY}, d, \Delta, R) - \delta$. Letting $\delta \downarrow 0$ gives the result. □

We are now in a position to prove Theorem 17. We will accomplish this by giving an upper bound on the probability of error by considering the error events separately.

Proof of Theorem 17. We start by noting that for n sufficiently large the constraint

of equation (4.11) is satisfied. Summing over sequences gives

$$
\begin{aligned}
&\Pr(d(X^n, \hat{X}^n) > \Delta) \\
&= \sum_{\mathbf{x},\mathbf{y},\mathbf{z}} \Pr(d(X^n, \hat{X}^n) > \Delta | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \times \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \\
&\leq \sum_{\mathcal{E}_b} \Big[\Pr(d(X^n, \hat{X}^n) > \Delta | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \times \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \Big] \\
&\quad + \sum_{\mathcal{E}_c} \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \\
&\quad + \sum_{\mathcal{E}_d} \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z})
\end{aligned}
$$

where the last inequality followed from upper bounding the conditional error probability by 1 in the summations over $\mathcal{E}_c$ and $\mathcal{E}_d$, and by zero (Lemma 50) on $(\mathcal{E}_b \cup \mathcal{E}_c \cup \mathcal{E}_d)^c$ (the sequences omitted from the sum). Next, we bound the sequence probabilities using Lemma 47 on $\mathcal{E}_b$ and $\mathcal{E}_d$ and Lemma 48 on $\mathcal{E}_c$. We bound the conditional error probability on $\mathcal{E}_b$ using Lemma 50.

$$
\begin{aligned}
&\Pr(d(X^n, \hat{X}^n) > \Delta) \\
&\leq \sum_{\mathcal{E}_b} \left[\exp\left(-n\left(R - J(Q_{\mathbf{xyz}}) - \delta_b^n\right)^+\right) \times P_{XY}^n(\mathbf{x}, \mathbf{y}) \frac{1}{|T^n_{Q^*_{Z|X}(Q_{\mathbf{x}})}(\mathbf{x})|} \right] \\
&\quad + \sum_{\mathcal{E}_c} \exp(-(n+1)^2) \\
&\quad + \sum_{\mathcal{E}_d} P_{XY}^n(\mathbf{x}, \mathbf{y}) \frac{1}{|T^n_{Q^*_{Z|X}(Q_{\mathbf{x}})}(\mathbf{x})|}
\end{aligned}
$$

We can rewrite the above by first summing over types and then over sequences

within each type class. This gives us

$$\begin{aligned}
&\Pr(d(X^n, \hat{X}^n) > \Delta) \\
&\leq \sum_{Q_X} \sum_{Q_Y} \Bigg[\Bigg(\sum_{Q_{XYZ} \in \mathcal{D}_b} \sum_{(\mathbf{x},\mathbf{y},\mathbf{z}) \in T^n_{Q_{XYZ}}} P^n_{XY}(\mathbf{x},\mathbf{y}) \frac{1}{|T^n_{Q^*_{Z|X}(Q_\mathbf{x})}(\mathbf{x})|} \\
&\qquad \times \exp\left(-n\left(R - J(Q_{\mathbf{xyz}}) - \delta^n_b\right)^+\right) \Bigg) \\
&+ \left(\sum_{Q_{XYZ} \in \mathcal{D}_d} \sum_{(\mathbf{x},\mathbf{y},\mathbf{z}) \in T^n_{Q_{XYZ}}} P^n_{XY}(\mathbf{x},\mathbf{y}) \frac{1}{|T^n_{Q^*_{Z|X}(Q_\mathbf{x})}(\mathbf{x})|} \right) \\
&+ \sum_{Q_{XYZ} \in \mathcal{D}_c} \sum_{(\mathbf{x},\mathbf{y},\mathbf{z}) \in T^n_{Q_{XYZ}}} \exp(-(n+1)^2) \Bigg].
\end{aligned}$$

Note that in the summation over joint types Q_{XYZ}, the marginal types of X and Y are fixed to be those set by the earlier summations. Proceeding in a similar manner as was taken in going from (B.8) to (B.11) in the SCPSI proof (with Z taking the role of S) we get

$$\begin{aligned}
&\Pr(d(X^n, \hat{X}^n) > \Delta) \\
&\leq \sum_{Q_X} \sum_{Q_Y} \Bigg[\sum_{Q_{XYZ} \in \mathcal{D}_b} \exp\Big(-n\big(D(Q_{XYZ}||P_{XY}Q_{Z|X}) \\
&\quad + R - J(Q_{XYZ}) - \delta^n_b\big)^+\Big) \\
&+ \sum_{Q_{XYZ} \in \mathcal{D}_d} \exp\Big(-nD(Q_{XYZ}||P_{XY}Q_{Z|X})\Big) \\
&+ \exp(-(n+1)^2 + n\log(|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|)) \Bigg]
\end{aligned}$$

Next, we use $a + b \leq 2\max(a, b)$ to combine the first two terms. We can then upper bound the summations by maximizing over the types, and since the choice of test channel $Q^*_{Z|X}$ and estimation function f were arbitrary, we can

optimize to give

$$\begin{aligned}
&\Pr(d(X^n, \hat{X}^n) > \Delta) \\
&\leq \Big[|\mathcal{P}^n(\mathcal{X})| \max_{Q_X} \min_{Q^*_{Z|X}} |\mathcal{P}^n(\mathcal{Y})| \max_{Q_Y} 2|\mathcal{P}^n(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})| \\
&\quad \min_{f \in \mathcal{F}} \max_{\substack{Q_{XYZ}: \\ Q_{Z|X} = Q^*_{Z|X}}} \tilde{G}^n[Q_{XYZ}, P_{XY}, f, \Delta, R, n] \Big] \\
&+ |\mathcal{P}^n(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})| \exp(-(n+1)^2 + n \log(|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|))
\end{aligned}$$

where we used the definition of G^n from Lemma 51. Moving the optimizations into the exponent we get

$$\begin{aligned}
&\Pr(d(X^n, \hat{X}^n) > \Delta) \\
&\leq 2|\mathcal{P}^n(\mathcal{X})||\mathcal{P}^n(\mathcal{Y})||\mathcal{P}^n(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})| \exp\Big(- n \Big[\min_{Q_X} \max_{Q^*_{Z|X}} \min_{Q_Y} \\
&\quad \max_{f \in \mathcal{F}} \min_{\substack{Q_{XYZ}: \\ Q_{Z|X} = Q^*_{Z|X}}} G^n[Q_{XYZ}, P_{XY}, f, d, \Delta, R] \Big] \Big) \\
&+ |\mathcal{P}^n(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z})| \exp(-(n+1)^2 + n \log(|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|))
\end{aligned}$$

We can absorb the set cardinalities $\delta_2 = \frac{1}{n}[1+\log(n+1)^{|\mathcal{X}|+|\mathcal{Y}|+|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|}]$ and observe that in the limit as $n \to \infty$, δ_2 vanishes, as does the second summand. Hence we have

$$\begin{aligned}
&\liminf_{n\to\infty} -\frac{1}{n} \log \Pr(d(X^n, \hat{X}^n) > \Delta) \\
&\geq \liminf_{n\to\infty} \min_{Q_X} \max_{\substack{Q_{Z|X} \in \\ \mathcal{C}^n(Q_X, \mathcal{Z})}} \min_{Q_Y} \max_{f \in \mathcal{F}} \min_{Q_{XYZ}} \\
&\qquad G^n\left[Q_{XYZ}, P_{XY}, f, d, \Delta, R\right] \\
&\geq \inf_{Q_X} \sup_{Q_{Z|X}} \inf_{Q_Y} \sup_{f \in \mathcal{F}} \inf_{Q_{XYZ}} G\left[Q_{XYZ}, P_{XY}, f, \Delta, R\right],
\end{aligned}$$

where the final line followed from application of Lemma 51. □

B.4 Gaussian Type-classes

For the Gaussian case ($\mathcal{X} = \mathcal{Y} = \mathbb{R}$), we need the following definitions[3]. These are a modification of the Gaussian types used by Arikan and Merhav [92]. The difference is that here the type-classes are disjoint and the conditions specifying joint types are independent. This significantly simplifies the subsequence analysis and might prove useful in other applications.

Definition 13. For a given $0 < \epsilon < 1$ and $\sigma_X^2 > 0$, a Gaussian type-class $T^{\epsilon}_{\sigma_X^2}$ is defined as the set of n-sequences

$$T^{\epsilon}_{\sigma_X^2} = \left\{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}^t\mathbf{x} - n\sigma_X^2| \leq n\epsilon\right\}.$$

For such a type-class, it can be shown (see Appendix B.4) that

$$\left(1 - \frac{2\sigma_X^4}{n\epsilon^2}\right) \exp\left(n\left(h(\sigma_X^2) - \frac{\epsilon}{2\sigma_X^2}\right)\right) \leq \mathrm{Vol}(T^{\epsilon}_{\sigma_X^2}) \leq \exp\left(n\left(h(\sigma_X^2) + \frac{\epsilon}{2\sigma_X^2}\right)\right). \tag{B.32}$$

Similarly, for a given $0 < \epsilon < 1$ and covariance matrix

$$K = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_X & \sigma_Y^2 \end{bmatrix},$$

with non-zero variances, a joint Gaussian type-class T^{ϵ}_K is defined as the set of

[3]For more than two jointly Gaussian random variables, these definitions can be extended in the obvious way.

pairs of n-sequences

$$\begin{aligned} T_K^\epsilon = \big\{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n : |\mathbf{x}^t\mathbf{x} - n\sigma_X^2| \le n\epsilon \\ |\mathbf{y}^t\mathbf{y} - n\sigma_Y^2| \le n\epsilon \\ |\mathbf{x}^t\mathbf{y} - \rho\sqrt{\mathbf{x}^t\mathbf{x}\mathbf{y}^t\mathbf{y}}| \le \epsilon\sqrt{\mathbf{x}^t\mathbf{x}\mathbf{y}^t\mathbf{y}}\}. \end{aligned}$$

This set has the corresponding volume bound

$$\mathrm{Vol}(T_K^\epsilon) \le \exp\left(n\left(h(K) + o_\epsilon(1)\right)\right), \tag{B.33}$$

where we use $o_\epsilon(1)$ to denote a quantity $g(\epsilon) > 0$ having the property that $\lim_{\epsilon\to 0} g(\epsilon) = 0$.

Furthermore, for a given $\mathbf{x} \in T^\epsilon_{\sigma_X^2}$, we define the conditional Gaussian type-class $T_K^\epsilon(\mathbf{x})$ as the $\mathbf{x}$-*set* of n-sequences

$$T_K^\epsilon(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n : (\mathbf{x},\mathbf{y}) \in T_K^\epsilon\}.$$

For this set one can show (see Appendix B.4) that

$$\begin{aligned} &\mathrm{Vol}(T_K^\epsilon(\mathbf{x})) \\ &\ge \left(1 - \frac{1}{no_\epsilon(1)} + o_\epsilon(1)\right)\exp(n(h(K_{Y|X}) - \tilde{f}_\epsilon)). \end{aligned} \tag{B.34}$$

where $\tilde{f}_\epsilon$ is an $o_\epsilon(1)$ term whose value is determined in the proof. In Appendix B.4 we show for a Gaussian distribution $f_K(\cdot,\cdot)$, if $(\mathbf{x},\mathbf{y}) \in T^\epsilon_{\tilde{K}}$, where $\tilde{K}$ is any positive definite covariance matrix, then

$$f_{XY}^n(\mathbf{x},\mathbf{y}) \le \exp\Big(-n\Big(D(\tilde{K}||K) + h(Q_{\tilde{K}}) - o_\epsilon(1)\Big)\Big). \tag{B.35}$$

The analysis for the Gaussian case requires that we "quantize" the space of 3×3 covariance matrices. Unlike discrete memoryless sources, Gaussian sources require use of a "bounding box" to limit the number of types. To this end, fix $0 <$

$M_L < 1$ and $M_U > M_L$; both will be chosen later. For a fixed $0 < \epsilon < M_L$ define $\sigma^2(i) = M_L + 2i\epsilon$ and for i, j, ϵ, given define $\eta_{ij}(r) = \sqrt{\sigma^2(i)\sigma^2(j)}(-1 + 2\epsilon(r-1))$. We will consider type-classes indexed by matrices of the form

$$K_\epsilon(i,j,k,r,s,t) = \begin{bmatrix} \sigma^2(i) & \eta_{ij}(r) & \eta_{ik}(s) \\ \eta_{ij}(r) & \sigma^2(j) & \eta_{jk}(t) \\ \eta_{ik}(s) & \eta_{jk}(t) & \sigma^2(k) \end{bmatrix}$$

and $i, j, k, r, s, t \geq 1$; note that not all of these matrices are positive semidefinite.

We let $\mathcal{P}_X^\epsilon = \{i : \exists \mathbf{x} \in T^\epsilon_{\sigma^2(i)}$ with $\mathbf{x}^t\mathbf{x} \leq M_U\}$ and similarly $\mathcal{P}^\epsilon_{XYZ} = \{(i,j,k,r,s,t) : \exists(\mathbf{x},\mathbf{y},\mathbf{z}) \in T_{K(i,j,k,r,s,t)}$ with $\mathbf{x}^t\mathbf{x} \leq nM_U$ and $\mathbf{y}^t\mathbf{y} \leq nM_U$ and $\mathbf{z}^t\mathbf{z} \leq nM_U\}$, where $M_U \gg M_L$. With $\mathcal{S}_L = \{(\mathbf{x},\mathbf{y},\mathbf{z}) : \mathbf{x}^t\mathbf{x} \leq n(M_L+\epsilon)$ or $\mathbf{y}^t\mathbf{y} \leq n(M_L+\epsilon)$ or $\mathbf{z}^t\mathbf{z} \leq n(M_L+\epsilon)\}$, $\mathcal{S}_U = \{(\mathbf{x},\mathbf{y},\mathbf{z}) : \mathbf{x}^t\mathbf{x} > nM_U$ or $\mathbf{y}^t\mathbf{y} > nM_U$ or $\mathbf{z}^t\mathbf{z} > nM_U\}$, the union of the shells $T^\epsilon_{K(i,j,k,r,s,t)}$, and the set $\mathcal{S}_L$ cover $\mathbb{R}^{3n}$ entirely and we define $\mathcal{R}^{3n} = \mathbb{R}^{3n} \backslash (\mathcal{S}_L \cup \mathcal{S}_U)$. We denote by $\nu(\mathbf{x})$ the *index* of the shell containing the string $\mathbf{x}$, i.e. $\mathbf{x} \in T^\epsilon_{\sigma^2(\nu(\mathbf{x}))}$, which is uniquely defined almost everywhere in $\mathcal{R}^{3n}$.

Proof of (B.32)

Let $X \sim \mathcal{N}(0, \sigma_X^2)$. Then

$$\begin{aligned} 1 &\geq \int_{T^\epsilon_{\sigma_X^2}} (2\pi\sigma_X^2)^{-\frac{n}{2}} \exp\left(-\frac{\mathbf{x}^t\mathbf{x}}{2\sigma_X^2}\right) d\mathbf{x} \\ &\geq \int_{T^\epsilon_{\sigma_X^2}} (2\pi\sigma_X^2)^{-\frac{n}{2}} \exp\left(-\frac{n(\sigma_X^2+\epsilon)}{2\sigma_X^2}\right) d\mathbf{x} \\ &= \exp\left(-n\left(\frac{1}{2}\log(2\pi\sigma_X^2) + \frac{1}{2}\right) - \frac{n\epsilon}{2\sigma_X^2}\right) \mathrm{Vol}(T^\epsilon_{\sigma_X^2}), \end{aligned}$$

which gives the upper bound. For the lower bound,

$$\begin{aligned}\Pr(T^{\epsilon}_{\sigma^2_X}) &= \int_{T^{\epsilon}_{\sigma^2_X}} (2\pi\sigma_X^2)^{-\frac{n}{2}} \exp\left(-\frac{\mathbf{x}^t\mathbf{x}}{2\sigma_X^2}\right) d\mathbf{x} \\ &\leq \int_{T^{\epsilon}_{\sigma^2_X}} (2\pi\sigma_X^2)^{-\frac{n}{2}} \exp\left(-\frac{n(\sigma_X^2-\epsilon)}{2\sigma_X^2}\right) d\mathbf{x}\end{aligned}$$

$$= \mathrm{Vol}(T^{\epsilon}_{\sigma^2_X}) \exp\left(-n\left(\frac{1}{2}\log(2\pi e\sigma_X^2)\right) + \frac{n\epsilon}{2\sigma_X^2}\right).$$

Conversely, by Chebyshev's inequality

$$\begin{aligned}1 - \Pr(T^{\epsilon}_{\sigma^2_X}) &= \Pr\left(|\mathbf{x}^t\mathbf{x} - n\sigma_X^2| > n\epsilon\right) \\ &\leq \mathbb{E}\left[\frac{(\mathbf{x}^t\mathbf{x} - n\sigma_X^2)^2}{n^2\epsilon^2}\right] \\ &= \frac{2\sigma_X^4}{n\epsilon^2}\end{aligned}$$

Combining these two calculations gives the lower bound. □

Proof of (B.34)

Let $\mathbf{x} \in T^{\epsilon}_{\sigma^2_X}$, then

$$\begin{aligned}T_K^{\epsilon}(\mathbf{x}) = \Big\{\mathbf{y} \in \mathbb{R}^n : &|\mathbf{y}^t\mathbf{y} - n\sigma_Y^2| \leq n\epsilon \\ &\left|\frac{\mathbf{y}^t\mathbf{x}}{n} - \rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}\frac{\mathbf{y}^t\mathbf{y}}{n}}\right| \leq \epsilon\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}\frac{\mathbf{y}^t\mathbf{y}}{n}}\Big\}.\end{aligned}$$

By the triangle inequality

$$\left|\frac{\mathbf{y}^t\mathbf{x}}{n} - \rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}\frac{\mathbf{y}^t\mathbf{y}}{n}}\right| \leq \left|\frac{\mathbf{y}^t\mathbf{x}}{n} - \rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}}\sigma_Y\right| + \left|\rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}}\sigma_Y - \rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}\frac{\mathbf{y}^t\mathbf{y}}{n}}\right|,$$

whence

$$T_K^\epsilon(\mathbf{x}) \supset A(\mathbf{x}) \triangleq \Big\{\mathbf{y} \in \mathbb{R}^n : |\mathbf{y}^t\mathbf{y} - n\sigma_Y^2| \le n\epsilon \\ \left|\frac{\mathbf{y}^t\mathbf{x}}{n} - \rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}}\sigma_Y\right| \le \frac{\epsilon}{2}\sqrt{\sigma_Y^2 - \epsilon}\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}} \\ \left|\rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}}\sigma_Y - \rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}\frac{\mathbf{y}^t\mathbf{y}}{n}}\right| \le \frac{\epsilon|\rho|}{2}\sqrt{\sigma_Y^2 - \epsilon}\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}}\Big\}.$$

Let $\mathbf{V}$ be a Gaussian random vector whose law is $\mathcal{N}(0, I\sigma_Y^2(1-\rho^2))$, and let $\mathbf{Y} = \frac{\rho\sigma_Y}{\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}}}\mathbf{x} + \mathbf{V}$. Applying the union bound gives

$$\Pr(A(\mathbf{x})^c) \le \Pr\left(|\mathbf{Y}^t\mathbf{Y} - n\sigma_Y^2| > n\epsilon\right) + \Pr\left(\left|\frac{\mathbf{Y}^t\mathbf{x}}{n} - \rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}}\sigma_Y\right| > \frac{\epsilon}{2}\sqrt{\sigma_Y^2 - \epsilon}\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}}\right) \\ + \Pr\left(\left|\rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}}\sigma_Y - \rho\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}\frac{\mathbf{Y}^t\mathbf{Y}}{n}}\right| > \frac{\epsilon|\rho|}{2}\sqrt{\sigma_Y^2 - \epsilon}\sqrt{\frac{\mathbf{x}^t\mathbf{x}}{n}}\right).$$

The event in the third probability on the right is equivalent to

$$\left\{\left|\frac{\mathbf{Y}^t\mathbf{Y}}{n} - \sigma_Y^2 - \frac{\epsilon^2(\sigma_Y^2 - \epsilon)}{4}\right| > \epsilon\sigma_Y\sqrt{\sigma_Y^2 - \epsilon}\right\}.$$

Using this fact and bounding each of the probabilities using Chebyshev's inequality yields

$$\begin{aligned}\Pr(A(\mathbf{x})^c) &\le \mathbb{E}\left[\frac{(\mathbf{Y}^t\mathbf{Y} - n\sigma_Y^2)^2}{n^2\epsilon^2}\right] + \mathbb{E}\left[\frac{(\frac{\mathbf{Y}^t\mathbf{x}}{n}\sqrt{\frac{n}{\mathbf{x}^t\mathbf{x}}} - \rho\sigma_Y)^2}{\epsilon^2(\sigma_Y^2 - \epsilon)/4}\right] \\ &\quad + \mathbb{E}\left[\frac{(\frac{\mathbf{Y}^t\mathbf{Y}}{n} - \sigma_Y^2 - \epsilon^2(\sigma_Y^2 - \epsilon)/4)^2}{\epsilon^2\sigma_Y^2(\sigma_Y^2 - \epsilon)}\right] \\ &= \frac{2\sigma_Y^4}{n\epsilon^2} + \frac{\sigma_Y^2(1-\rho^2)}{n(\epsilon^2(\sigma_Y^2 - \epsilon))/4} + \frac{2\sigma_Y^4}{n\epsilon^2\sigma_Y^2(\sigma_Y^2 - \epsilon)} + \frac{\epsilon^2(\sigma_Y^2 - \epsilon)}{16\sigma_Y^2} \\ &= \frac{1}{no_\epsilon(1)} + o_\epsilon(1) \qquad (*)\end{aligned}$$

To bound the volume we note that under the law above

$$\begin{aligned}\Pr(A(\mathbf{x})) &= \int_{A(\mathbf{x})} f_{\mathbf{V}}\Big(\mathbf{y} - \frac{\sqrt{n}\rho\sigma_Y\mathbf{x}}{\sqrt{\mathbf{x}^t\mathbf{x}}}\Big)\, d\mathbf{y} \\ &= \int_{A(\mathbf{x})} (2\pi\sigma_Y^2(1-\rho^2))^{-n/2} \exp\left(\frac{-\sum_i (y_i - \frac{\sqrt{n}\rho\sigma_Y}{\sqrt{\mathbf{x}^t\mathbf{x}}}x_i)^2}{2(\sigma_Y^2(1-\rho^2))}\right) d\mathbf{y}.\end{aligned}$$

We can get an upper bound on the density by lower bounding the summand in the exponent

$$\begin{aligned}\sum_i \Big(y_i - \frac{\sqrt{n}\rho\sigma_Y}{\sqrt{\mathbf{x}^t\mathbf{x}}}x_i\Big)^2 &= \mathbf{y}^t\mathbf{y} - 2\rho\sigma_Y\frac{\sqrt{n}}{\sqrt{\mathbf{x}^t\mathbf{x}}}\mathbf{y}^t\mathbf{x} + n\rho^2\sigma_Y^2 \\ &\geq n(\sigma_Y^2 - \epsilon) + n\rho^2\sigma_Y^2 - 2\rho\sigma_Y n\big(\rho\sigma_Y + \operatorname{sgn}(\rho)\epsilon/2\sqrt{\sigma_Y^2 - \epsilon}\big) \\ &= n(\sigma_Y^2(1-\rho^2) - f_\epsilon(\rho, \sigma_Y))\end{aligned}$$

where $f_\epsilon(\rho, \sigma_Y) = \epsilon(1 + \rho\operatorname{sgn}(\rho)\sigma_Y\sqrt{\sigma_Y^2 - \epsilon})$ goes to zero with ϵ. Thus

$$\begin{aligned}\Pr(A(\mathbf{x})) &\leq \operatorname{Vol}(A(\mathbf{x}))\exp\left(-n\left(\frac{1}{2}\log(2\pi\sigma_Y^2(1-\rho^2)) - \frac{1}{2} - \tilde{f}_\epsilon(\rho, \sigma_Y)\right)\right) \\ &= \operatorname{Vol}(A(\mathbf{x}))\exp\left(-n\left(\frac{1}{2}\log(2\pi e\sigma_Y^2(1-\rho^2)) - \tilde{f}_\epsilon(\rho, \sigma_Y)\right)\right),\end{aligned}$$

where $\tilde{f}_\epsilon = f_\epsilon/(2(\sigma_Y^2(1-\rho^2)))$. Combining this with $(*)$ and using the fact that $\operatorname{Vol}(T_K^\epsilon(\mathbf{x})) \geq \operatorname{Vol}(A(\mathbf{x}))$ gives the result. $\square$

Proof of (B.35)

Let $(X, Y) \sim \mathcal{N}(0, K)$ and $(\mathbf{x}, \mathbf{y}) \in T_{\tilde{K}}^\epsilon$. Then

$$\begin{aligned}f(\mathbf{x}, \mathbf{y}) &= [(2\pi)^2|K|]^{-\frac{n}{2}} \\ &\times \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{\mathbf{x}^t\mathbf{x}}{\sigma_X^2} + \frac{\mathbf{y}^t\mathbf{y}}{\sigma_Y^2} - \frac{2\rho\mathbf{x}^t\mathbf{y}}{\sigma_X\sigma_Y}\right)\right).\end{aligned}$$

Applying the bounds from the definition of $T_{\tilde{K}}^\epsilon$ allows us to continue the inequality with

$$\begin{aligned}&\leq \exp\Bigg(-\frac{n}{2}\log((2\pi)^2|K|) - \frac{1}{2(1-\rho^2)}\Big(\frac{n(\tilde{\sigma}_X^2 - \epsilon)}{\sigma_X^2} \\ &+ \frac{n(\tilde{\sigma}_Y^2 - \epsilon)}{\sigma_Y^2} - \frac{2\rho n\sqrt{(\tilde{\sigma}_X^2 + \epsilon)(\tilde{\sigma}_Y^2 + \epsilon)}(\tilde{\rho} + \operatorname{sgn}(\rho)\epsilon)}{\sigma_X\sigma_Y}\Big)\Bigg).\end{aligned}$$

For $f_\epsilon(\sigma_X, \sigma_Y, \tilde{\sigma}_X, \tilde{\sigma}_Y, \rho, \tilde{\rho})$ (which goes to zero with ϵ), we can write

$$\leq \exp -n\Bigg(\frac{1}{2}\Big(\log((2\pi)^2|K|) + \frac{\tilde{\sigma}_X^2}{\sigma_X^2(1-\rho^2)} + \frac{\tilde{\sigma}_Y^2}{\sigma_Y^2(1-\rho^2)} - \frac{2\rho\tilde{\sigma}_X\tilde{\sigma}_Y\tilde{\rho}}{\sigma_X\sigma_Y(1-\rho^2)} - f_\epsilon(\sigma_X, \sigma_Y, \tilde{\sigma}_X, \tilde{\sigma}_Y, \rho, \tilde{\rho})\Big)\Bigg).$$

Finally, using the identity

$$D(\tilde{K}||K) = \frac{1}{2}\left(\log\frac{|K|}{|\tilde{K}|} + \mathrm{Tr}(K^{-1}\tilde{K}) - 2\right)$$

gives

$$f(\mathbf{x}, \mathbf{y}) \leq \exp -n\Big(D(\tilde{K}||K) + \frac{1}{2}\log(2\pi e)^2|\tilde{K}| - f_\epsilon(\sigma_X, \sigma_Y, \tilde{\sigma}_X, \tilde{\sigma}_Y, \rho, \tilde{\rho})\Big) \quad \square$$

B.5 Proof of Theorem 19

B.5.1 Scheme

Let $\epsilon > 0$ and M_L, M_U as defined in Appendix B.4. For each blocklength n, and for each shell of n-length $\mathbf{x}$ sequences, $T^\epsilon_{\sigma^2(i)}$ we choose a Gaussian test channel. The test channel is specified by selecting integers $k(i)$ and $s(i)$ (such that $\sigma^2(k(i)) < M_U$) so that if $X \sim \mathcal{N}(0, \sigma^2(i))$ is the input to the channel then $(X, Z) \sim \mathcal{N}(0, \overline{\sigma^2(i)})$; where the bar applied to a scalar results in

$$\overline{\sigma^2(i)} = \begin{bmatrix} \sigma^2(i) & \eta_{i,k(i)}(s(i)) \\ \eta_{i,k(i)}(s(i)) & \sigma^2(k(i)) \end{bmatrix}. \tag{B.36}$$

The codebook for the ith shell of $\mathbf{x}$ sequences is a randomly chosen set of codewords, $B^n(i)$, selected in the following way. The size of $B^n(i)$ is an integer

satisfying

$$\exp(n(I_{\overline{\sigma^2(i)}}(X;Z)+2g_\epsilon)) \leq |B^n(i)| \leq \exp(n(I_{\overline{\sigma^2(i)}}(X;Z)+3g_\epsilon)) \tag{B.37}$$

where $g_\epsilon = \tilde{f}_\epsilon + \epsilon/2\sigma^2(k(i))$ (c.f. (B.34)) and the codewords are chosen uniformly from the shell $T_{\sigma^2(k(i))}$.

For $\mathbf{x} \in T^\epsilon_{\sigma^2(i)}$, define $Z(\mathbf{x}) : T^\epsilon_{\sigma^2(i)} \to B^n(i)$ as follows. We can cover the shell $T^\epsilon_{\sigma^2(i)}$ with conditional type-classes $T^\epsilon_{\overline{\sigma^2(i)}}(B^n(i)[j])$, where $B^n(i)[j]$ is the jth codeword. This covering induces a partition of sequences in $T^\epsilon_{\sigma^2(i)}$, with the partition being based on the set of possible codewords in $B^n(i)$ that have the correct joint type with the sequences. For each set generated by this partition, we chose the codeword for that set uniformly among the covering conditional type-classes. For the sets not covered by any class, the codeword is selected at random from $B^n(i)$. We define $Z^n = Z(X^n)$. Finally, let the encoder's message set be defined as $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2$, where

$$\mathcal{M}_1 = \{1, \ldots, M_1 \triangleq \exp(nR)\}, \mathcal{M}_2 = \{1, 2, \ldots, |\mathcal{P}^\epsilon_X|\}.$$

Operation of the Encoder: To encode a sequence $\mathbf{x} \in T^\epsilon_{\sigma^2(i)}$, the encoder sends i, the "type" of $\mathbf{x}$ and an index, $U(Z(\mathbf{x}))$, of the codeword $Z(\mathbf{x})$. If $\log|B^n(i)| \geq nR$ we use random binning of the codewords, and $U(Z(\mathbf{x}))$ denotes the element of $\mathcal{M}_1$ to which $Z(\mathbf{x})$ is mapped. For sequences with $\mathbf{x}^t\mathbf{x} \notin (n(M_L+\epsilon), nM_U]$ the encoder declares an error. The encoder can be expressed mathematically as

$$\psi(\mathbf{x}) = (U(Z(\mathbf{x})), i) \text{ for } \mathbf{x} \in T^\epsilon_{\sigma^2(i)} \tag{B.38}$$

Operation of the Decoder: The decoder operates in a two-step manner. First it attempts to recover the codeword Z^n:

1. If $\log|B^n(i)| < nR$ then Z^n can be decoded without error,
2. If $\log|B^n(i)| \geq nR$ the decoder receives a bin index and uses the side information to pick the $\mathbf{z}$ from the bin by searching for a $\hat{\mathbf{z}}$ in the received bin so that among all $\tilde{\mathbf{z}}$ in the bin, $\rho^2_{\tilde{\mathbf{z}},\mathbf{y}} < \rho^2_{\hat{\mathbf{z}},\mathbf{y}}$. If there is no such $\hat{\mathbf{z}}$, the encoder picks uniformly at random from the bin.

Let

$$\varphi_1(l,i,\mathbf{y}) = \begin{cases} \hat{\mathbf{z}} & \hat{\mathbf{z}} \in \mathrm{Bin}(l) \text{ and } \forall \tilde{\mathbf{z}} \neq \hat{\mathbf{z}} \in \mathrm{Bin}(l), \\ & \quad \rho^2_{\tilde{\mathbf{z}},\mathbf{y}} < \rho^2_{\hat{\mathbf{z}},\mathbf{y}} \\ \text{any } \tilde{\mathbf{z}} & \text{if no such } \hat{\mathbf{z}} \in \mathrm{Bin}(l) \end{cases} \tag{B.39}$$

where $\mathrm{Bin}(l) = \{\mathbf{z} : \mathbf{z} \in B^n(i) \text{ and } U(\mathbf{z}) = l\}$ denotes the set of codewords that are assigned to bin l. The *marginal* types i, j of $\mathbf{x}$ and $\mathbf{y}$ are known, and for each pair i, j we choose an estimation function. We restrict our attention to estimation functions that are linear in the side information and the codeword, i.e. $\lambda_{i,j}(y,z) = \alpha(i,j)y + \beta(i,j)z$, where $\alpha(i,j) = \nu\epsilon, \beta(i,j) = \kappa\epsilon$ for integers ν, κ so that $\alpha(i,j), \beta(i,j) \in [-M_\lambda, M_\lambda]$. α and γ will be optimized later and $M_\lambda > 0$ is an arbitrary positive constant. For the second step the decoder uses the estimation function, λ, to combine the side information $\mathbf{y}$ with codeword $\mathbf{z}$ to give the reproduction $\hat{\mathbf{x}}$. This is expressed mathematically as

$$\begin{aligned} &\varphi(l,i,\mathbf{y}) = \hat{\mathbf{x}} \\ &\text{s.t. } \hat{\mathbf{x}}_m = \alpha(i,\nu(\mathbf{y}))\mathbf{y}_m + \beta(i,\nu(\mathbf{y}))\varphi_1(l,i,\mathbf{y})_m. \end{aligned} \tag{B.40}$$

B.5.2 Key events

The following subsets of $\mathbb{R}^{3n}$ will be of interest.

$$\begin{aligned}
\mathcal{E}_b &= \Big\{(\mathbf{x},\mathbf{y},\mathbf{z}) \in \mathcal{R}^{3n} : \mathbf{z} \in T^{\epsilon}_{\overline{\sigma^2(\nu(\mathbf{x}))}}(\mathbf{x}), \\
&\qquad \frac{1}{n}\|\mathbf{x} - \lambda_{\nu(\mathbf{x}),\nu(\mathbf{y})}(\mathbf{y},\mathbf{z})\|_2^2 < \Delta, \log|B^n(\nu(\mathbf{x}))| \geq nR\Big\} \\
\mathcal{E}_c &= \Big\{(\mathbf{x},\mathbf{y},\mathbf{z}) \in \mathcal{R}^{3n} : \mathbf{z} \notin T^{\epsilon}_{\overline{\sigma^2(\nu(\mathbf{x}))}}(\mathbf{x})\Big\} \\
\mathcal{E}_d &= \Big\{(\mathbf{x},\mathbf{y},\mathbf{z}) \in \mathcal{R}^{3n} : \mathbf{z} \in T^{\epsilon}_{\overline{\sigma^2(\nu(\mathbf{x}))}}(\mathbf{x}), \\
&\qquad \frac{1}{n}\|\mathbf{x} - \lambda_{\nu(\mathbf{x}),\nu(\mathbf{y})}(\mathbf{y},\mathbf{z})\|_2^2 \geq \Delta\Big\}.
\end{aligned}$$

On $\mathcal{E}_b$, the distortion constraint is violated only if there is a decoding error. On $\mathcal{E}_c$ we say there is a "covering" error: the encoder cannot find a codeword with the desired joint type with the source sequence. On $\mathcal{E}_d$, the distortion constraint will be violated even if the codeword is decoded correctly by the decoder.

For $\mathbf{x} \in T^{\epsilon}_{\sigma^2(i)}$, F is defined to be the event that there exists $\tilde{\mathbf{z}} \in B^n(i)$ such that $\tilde{\mathbf{z}} \in T^{n}_{\overline{\sigma^2(i)}}(\mathbf{x})$.

B.5.3 Error Probability Calculation

We will first state several useful lemmas, which are "Gaussian versions" of the discrete memoryless Wyner-Ziv lemmas.

Lemma 52. *Let $X^n, Y^n, Z^n = Z(X^n)$ be generated according to our scheme and suppose that $A \subset (\mathcal{E}_c)^c \cap \mathcal{R}^{3n}$. Then*

$$\begin{aligned}
&\Pr((X^n,Y^n,Z^n) \in A) \\
&\qquad \leq \int_A f^n_{XY}(\mathbf{x},\mathbf{y}) \frac{1}{\mathrm{Vol}(T^{\epsilon}_{\overline{\sigma^2(\nu(\mathbf{x}))}}(\mathbf{x}))}\, d\mathbf{xyz}. \qquad \text{(B.41)}
\end{aligned}$$

Proof. For the $\mathbf{x}, \mathbf{y}, \mathbf{z} \in A$ in this lemma, $\{X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}\}$ implies that the event F has occurred. Let A_{XY} be the projection of A onto XY space, i.e. $A_{XY} = \{(\mathbf{x}, \mathbf{y}) : (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in A \text{ for some } \mathbf{z}\}$ and $A^{\mathbf{x},\mathbf{y}} = \{\mathbf{z} : (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in A\}$. Then

$$\begin{aligned}
&\Pr((X^n,Y^n, Z^n) \in A) \\
&\quad = \Pr((X^n, Y^n, Z^n) \in A, F) \\
&\quad = \int_{A_{XY}} f_{XY}^n(\mathbf{x}, \mathbf{y}) \Pr(F | X^n = \mathbf{x}, Y^n = \mathbf{y}) \\
&\qquad \times \Pr(Z^n \in A^{\mathbf{x},\mathbf{y}} | X^n = \mathbf{x}, Y^n = \mathbf{y}, F) d\mathbf{x}\mathbf{y} \\
&\quad \leq \int_{A_{XY}} f_{XY}^n(\mathbf{x}, \mathbf{y}) \\
&\qquad \times \Pr(Z^n \in A^{\mathbf{x},\mathbf{y}} | X^n = \mathbf{x}, Y^n = \mathbf{y}, F) d\mathbf{x}\mathbf{y} \\
&\quad = \int_{A_{XY}} f_{XY}^n(\mathbf{x}, \mathbf{y}) \int_{A^{\mathbf{x},\mathbf{y}}} f_{Z|X,Y,F}(\mathbf{z}|\mathbf{x}, \mathbf{y}) d\mathbf{z} d\mathbf{x}\mathbf{y} \\
&\quad = \int_A f_{XY}^n(\mathbf{x}, \mathbf{y}) \frac{1}{\mathrm{Vol}(T^{\epsilon}_{\sigma^2(\nu(\mathbf{x}))}(\mathbf{x}))} d\mathbf{x}\mathbf{y}\mathbf{z}
\end{aligned}$$

where in the final line we used that conditional on F and $X^n = \mathbf{x}$, Z^n is uniformly distributed over $T^{\epsilon}_{\sigma^2(\nu(\mathbf{x}))}(\mathbf{x})$ and independent of Y. □

Lemma 53. *Let $X^n, Y^n, Z^n = Z(X^n)$ be generated according to our scheme. Then for n sufficiently large*

$$\Pr((X^n,Y^n, Z^n) \in \mathcal{E}_c) \leq |\mathcal{P}_X^{\epsilon}| \exp(-\exp(no_{\epsilon}(1))) \tag{B.42}$$

Proof. For $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{E}_c$, $\{X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}\}$ implies that the event F^c has

occurred. Thus

$$\begin{aligned}
&\Pr((X^n,Y^n, Z^n) \in \mathcal{E}_c) \\
&\quad = \Pr((X^n, Y^n, Z^n) \in \mathcal{E}_c, F^c) \\
&\quad \leq \sum_{T^\epsilon_{\sigma^2(i)} \in \mathcal{P}^\epsilon_X} \Pr(X^n \in T^\epsilon_{\sigma^2(i)}) \Pr(F^c | X^n \in T^\epsilon_{\sigma^2(i)}) \\
&\qquad \times \Pr((X^n, Y^n, Z^n) \in \mathcal{E}_c | X^n \in T^\epsilon_{\sigma^2(i)}, F^c) \\
&\quad \leq \sum_{T^\epsilon_{\sigma^2(i)} \in \mathcal{P}^\epsilon_X} \Pr(F^c | X^n \in T^\epsilon_{\sigma^2(i)}) \Pr(X^n \in T^\epsilon_{\sigma^2(i)}).
\end{aligned}$$

$\Pr(F^c | X^n \in T^\epsilon_{\sigma^2(i)})$ is the probability that there is no $\tilde{\mathbf{z}} \in B^n(i)$ so that $\tilde{\mathbf{z}} \in T^\epsilon_{\overline{\sigma(i)}}(X^n)$. We will now give an upper bound on this probability using the properties of the codeword set. Let $m = |B^n(i)|$ and $B^n(i)[j]$ be the jth codeword in the set $B^n(i)$. Then

$$\begin{aligned}
\Pr(F^c | X^n \in T^\epsilon_{\sigma^2(i)}) &= \prod_{j=1}^{m} \Pr(B^n(i)[j] \notin T^\epsilon_{\overline{\sigma^2(i)}}(X^n)) \\
&= \prod_{j=1}^{m} [1 - \Pr(B^n(i)[j] \in T^\epsilon_{\overline{\sigma^2(i)}}(X^n))] \\
&= \left(1 - \frac{\mathrm{Vol}(T^\epsilon_{\overline{\sigma^2(i)}}(X^n))}{\mathrm{Vol}(T^\epsilon_{\sigma^2(k(i))})}\right)^m \\
&\leq \exp\left(-\frac{\mathrm{Vol}(T^\epsilon_{\overline{\sigma^2(i)}}(X^n))}{\mathrm{Vol}(T^\epsilon_{\sigma^2(k(i))})} m\right)
\end{aligned}$$

where the last line followed by applying the inequality $(1-t)^m \leq \exp(-tm)$. Next, using (B.32) and (B.34) to bound the volume of the shells,

$$\begin{aligned}
&\Pr(F^c | X^n \in T^\epsilon_{\sigma^2(i)}) \\
&\leq \exp\left(-\left(1 - \frac{1}{no_\epsilon(1)} - o_\epsilon(1)\right) m \exp\left(-n\left(I_{\overline{\sigma^2(i)}}(X;Z) + g_\epsilon\right)\right)\right) \\
&\leq \exp(-\exp(no_\epsilon(1)))
\end{aligned}$$

where the final line followed by substitution our choice of m from (B.37). $\square$

Lemma 54. *For any positive definite covariance matrix K,*

$$\begin{aligned}&\Pr(T_K^\epsilon \cap (\mathcal{E}_d \cup \mathcal{E}_b))\\ &\leq \exp\Big(-n\Big(D(K||\bar{K}) - o_\epsilon(1) - \delta_p\Big)\Big)\end{aligned} \tag{B.43}$$

where $\bar{K}$ is defined in (4.16) *and*

$$\text{and } \delta_p = \frac{1}{n}\log\left(1 - \frac{1}{no_\epsilon(1)} - o_\epsilon(1)\right)^{-1}$$

Proof. Lemma 52 gives an upper bound for the probability density on $\mathcal{E}_b$ and $\mathcal{E}_d$. Applying this lemma with (B.32) and (B.35), we get

$$\begin{aligned}\Pr(T_K^\epsilon \cap (\mathcal{E}_d \cup \mathcal{E}_b)) &\leq \int_{T_K^\epsilon} f_\Sigma^n(\mathbf{x},\mathbf{y})\frac{1}{\mathrm{Vol}(T^\epsilon_{\sigma^2(\nu(\mathbf{x}))}(\mathbf{x}))}d\mathbf{xyz}\\ &\leq \int_{T_K^\epsilon} \exp\Big(-n\Big(D(K_{XY}||\Sigma) + h(K_{XY}) - o_\epsilon(1)\Big)\Big)\\ &\quad\times\left(1 - \frac{1}{no_\epsilon(1)} - o_\epsilon(1)\right)^{-1}\exp(-n(h(K_{Z|X}) - o_\epsilon(1)))d\mathbf{xyz}\\ &= \mathrm{Vol}(T_K^\epsilon)\exp\Big(-n\Big(D(K_{XY}||\Sigma) + h(K_{XY}) - o_\epsilon(1)\Big)\Big)\\ &\quad\times\left(1 - \frac{1}{no_\epsilon(1)} - o_\epsilon(1)\right)^{-1}\exp(-n(h(K_{Z|X}) - o_\epsilon(1)))d\mathbf{xyz}.\end{aligned}$$

Bounding the volume term using (B.33) and applying the identity

$$D(K||\bar{K}) = D(K_{XY}||\Sigma) + h(K_{Z|X}) - h(K_{Z|XY})$$

gives the result. □

Lemma 55. *Let $\mathbf{y}, \mathbf{z}$ be two strings with empirical correlation $\rho_{\mathbf{z},\mathbf{y}}$ and let*

$$A(\mathbf{z},\mathbf{y}) = \{\tilde{\mathbf{z}} \in T^\epsilon_{\sigma^2} : \rho^2_{\tilde{\mathbf{z}},\mathbf{y}} \geq \rho^2_{\mathbf{z},\mathbf{y}}\}.$$

Then

$$\mathrm{Vol}(A(\mathbf{z},\mathbf{y})) \leq 2\exp\left(\frac{n}{2}\log\left(2\pi e\sigma^2(1-\rho^2_{\mathbf{z},\mathbf{y}})\right) + no_\epsilon(1)\right).$$

Proof. The empirical correlation does not change if we scale the vectors, so we may assume that $\mathbf{z}^t\mathbf{z} = \mathbf{y}^t\mathbf{y} = n(\sigma^2 + \epsilon)$. Suppose $\mathbf{z}^t\mathbf{y} \geq 0$, in which case

$$A(\mathbf{z}, \mathbf{y}) = \{\tilde{\mathbf{z}} \in T^{\epsilon}_{\sigma^2} : \rho_{\tilde{\mathbf{z}},\mathbf{y}} \geq \rho_{\mathbf{z},\mathbf{y}}\} \cup \{\tilde{\mathbf{z}} \in T^{\epsilon}_{\sigma^2} : \rho_{\tilde{\mathbf{z}},\mathbf{y}} \leq -\rho_{\mathbf{z},\mathbf{y}}\}$$

and by symmetry the two sets on the right hand side have the same volume and it suffices to consider one of them.

$$\begin{aligned}
\{\tilde{\mathbf{z}} \in T^{\epsilon}_{\sigma^2} : \rho_{\tilde{\mathbf{z}},\mathbf{y}} \geq \rho_{\mathbf{z},\mathbf{y}}\} &= \{\tilde{\mathbf{z}} \in T^{\epsilon}_{\sigma^2} : \frac{\tilde{\mathbf{z}}^t\mathbf{y}}{\sqrt{\tilde{\mathbf{z}}^t\tilde{\mathbf{z}}\mathbf{y}^t\mathbf{y}}} \geq \frac{\mathbf{z}^t\mathbf{y}}{\sqrt{\mathbf{z}^t\mathbf{z}\mathbf{y}^t\mathbf{y}}}\} \\
&= \Big\{\tilde{\mathbf{z}} \in T^{\epsilon}_{\sigma^2} : -2\,\rho_{\mathbf{z},\mathbf{y}}\tilde{\mathbf{z}}^t\mathbf{y} \leq -2\,\rho_{\mathbf{z},\mathbf{y}}\mathbf{z}^t\mathbf{y}\frac{\sqrt{\tilde{\mathbf{z}}^t\tilde{\mathbf{z}}}}{\sqrt{\mathbf{z}^t\mathbf{z}}}\Big\} \\
&\triangleq B(\mathbf{z}, \mathbf{y})
\end{aligned}$$

We now bound the volume of $B(\mathbf{z}, \mathbf{y})$. Let $\mathbf{X} \sim \mathcal{N}(\rho_{\mathbf{x},\mathbf{y}}\mathbf{y}, \sigma^2(1 - \rho^2_{\mathbf{z},\mathbf{y}})I)$. Then

$$\begin{aligned}
1 &\geq \int_{B(\mathbf{z},\mathbf{y})} f_{\mathbf{X}}(\mathbf{x})\,d\mathbf{x} \\
&= \int_{B(\mathbf{z},\mathbf{y})} (2\pi\sigma^2(1 - \rho^2_{\mathbf{z},\mathbf{y}}))^{-n/2} \exp\Big(-\frac{\sum(x_i - \rho_{\mathbf{z},\mathbf{y}}y_i)^2}{2(\sigma^2(1 - \rho^2_{\mathbf{z},\mathbf{y}}))}\Big)\,d\mathbf{x}. \qquad (*)
\end{aligned}$$

To continue we upper bound the summand in the exponent as follows

$$\begin{aligned}
\sum(x_i - \rho_{\mathbf{x},\mathbf{y}}y_i)^2 &= \mathbf{x}^t\mathbf{x} - 2\rho_{\mathbf{z},\mathbf{y}}\mathbf{x}^t\mathbf{y} + \rho^2_{\mathbf{z},\mathbf{y}}\mathbf{y}^t\mathbf{y} \\
&\leq n(\sigma^2 + \epsilon) - 2\rho_{\mathbf{z},\mathbf{y}}\mathbf{x}^t\mathbf{y} + \rho^2_{\mathbf{z},\mathbf{y}}n(\sigma^2 + \epsilon) \\
&\leq n(\sigma^2 + \epsilon) - 2\rho^2_{\mathbf{z},\mathbf{y}}n(\sigma^2 + \epsilon)(1 - o_\epsilon(1)) + \rho^2_{\mathbf{z},\mathbf{y}}n(\sigma^2 + \epsilon) \\
&\leq n(\sigma^2(1 - \rho^2_{\mathbf{z},\mathbf{y}}) + o_\epsilon(1)).
\end{aligned}$$

Substituting the above into (*) gives

$$\begin{aligned}
\mathrm{Vol}(B(\mathbf{z}, \mathbf{y})) &\leq \exp\Big(n\big(\frac{1}{2}\log(2\pi\sigma^2(1 - \rho^2_{\mathbf{z},\mathbf{y}})) + \frac{1}{2} + o_\epsilon(1)\big)\Big) \\
&= \exp\Big(n\big(\frac{1}{2}\log(2\pi e\sigma^2(1 - \rho^2_{\mathbf{z},\mathbf{y}})) + o_\epsilon(1)\big)\Big).
\end{aligned}$$

Observing that an identical argument holds for $\mathbf{z}^t\mathbf{y} \leq 0$ we are done. □

Lemma 56. *Let* $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (\mathcal{E}_c \cup \mathcal{E}_d)^c \cap \mathcal{R}^{3n}$. *Then*

$$\begin{aligned}\Pr\left(\frac{1}{n}\|X^n - \hat{X}^n\|_2^2 > \Delta | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}\right) \\ \leq \exp\left(-n\left(R - J(K) - o_\epsilon(1) - \delta_b\right)^+\right)\end{aligned} \tag{B.44}$$

where $K = K(i, j, k(i), r, s(i), t)$ *is the type containing* $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ *and*

$$J(K) = I_K(X; Z) - I_K(Y; Z),$$
$$\delta_b = \frac{1}{n}\log\left(2\left(1 - \frac{2\sigma^4(k(i))}{n\epsilon^2}\right)^{-1}\right)$$

Moreover, if $\log|B^n(\nu(\mathbf{x}))| < nR$ *then*

$$\Pr\left(\frac{1}{n}\|X^n - \hat{X}^n\|_2^2 > \Delta | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}\right) = 0.$$

Proof. For a given sequence $\mathbf{x}, \mathbf{y}, \mathbf{z}$, let L be the event that $\mathbf{z} \neq \varphi_1(\psi(\mathbf{x}), \mathbf{y})$. Observe that L occurs when the decoder decodes the wrong codeword and that $\Pr\left(\frac{1}{n}\|X^n - \hat{X}^n\|_2^2 > \Delta | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}\right)$ is upper bounded by $\Pr(L | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z})$.

If i is such that $\log|B^n(i)| < nR$, then

$$\Pr(L | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) = 0.$$

For the case in which $\log|B^n(i)| \geq nR$, we invoke the union bound over the slots of the codebook. From the perspective of the decoder, given that the true sequence is $(\mathbf{x}, \mathbf{y}, \mathbf{z})$, a codeword $\tilde{\mathbf{z}}$ is "bad" if it has higher empirical correlation with $\mathbf{y}$ and ends up in the same bin as $\mathbf{z}$. Mathematically,

$$\begin{aligned}&\Pr(L | X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \\ &\leq \sum_{|B^n(i)|} \int_{T^\epsilon_{\sigma^2(k(i))}} \Pr(\rho^2_{\tilde{\mathbf{z}},\mathbf{y}} > \rho^2_{\mathbf{z},\mathbf{y}}, U(\tilde{\mathbf{z}}) = U(\mathbf{z})) |T^\epsilon_{\sigma^2(k(i))}|^{-1} d\tilde{\mathbf{z}}\end{aligned}$$

where we used the fact that the codewords are chosen uniformly from the shell. Now using the result from lemma 55, we can bound the volume of the set of bad $\tilde{\mathbf{z}}$s and use (B.32) to bound the density.

$$
\begin{aligned}
&\leq \sum_{|B^n(i)|} 2\exp\Big(\frac{n}{2}\log 2\pi e\sigma^2(k(i))(1-\rho^2_{\mathbf{z},\mathbf{y}}) + no_\epsilon(1)\Big) \\
&\quad \times \left(1 - \frac{2\sigma^4(k(i))}{n\epsilon^2}\right)^{-1} \exp(-n(h(\sigma^2(i)) + \epsilon/2 + R)) \\
&\leq 2\left(1 - \frac{2\sigma^4(k(i))}{n\epsilon^2}\right)^{-1} \exp\Big(-n\Big(R + I(\mathbf{y};\mathbf{z}) \\
&\quad - I_{\overline{\sigma^2(k(i))}}(X;Z) - o_\epsilon(1)\Big)\Big) \\
&\leq 2\left(1 - \frac{2\sigma^4(k(i))}{n\epsilon^2}\right)^{-1} \exp\Big(-n\Big(R - J(\mathbf{x},\mathbf{y},\mathbf{z}) - o_\epsilon(1)\Big)\Big).
\end{aligned}
$$

Also, since $\Pr(L|X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \leq 1$ we get

$$
\begin{aligned}
&\Pr(L|X^n = \mathbf{x}, Y^n = \mathbf{y}, Z^n = \mathbf{z}) \\
&\qquad \exp\left(-n\left((R - J(\mathbf{x},\mathbf{y},\mathbf{z}) - o_\epsilon(1) - \delta_b)^+\right)\right).
\end{aligned}
$$

□

Lemma 57. *Let δ_p, δ_b be sequences going to zero as $n \to \infty$,*

$$
G^n_\epsilon(K,\Sigma,\lambda,\Delta,R) =
\begin{cases}
D(K||\bar{K}) - o_\epsilon(1) - \delta_p & \mathbb{E}_K[(X - \lambda(Y,Z))^2] \geq \Delta - o_\epsilon(1) \\
D(K||\bar{K}) - o_\epsilon(1) - \delta_p & \\
+\big(R - I_K(X;Z) & \mathbb{E}_K[(X - \lambda(Y,Z))^2] < \Delta - o_\epsilon(1) \\
+I_K(Y;Z) - o_\epsilon(1) - \delta_b\big)^+ & \textit{and } I_K(X;Z) \geq R - o_\epsilon(1) \\
\infty & \textit{otherwise,}
\end{cases}
$$

$$
\pi^n_\epsilon(R,\Delta,\Sigma) = \min_i \max_{k,s} \min_j \max_\lambda \min_{r,t} G^n_\epsilon(K_\epsilon,\Sigma,\lambda,\Delta,R),
$$

and

$$\pi(R,\Delta,\Sigma) = \inf_{\sigma_X}\sup_{\sigma_Z,\rho_{xz}}\inf_{\sigma_Y}\sup_{\lambda}\inf_{\rho_{xy},\rho_{yz}} G_G(K,\Sigma,\lambda,\Delta,R),$$

where K_ϵ is shorthand for $K_\epsilon(i,j,k,r,s,t)$ and K is a covariance matrix with entries $(\sigma_X,\sigma_Y,\sigma_Z,\rho_{xy},\rho_{xz},\rho_{yz})$. Then

$$\liminf_{\epsilon\to 0}\liminf_{n\to\infty}\pi_\epsilon^n(R,\Delta,\Sigma) \geq \pi(R,\Delta,\Sigma).$$

Proof. Let $\delta > 0$. For $\epsilon > 0$ define

$$G_\epsilon(K,\Sigma,\lambda,\Delta,R) =$$
$$\begin{cases} D(K||\bar{K}) - o_\epsilon(1) & \mathbb{E}_K[(X-\lambda(Y,Z))^2] \geq \Delta - o_\epsilon(1) \\ D(K||\bar{K}) - o_\epsilon(1) & \\ +\big(R - I_K(X;Z) & \mathbb{E}_K[(X-\lambda(Y,Z))^2] < \Delta - o_\epsilon(1) \\ +I_K(Y;Z) - o_\epsilon(1)\big)^+ & \text{and } I_K(X;Z) \geq R - o_\epsilon(1) \\ \infty & \text{otherwise,} \end{cases}$$

and

$$\pi_\epsilon(R,\Delta,\Sigma) \triangleq \min_i \max_{k,s} \min_j \max_\lambda \min_{r,t} G_\epsilon(K_\epsilon,\Sigma,\lambda,\Delta,R).$$

Then for any choice of arguments and n sufficiently large $G_\epsilon - G_\epsilon^n \leq \frac{\delta}{3}$. Hence

$$\liminf_{n\to\infty}\pi_\epsilon^n(R,\Delta,\Sigma) \geq \pi_\epsilon(R,\Delta,\Sigma) - \frac{\delta}{3}.$$

Via the use of the functions $\sigma^2(\cdot)$ and $\eta(\cdot,\cdot,\cdot)$ we write the optimization above as follows

$$\pi_\epsilon(R,\Delta,\Sigma) = \min_{\sigma_X}\max_{\sigma_Z,\rho_{xz}}\min_{\sigma_Y}\max_{\lambda}\min_{\rho_{xy},\rho_{yz}} G_\epsilon(K_\epsilon,\Sigma,\lambda,\Delta,R),$$

where the use of $\max,\min$ are justified since we optimizing over finite sets.

Take any sequence $\epsilon_m \to 0$. Let $K^{(m)} = K^{(m)}(\sigma_X^{(m)}, \sigma_Z^{(m)}, \rho_{xz}^{(m)}, \sigma_Y^{(m)}, \rho_{xy}^{(m)}, \rho_{yz}^{(m)})$ and $\lambda^{(m)}$ be such that

$$\pi_{\epsilon_m}(R, \Delta, \Sigma) = G_{\epsilon_m}(K^{(m)}, \Sigma, \lambda^{(m)}, \Delta, R).$$

By considering subsequences, we may assume that $K^{(m)} \to K^\infty$ and $\lambda^{(m)} \to \lambda^\infty$. Then there exists $\tilde{\sigma}_Z^\infty, \tilde{\rho}_{xz}^\infty$ so that

$$\begin{aligned}
&\inf_{\sigma_Y} \sup_{\lambda} \inf_{\rho_{xy},\rho_{yz}} G_G(K(\sigma_X^\infty, \sigma_Y, \tilde{\sigma}_Z^\infty, \tilde{\rho}_{xz}^\infty, \rho_{xy}, \rho_{yz}), \Sigma, \lambda, \Delta, R) \\
\geq &\sup_{\sigma_Z,\rho_{xz}} \inf_{\sigma_Y} \sup_{\lambda} \inf_{\rho_{xy},\rho_{yz}} G_G(K(\sigma_X^\infty, \sigma_Y, \sigma_Z, \rho_{xz}, \rho_{xy}, \rho_{yz}), \Sigma, \lambda, \Delta, R) - \frac{\delta}{3}
\end{aligned}$$

and there are sequences $\tilde{\rho}_{xz}^{(m)}, \tilde{\sigma}_Z^{(m)}$ converging to $\tilde{\rho}_{xz}^\infty$ and $\tilde{\sigma}_Z^\infty$ respectively. Let

$$\tilde{\sigma}_Y^{(m)} \in \arg\min_{\sigma_Y} \max_{\lambda} \min_{\rho_{xy},\rho_{yz}} G_{\epsilon_m}(K(\sigma_X^{(m)}, \sigma_Y, \tilde{\sigma}_Z^{(m)}, \tilde{\rho}_{xz}^{(m)}, \rho_{xy}, \rho_{yz}), \Sigma, \lambda, \Delta, R)$$

and by taking a further subsequence we can assume $\tilde{\sigma}_Y^{(m)} \to \tilde{\sigma}_Y^\infty$. Then there exists $\tilde{\lambda}^\infty$ such that

$$\begin{aligned}
&\inf_{\rho_{xy},\rho_{yz}} G_G(K(\sigma_X^\infty, \tilde{\sigma}_Y^\infty, \tilde{\sigma}_Z^\infty, \tilde{\rho}_{xz}^\infty, \rho_{xy}, \rho_{yz}), \Sigma, \tilde{\lambda}^\infty, \Delta, R) \\
\geq &\sup_{\lambda} \inf_{\rho_{xy},\rho_{yz}} G_G(K(\sigma_X^\infty, \tilde{\sigma}_Y^\infty, \tilde{\sigma}_Z^\infty, \tilde{\rho}_{xz}^\infty, \rho_{xy}, \rho_{yz}), \Sigma, \lambda, \Delta, R) - \frac{\delta}{3}
\end{aligned}$$

and we let $\tilde{\lambda}^{(m)}$ be a sequence converging to $\tilde{\lambda}^\infty$. Let

$$(\tilde{\rho}_{xz}^{(m)}, \tilde{\rho}_{yz}^{(m)}) \in \arg\min_{\rho_{xy},\rho_{yz}} G_{\epsilon_m}(K(\sigma_X^{(m)}, \tilde{\sigma}_Y^{(m)}, \tilde{\sigma}_Z^{(m)}, \tilde{\rho}_{xz}^{(m)}, \rho_{xy}, \rho_{yz}), \Sigma, \lambda^{(m)}, \Delta, R).$$

Define $\tilde{K}^{(m)} \triangleq \tilde{K}^{(m)}(\sigma_X^{(m)}, \tilde{\sigma}_Y^{(m)}, \tilde{\sigma}_Z^{(m)}, \tilde{\rho}_{xz}^{(m)}, \rho_{xy}^{(m)}, \rho_{yz}^{(m)})$, then observe that

$$\begin{aligned}
&\pi_{\epsilon_m}(R, \Delta, \Sigma) \\
=&\max_{\sigma_Z,\rho_{xz}} \min_{\sigma_Y} \max_{\lambda} \min_{\rho_{xy},\rho_{yz}} G_{\epsilon_m}(K^{(m)}(\sigma_X^{(m)}, \sigma_Y, \sigma_Z, \rho_{xz}, \rho_{xy}, \rho_{yz}), \Sigma, \lambda, \Delta, R) \\
\geq&\min_{\sigma_Y} \max_{\lambda} \min_{\rho_{xy},\rho_{yz}} G_{\epsilon_m}(K^{(m)}(\sigma_X^{(m)}, \sigma_Y, \tilde{\sigma}_Z^{(m)}, \tilde{\rho}_{xz}^{(m)}, \rho_{xy}, \rho_{yz}), \Sigma, \lambda, \Delta, R) \\
=&\max_{\lambda} \min_{\rho_{xy},\rho_{yz}} G_{\epsilon_m}(K^{(m)}(\sigma_X^{(m)}, \tilde{\sigma}_Y^{(m)}, \tilde{\sigma}_Z^{(m)}, \tilde{\rho}_{xz}^{(m)}, \rho_{xy}, \rho_{yz}), \Sigma, \lambda, \Delta, R) \\
\geq&\min_{\rho_{xy},\rho_{yz}} G_{\epsilon_m}(K^{(m)}(\sigma_X^{(m)}, \tilde{\sigma}_Y^{(m)}, \tilde{\sigma}_Z^{(m)}, \tilde{\rho}_{xz}^{(m)}, \rho_{xy}, \rho_{yz}), \Sigma, \tilde{\lambda}^{(m)}, \Delta, R) \\
=&G_{\epsilon_m}(\tilde{K}^{(m)}, \Sigma, \tilde{\lambda}^{(m)}, \Delta, R)
\end{aligned}$$

By examining the various cases and using the continuity of expectation and the information measures, one can show that

$$\liminf_{m\to\infty} G_{\epsilon_m}(\tilde{K}^{(m)}, \Sigma, \tilde{\lambda}^{(m)}, R, \Delta) \geq G_G(\tilde{K}^{\infty}, \Sigma, \tilde{\lambda}^{\infty}, R, \Delta).$$

Furthermore,

$$\begin{aligned}
&G_G(\tilde{K}^{\infty}, \Sigma, \tilde{\lambda}^{\infty}, R, \Delta) \\
\geq &\inf_{\rho_{xz},\rho_{yz}} G_G(K(\sigma_X^{\infty}, \tilde{\sigma}_Y^{\infty}, \tilde{\sigma}_Z^{\infty}, \tilde{\rho}_{xz}^{\infty}, \rho_{xy}, \rho_{yz}), \Sigma, \tilde{\lambda}^{\infty}, R, \Delta) \\
\geq &\sup_{\lambda} \inf_{\rho_{xy},\rho_{yz}} G_G(K(\sigma_X^{\infty}, \tilde{\sigma}_Y^{\infty}, \tilde{\sigma}_Z^{\infty}, \tilde{\rho}_{xz}^{\infty}, \rho_{xy}, \rho_{yz}), \Sigma, \lambda, \Delta, R) - \frac{\delta}{3} \\
\geq &\inf_{\sigma_Y} \sup_{\lambda} \inf_{\rho_{xy},\rho_{yz}} G_G(K(\sigma_X^{\infty}, \sigma_Y, \tilde{\sigma}_Z^{\infty}, \tilde{\rho}_{xz}^{\infty}, \rho_{xy}, \rho_{yz}), \Sigma, \lambda, \Delta, R) - \frac{\delta}{3} \\
\geq &\sup_{\sigma_Z,\rho_{xz}} \inf_{\sigma_Y} \sup_{\lambda} \inf_{\rho_{xy},\rho_{yz}} G_G(K(\sigma_X^{\infty}, \sigma_Y, \sigma_Z, \rho_{xz}, \rho_{xy}, \rho_{yz}), \Sigma, \lambda, \Delta, R) - \frac{2\delta}{3} \\
\geq &\pi(R, \Delta, \Sigma) - \frac{2\delta}{3}
\end{aligned}$$

Hence

$$\liminf_{m\to\infty} \liminf_{n\to\infty} \pi_{\epsilon_m}^n(R, \Delta, \Sigma) \geq \pi(R, \Delta, \Sigma) - \delta.$$

But $\epsilon \to 0$ and $\delta > 0$ were arbitrary. □

Proof of Theorem 19.

$$\begin{aligned}
&\Pr\left(\frac{1}{n}\|X^n - \hat{X}^n\|_2^2 > \Delta\right) \qquad \text{(B.45)} \\
&= \Pr\left(\frac{1}{n}\|X^n - \hat{X}^n\|_2^2 > \Delta | (X^n, Y^n, Z^n) \in (\mathcal{R}^{3n})^c\right) \Pr((\mathcal{R}^{3n})^c\Big) \\
&\quad + \Pr\left(\frac{1}{n}\|X^n - \hat{X}^n\|_2^2 > \Delta | (X^n, Y^n, Z^n) \in \mathcal{R}^{3n}\right) \Big) \Pr(\mathcal{R}^{3n}) \\
&\leq \int_{\mathcal{R}^{3n}} \Pr\left(\frac{1}{n}\|X^n - \hat{X}^n\|_2^2 > \Delta | \mathbf{x}, \mathbf{y}, \mathbf{z}\right) \Big) dF(\mathbf{xyz}) + \Pr((\mathcal{R}^{3n})^c)
\end{aligned}$$

For now we focus on the integral and will deal with $\Pr((\mathcal{R}^{3n})^c)$ separately.

Observe first that the error probability on $(\mathcal{E}_b \cup \mathcal{E}_c \cup \mathcal{E}_d)^c$ is zero, thus we can we can split the integral as follows, allowing us to deal with the various key

events defined in Section B.5.2.

$$\begin{aligned}
&\int_{\mathcal{R}^{3n}\cap\mathcal{E}_c} \Pr\left(\frac{1}{n}\|X^n-\hat{X}^n\|_2^2 > \Delta|\mathbf{x},\mathbf{y},\mathbf{z}\right) dF(\mathbf{xyz}) \\
&\quad + \int_{\mathcal{R}^{3n}\cap\mathcal{E}_d} \Pr\left(\frac{1}{n}\|X^n-\hat{X}^n\|_2^2 > \Delta|\mathbf{x},\mathbf{y},\mathbf{z}\right) dF(\mathbf{xyz}) \\
&\quad + \int_{\mathcal{R}^{3n}\cap\mathcal{E}_b} \Pr\left(\frac{1}{n}\|X^n-\hat{X}^n\|_2^2 > \Delta|\mathbf{x},\mathbf{y},\mathbf{z}\right) dF(\mathbf{xyz}).
\end{aligned}$$

Bounding the error probability on $\mathcal{E}_c$ and $\mathcal{E}_d$ by 1 gives

$$\begin{aligned}
&\Pr(\mathcal{E}_c\cap\mathcal{R}^{3n}) + \Pr(\mathcal{E}_d\cap\mathcal{R}^{3n}) \\
&+ \int_{\mathcal{R}^{3n}\cap\mathcal{E}_b} \Pr\left(\frac{1}{n}\|X^n-\hat{X}^n\|_2^2 > \Delta|\mathbf{x},\mathbf{y},\mathbf{z}\right) dF(\mathbf{xyz}).
\end{aligned} \tag{B.46}$$

By Lemma 53, $\Pr(\mathcal{E}_c\cap\mathcal{R}^{3n})$ tends to zero double exponentially with the block length and can therefore also be neglected. Let

$$\mathcal{D}_d = \{K : T_K^\epsilon \cap \mathcal{E}_d \neq \emptyset\}.$$

Then applying Lemma 54 gives

$$\begin{aligned}
\Pr(\mathcal{E}_d\cap\mathcal{R}^{3n}) &\leq \sum_i \sum_j \sum_{r,t:K(i,j,k(i),r,s(i),t)\in\mathcal{D}_d} \\
&\quad \exp(-n(D(K||\bar{K}) - o_\epsilon(1) - \delta_p)) \\
&\leq \sum_i \sum_j |\mathcal{P}_{XYZ}^\epsilon| \max_{r,t:K(i,j,k(i),r,s(i),t)\in\mathcal{D}_d} \\
&\quad \exp(-n(D(K||\bar{K}) - o_\epsilon(1) - \delta_p)).
\end{aligned}$$

where we have written K for $K_\epsilon(i,j,k,r,s(i),t)$ and likewise $\bar{K}$ for $\overline{K_\epsilon(i,j,k(i),r,s(i),t)}$. Next let

$$\mathcal{D}_b = \{K : T_K^\epsilon \cap \mathcal{E}_b \neq \emptyset\}.$$

Addressing the integral in (B.46),

$$
\begin{aligned}
&\int_{\mathcal{R}^{3n}\cap\mathcal{E}_b} \Pr\left(\frac{1}{n}\|X^n-\hat{X}^n\|_2^2>\Delta|\mathbf{x},\mathbf{y},\mathbf{z}\right)dF(\mathbf{xyz})\\
&\overset{(a)}{\leq}\sum_{K\in\mathcal{D}_b}\int_{T_K\cap\mathcal{E}_b}\exp(-n(R-J(K)-o_\epsilon(1)-\delta_b)^+)\\
&\quad dF(\mathbf{xyz})\\
&\leq\sum_{K\in\mathcal{D}_b}\exp(-n(D(K||\overline{K})-o_\epsilon(1)+(R-J(K)-o_\epsilon(1)-\delta_b)^+-\delta_p))\\
&=\sum_i\sum_j\sum_{(r,t):K(i,j,k(i),r,s(i),t)\in\mathcal{D}_b}\\
&\quad\exp(-n(D(K||\overline{K})-o_\epsilon(1)+(R-J(K)-o_\epsilon(1)-\delta_b)^+-\delta_p))\\
&\leq\sum_i\sum_j|\mathcal{P}^\epsilon_{XYZ}|\max_{(r,t):K(i,j,k(i),r,s(i),t)\in\mathcal{D}_b}\\
&\quad\exp(-n(D(K||\overline{K})-o_\epsilon(1)+(R-J(K)-o_\epsilon(1)-\delta_b)^+-\delta_p)),
\end{aligned}
$$

where (a) follows from Lemma 56 and (b) follows from Lemma 54.

Turning to $(\mathcal{R}^{3n})^c$, using well-known large-deviations results for the Gaussian distribution, we obtain

$$
\begin{aligned}
\Pr((\mathcal{R}^{3n})^c)&\leq 2\Pr(\mathbf{x}^t\mathbf{x}<n(M_L+\epsilon))+2\Pr(\mathbf{x}^t\mathbf{x}>nM_U)\\
&\leq 2\exp\left(-\frac{n}{2}\left((M_L+\epsilon)-\log(M_L+\epsilon)-1-o_\epsilon(1)\right)\right)\\
&\quad+2\exp\left(-\frac{n}{2}\left(M_U-\log M_U-1-o_\epsilon(1)\right)\right).
\end{aligned}
$$

Now M_U and M_L can be chosen so that this term does not dominate the exponent and can therefore be neglected. Combining the various bounds (and

neglecting the terms in the previous equation) gives

$$\begin{aligned}
&\Pr\left(\frac{1}{n}\|X^n - \hat{X}^n\|_2^2 > \Delta \cap \mathcal{R}^{3n}\right) \\
&\leq \sum_{i,j} |\mathcal{P}^{\epsilon}_{XYZ}|\Big[\max_{(r,t):K\in\mathcal{D}_d} \exp(-n(D(K||\overline{K}) - \delta_p - o_\epsilon(1))) \\
&+ \max_{(r,t):K\in\mathcal{D}_b} \exp(-n(D(K||\overline{K}) - o_\epsilon(1) + (R - J(K) - o_\epsilon(1) \\
&\qquad - \delta_b)^+) - \delta_p)\Big].
\end{aligned}$$

Using the formula $a + b \leq 2\max(a, b)$, we can upper bound the quantity in square brackets by

$$\begin{aligned}
&2\max\Big(\max_{(r,t):K\in\mathcal{D}_d} \exp(-n(D(K||\overline{K}) - \delta_p - o_\epsilon(1))), \\
&\qquad \max_{(r,t):K\in\mathcal{D}_b} \exp(-n(D(K||\overline{K}) - o_\epsilon(1) + (R - J(K) - o_\epsilon(1) - \delta_b)^+) - \delta_p)\Big).
\end{aligned}$$

Note that the sets $\mathcal{D}_b$ and $\mathcal{D}_d$ may overlap. However, without loss of generality, we may assume that the $o_\epsilon(1)$ terms are such that the objective in the $\mathcal{D}_d$ max is no smaller than the objective in the $\mathcal{D}_b$ max. This quantity can then be further upper bounded by replacing the maximum over (r, t) such that $K \in \mathcal{D}_b$ with a maximum over (r, t) such that $K \in \mathcal{D}_b \backslash \mathcal{D}_d$. This yields

$$2|\mathcal{P}^{\epsilon}_{XYZ}| \max_{(r,t)} H(K),$$

with $H(K) = \exp(-nG^n_\epsilon(K))$, where $G^n_\epsilon(K)$ is as in Lemma 57.

Thus

$$P\left(\frac{1}{n}\|X^n - \hat{X}^n\|_2^2 > \Delta\right) \leq \sum_i \sum_j 2|\mathcal{P}^{\epsilon}_{XYZ}| \max_{r,t} H(K).$$

Since λ and the choice of the test channel were arbitrary, the right-hand side is upper bounded by

$$2|\mathcal{P}^{\epsilon}_{XYZ}|^3 \max_i \min_{k,s} \max_j \min_\lambda \max_{r,t} H(K).$$

We then let take logs, divide by n, and let n tend to infinity and ϵ tend to zero, invoking Lemma 57 to obtain the desired result. □

B.6 Proof of Theorem 20

Proof. Let f^n, g^n be a code for the two-sided Gaussian rate distortion problem with conditional rate distortion function $R_{X|Y}$ and define

$$\mathcal{E}_\Delta^n \triangleq \{(\mathbf{x}, \mathbf{y}) : \|\mathbf{x} - g^n(f^n(\mathbf{x}, \mathbf{y}), \mathbf{y})\|_2^2 > n\Delta\}$$

and

$$\mathcal{E}_K^n \triangleq \{(\mathbf{x}, \mathbf{y}) : nK \geq \|\mathbf{x} - g^n(f^n(\mathbf{x}, \mathbf{y}), \mathbf{y})\|_2^2\},$$

where $K \in \mathbb{R}^+$ is to be specified later. For R fixed, choose a covariance matrix Π so that

$$R_{X|Y}(f_\Pi, \Delta) > R. \tag{B.47}$$

Let Δ' be the solution to $R_{X|Y}(f_\Pi, \Delta') = R$ and define $\bar{\Delta}(f^n, g^n) \triangleq \mathbb{E}_\Pi[\frac{1}{n}\|X^n - g^n(f^n(X^n, Y^n), Y^n)\|_2^2]$. Then according to [79, section 4]

$$R_{X|Y}(f_\Pi, \Delta) > RR_{X|Y}(f_\Pi, \Delta') \geq R_{X|Y}(f_\Pi, \bar{\Delta}) \tag{B.48}$$

for every n and code (f^n, g^n) with rate at most R. Monotonicity of the rate distortion function implies that $\bar{\Delta}(f^n, g^n) \geq \Delta' > \Delta$.

To continue we modify our original code to give $(\tilde{f}^n, \tilde{g}^n)$. The modification comprises adding a new codeword such that the decoder emits the string $\mathbf{0}$ on receipt of this codeword. Encoder $\tilde{f}^n$, knowing the side information can choose to send this codeword if the choice by f_n results in a higher distortion than $\frac{1}{n}\|X^n\|_2^2$. If we let $\hat{X}^n = g^{(n)}(f^{(n)}(X^n, Y^n), Y^n)$ and $\tilde{\hat{X}}^n = \tilde{g}^{(n)}(\tilde{f}^{(n)}(X^n, Y^n), Y^n)$

then we see that $n^{-1}(X^n - \tilde{\hat{X}}^n)^2 \leq \frac{1}{n}\|X^n\|_2^2$ a.s. Modifying the code in this way only reduces the squared error, hence defining

$$\tilde{\mathcal{E}}_\Delta^n = \{(\mathbf{x}, \mathbf{y}) : \|\mathbf{x} - \tilde{g}^n(\tilde{f}^n(\mathbf{x}, \mathbf{y}), \mathbf{y})\|_2^2 > n\Delta\}$$

(and correspondingly $\tilde{\mathcal{E}}_K^n$) we see that $\mathcal{E}_\Delta \supset \tilde{\mathcal{E}}_\Delta$. In the following all expectations and probabilities are with respect to the law f_Π unless stated otherwise.

$$\begin{aligned}\mathbb{E}[\|X^n - \tilde{\hat{X}}^n\|_2^2 \mathbf{1}_{\tilde{\mathcal{E}}_\Delta^n \cap (\tilde{\mathcal{E}}_K^n)^c}] &\leq \mathbb{E}[\|X^n\|_2^2 \mathbf{1}_{\tilde{\mathcal{E}}_\Delta^n \cap (\tilde{\mathcal{E}}_K^n)^c}] \\ &\leq \mathbb{E}[\|X^n\|_2^2 \mathbf{1}_{\{\|X^n\|_2^2 > nK\}}].\end{aligned}$$

Next, applying the Cauchy-Schwarz inequality gives

$$\begin{aligned}&\leq \sqrt{\mathbb{E}[(\|X^n\|_2^2)^2] \Pr(\|X^n\|_2^2 > nK)} \\ &= \sqrt{\mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n X_i^2 X_j^2\right] \Pr(\|X^n\|_2^2 > nK)} \\ &= \sqrt{(n\mathbb{E}[X_1^4] + (n^2 - n)\mathbb{E}[X_1^2]\mathbb{E}[X_1^2]) \Pr(\|X^n\|_2^2 > nK)}.\end{aligned}$$

Choosing $K = \mathbb{E}[X_1^2] + \epsilon$ and applying Chebyshev's inequality to the probability allows us to further bound this quantity by

$$\begin{aligned}&\leq \sqrt{(n\mathbb{E}[X_1^4] + (n^2 - n)\mathbb{E}[X_1^2]\mathbb{E}[X_1^2])} \\ &\quad \times \sqrt{\frac{\mathbb{E}[X_1^4] - \mathbb{E}[X_1^2]^2}{n\epsilon^2}}.\end{aligned}$$

Hence

$$\begin{aligned}&\mathbb{E}[n^{-1}||X^n - \hat{\tilde{X}}^n||^2 \mathbf{1}_{\tilde{\mathcal{E}}_\Delta^n \cap (\tilde{\mathcal{E}}_K^n)^c}] \\ &\leq \sqrt{(n^{-1}\mathbb{E}[X_1^4] + (1 - n^{-1})\mathbb{E}[X_1^2]\mathbb{E}[X_1^2])} \\ &\quad \times \sqrt{\frac{\mathbb{E}[X_1^4] - \mathbb{E}[X_1^2]^2}{n^3\epsilon^2}}\end{aligned}$$

which goes to zero with n. We note that this new code has rate $\tilde{R} = R + n^{-1}\log(1+\exp(-nR)) = R + o_n(1)$. Let $\Delta' - \Delta > \delta_1 > 0$, and $\tilde{\Delta}$ be the solution to $\tilde{R} = R(f_\Pi, \tilde{\Delta})$. Then for n sufficiently large

$$\Delta' - \tilde{\Delta} < \delta_1.$$

We also note that $\bar{\Delta}(\tilde{f}^n, \tilde{g}^n) \geq \tilde{\Delta}$. One may decompose the space into different events to see that

$$\begin{aligned}
\bar{\Delta}(\tilde{f}^n, \tilde{g}^n) &= \mathbb{E}[n^{-1}\|X^n - \tilde{\hat{X}}^n\|_2^2] \\
&= \mathbb{E}[n^{-1}\|X^n - \tilde{\hat{X}}^n\|_2^2 \mathbf{1}_{(\tilde{\mathcal{E}}_\Delta^n)^c}] \\
&\quad + \mathbb{E}[n^{-1}\|X^n - \tilde{\hat{X}}^n\|_2^2 \mathbf{1}_{\tilde{\mathcal{E}}_\Delta^n \cap \tilde{\mathcal{E}}_K^n}] \\
&\quad + \mathbb{E}[n^{-1}\|X^n - \tilde{\hat{X}}^n\|_2^2 \mathbf{1}_{\tilde{\mathcal{E}}_\Delta^n \cap (\tilde{\mathcal{E}}_K^n)^c}] \\
&\leq \Delta \Pr((\tilde{\mathcal{E}}_\Delta^n)^c) + K \Pr(\tilde{\mathcal{E}}_\Delta^n \cap \tilde{\mathcal{E}}_K^n) \\
&\quad + \mathbb{E}[n^{-1}\|X^n - \tilde{\hat{X}}^n\|_2^2 \mathbf{1}_{\tilde{\mathcal{E}}_\Delta^n \cap (\tilde{\mathcal{E}}_K^n)^c}] \\
&\leq \Delta(1 - \Pr(\tilde{\mathcal{E}}_\Delta^n)) + K \Pr(\tilde{\mathcal{E}}_\Delta^n) \\
&\quad + \mathbb{E}[n^{-1}\|X^n - \tilde{\hat{X}}^n\|_2^2 \mathbf{1}_{\tilde{\mathcal{E}}_\Delta^n \cap (\tilde{\mathcal{E}}_K^n)^c}]
\end{aligned}$$

i.e.

$$\Pr(\tilde{\mathcal{E}}_\Delta^n) \geq \frac{\bar{\Delta}(\tilde{f}^n, \tilde{g}^n) - \Delta - \mathbb{E}[n^{-1}\|X^n - \tilde{\hat{X}}^n\|_2^2 \mathbf{1}_{\tilde{\mathcal{E}}_\Delta^n \cap (\tilde{\mathcal{E}}_K^n)^c}]}{K - \Delta}. \tag{B.49}$$

Thus

$$\begin{aligned}
\Pr(\mathcal{E}_\Delta^n) &\geq \Pr(\tilde{\mathcal{E}}_\Delta^n) \\
&\geq \frac{\tilde{\Delta} - \Delta - \mathbb{E}[n^{-1}\|X^n - \hat{X}^n\|_2^2 \mathbf{1}_{\tilde{\mathcal{E}}_\Delta^n \cap (\tilde{\mathcal{E}}_K^n)^c}]}{K - \Delta} \\
&\geq \frac{\Delta' - \Delta - \delta_2}{K - \Delta} \triangleq \alpha > 0
\end{aligned}$$

for all $n > n_1$ (where $\delta_2 \triangleq \delta_1 + \mathbb{E}[n^{-1}(X^n - \hat{X}^n)^2 \mathbf{1}_{\mathcal{E}_\Delta^n \cap (\mathcal{E}_K^n)^c}]$). Next, we set

$$G^n = \left\{ (\mathbf{x}, \mathbf{y}) : \left| \frac{1}{n} \log \frac{f_\Pi(\mathbf{x}, \mathbf{y})}{f_\Sigma(\mathbf{x}, \mathbf{y})} - D(\Pi||\Sigma) \right| < \delta_3 \right\}.$$

By the law of large numbers,

$$\int_{G^n} f_\Pi(\mathbf{x},\mathbf{y})\,d\mathbf{x}\mathbf{y} > 1 - \frac{1}{2}\alpha$$

for all n sufficiently large. Combining everything, this gives

$$\begin{aligned}
\Pr_\Sigma(\mathcal{E}_\Delta^n) &= \int_{\mathcal{E}_\Delta^n} f_\Sigma(\mathbf{x},\mathbf{y})\,d\mathbf{x}\mathbf{y} \\
&\geq \int_{\mathcal{E}_\Delta^n \cap G^n} f_\Sigma(\mathbf{x},\mathbf{y})\,d\mathbf{x}\mathbf{y} \\
&= \int_{\mathcal{E}_\Delta^n \cap G^n} f_\Pi(\mathbf{x},\mathbf{y}) \exp\left(-\log\frac{f_\Pi(\mathbf{x},\mathbf{y})}{f_\Sigma(\mathbf{x},\mathbf{y})}\right) d\mathbf{x}\mathbf{y} \\
&\geq \frac{1}{2}\alpha \exp(-n(D(\Pi||\Sigma) + \delta_3)).
\end{aligned}$$

We observe that this inequality holds for all codes of rate at most R and Π satisfying (B.47). To complete the proof it suffices to show that

$$\lim_{\epsilon\to 0} \inf_{\Pi: R(\Pi,\Delta) > R+\epsilon} D(\Pi||\Sigma) = \inf_{\Pi: R(\Pi,\Delta) > R} D(\Pi||\Sigma)$$

The first direction ($\geq$) is obvious. For the reverse inequality, choose Π^* to achieve within δ of the infimum on the right-hand side. Let $\Pi^{(\epsilon)}$ be a collection of covariance matrices converging to Π^* such that $R(\Pi^{(\epsilon)},\Delta) > R + \epsilon$. That such a choice is possible follows by continuity of the rate distortion function. Then

$$\lim_{\epsilon\to 0} \inf_{\Pi: R(\Pi,\Delta) > R+\epsilon} D(\Pi||\Sigma) \leq \lim_{\epsilon\to 0} D(\Pi^{(\epsilon)}||\Sigma) = D(\Pi^*||\Sigma) \leq \inf_{\Pi: R(\Pi,\Delta) > R} D(\Pi||\Sigma) + \delta$$

by continuity of relative entropy. But δ was arbitrary. □

APPENDIX C

CHAPTER 5 - PROOFS

C.1 Proof of Theorem 24

The key to the proof is Lemma 59, a bound on degree of the codebook graph which holds with exponentially high probability. With this fact established we give a scheme for coding when the bound holds and declare an error when the bound does not. Throughout this section the reader should keep in mind that besides the source, the randomness comes from the codebook construction.

C.1.1 Codebook Construction

Operating on blocks of length n, for each type Q_X choose a test channel $Q_{U^*|X} = Q^*_{U|X}(Q_X)$ and let $Q_{U^*} = Q^*_U(Q_X)$ denote the resulting induced marginal type[1]. The test channel is used to build a codebook $B^n(Q_X)$ as follows. For each $\mathbf{u} \in T_{Q_{U^*}}$, flip a coin with probability of heads

$$p \triangleq \exp\Big(-n\Big[H(Q_{U^*|X}|Q_X) - 3\frac{|\mathcal{U}||\mathcal{X}|\log(n+1)}{n}\Big]\Big),$$

and add $\mathbf{u}$ to the codebook only if the coin comes up heads. Define the distribution

$$Q_{UY}(u,y) = \sum_{x\in\mathcal{X}} P_{XY}(x,y)Q_{U^*|X}(u|x)$$

and let G_{U^*} be the resulting characteristic graph. The codeword for $\mathbf{x} \in T_{Q_X}$ is chosen as follows. If $\mathcal{G}(\mathbf{x}) \triangleq B^n(Q_X) \cap T_{Q^*_{U|X}}(\mathbf{x})$ is non-empty, choose uni-

[1]For brevity we will use the following conventions: The random variable U^* (resp. channel $Q_{U^*|X}$) refers to the random variable (resp. channel) defined by the choice of test channel for the particular Q_X under consideration.

formly from $\mathcal{G}(\mathbf{x})$. If $\mathcal{G}(\mathbf{x})$ is null, choose uniformly from $B^n(Q_X)$. We let $U(\mathbf{x})$ denote the chosen codeword. For each codebook, we define $b_{Q_X} : B^n(Q_X) \to [1, \ldots, \exp(nR)]$ (a binning function) as follows, for all $\mathbf{u} \in B^n(Q_X)$

$$\Pr(b_{Q_X}(\mathbf{u}) = i) = \exp(-nR), \text{for all } i \in [1, \ldots, \exp(nR)].$$

C.1.2 Scheme

In Lemmas 58 and 59 we establish that

$$\begin{aligned}\gamma(G_{U^*} \cap B^n(Q_X)) &\leq \Delta(G_{U^*} \cap B^n(Q_X)) + 1 \\ &\overset{\text{w.h.p.}}{\leq} \exp(n[\kappa_2(Q_X) + \lambda_n + \tilde{\delta}_n]) + 1,\end{aligned}$$

for some $\lambda_n > 0$, $\tilde{\delta}_n \to 0$ as $n \to \infty$ and where $w.h.p$ stands for probability tending to 1 as $n \to \infty$. For types Q_X in which the above bound fails to hold, we send an error message to the decoder. For types in which the bound holds, the scheme is as follows. To communicate the codeword to the decoder, the encoder may either give an index into the codeword set B^n or using the ideas from the improved lossless binning scheme, it can color the graph $G^n_{U^*} \cap B^n(Q_X)$ using a minimal coloring and send the color of the codeword.

Encoder:

The encoder first sends the type of the source sequence $Q_{\mathbf{x}}$. If $\exp(n[\kappa_2(Q_{\mathbf{x}}) + \lambda_n + \tilde{\delta}_n]) + 1 < \exp(nR)$, the encoder transmits the color of the codeword in the graph $G_{U^*} \cap B^n(Q_{\mathbf{x}})$. Otherwise it sends the bin index $b_{Q_{\mathbf{x}}}(U(\mathbf{x}))$. Formally, we denote the encoder by $f^n : \mathcal{X}^n \to \mathcal{M}$, where

$$\mathcal{M} = [1, \ldots, (n+1)^{|\mathcal{X}|}] \times [1, \ldots, \exp(nR)]$$

Decoder:

The decoder receives a type index, a message and the side information $\mathbf{y}$. If $\exp(n[\kappa_2(Q_{\mathbf{x}}) + \lambda]) + 1 < \exp(nR)$ then the codeword can be decoded without error. In the opposite case, the decoder searches the bin for a unique codeword $\hat{\mathbf{u}}$, so that among all $\tilde{\mathbf{u}}$ in the received bin, $H(\hat{\mathbf{u}}|\mathbf{y}) < H(\tilde{\mathbf{u}}|\mathbf{y})$. If there is no such unique codeword, the decoder chooses $\hat{\mathbf{u}}$ uniformly at randomly from the received bin. For each pair of types Q_X, Q_Y, the decoder picks an reproduction function ϕ, and declares the output as

$$\hat{\mathbf{x}} \text{ where } \hat{\mathbf{x}}_j = \phi(\hat{\mathbf{u}}_j, \mathbf{y}_j).$$

Thus the decoder $g^n : \mathcal{Y}^n \times \mathcal{M} \to \hat{\mathcal{X}}$ is specified.

Lemma 58. *Let*

$$\begin{aligned} \delta_n &= 3\frac{|\mathcal{U}||\mathcal{X}|\log(n+1)}{n} \text{ and } \tilde{\delta}_n = \frac{|\mathcal{U}||\mathcal{U}|}{n}\log(n+1) \\ \kappa_2^n(Q_X) &= \kappa_2(Q_X) + \tilde{\delta}_n \text{ and} \\ \lambda_n &= \frac{2}{n}\log(n+1) + \delta_n. \end{aligned}$$

Then for all n sufficiently large and for all types Q_X,

$$\begin{aligned} &\Pr(\Delta(G_{U^*}^n \cap B^n(Q_X)) > \exp(n[\kappa_2^n(Q_X) + \lambda_n]) \\ &\leq \exp_e(-(n+1)^2). \end{aligned}$$

Note the randomness in $\Delta(G_U^n \cap B^n(Q_X))$ comes from the fact that $B^n(Q_X)$ is a random set.

Proof. Let $K = 2^{n[\kappa_2^n(Q_X)+\lambda_n]}$, then

$$\begin{aligned}
&\Pr(\Delta(G_{U^*}^n \cap B^n(Q_X)) > K) \\
&= \Pr(\exists \mathbf{u} \in T_{Q_{U^*}} : \mathbf{u} \in B^n(Q_X), \Delta(\mathbf{u}) > K) \\
&\leq \sum_{\mathbf{u} \in T_{Q_U^*}} \Pr(\mathbf{u} \in B^n(Q_X)) \Pr(\Delta(\mathbf{u}) \geq K | \mathbf{u} \in B^n(Q_X)) \\
&\leq \sum_{\mathbf{u} \in T_{Q_U^*}} \Pr(\Delta(\mathbf{u}) \geq K | \mathbf{u} \in B^n(Q_X)).
\end{aligned}$$

Let $N(\mathbf{u})$ denote the neighbors of $\mathbf{u}$ in the graph G_U^n, then quantity in the previous line is upper bounded by

$$\sum_{\mathbf{u} \in T_{Q_U^*}} \Pr\Big(\sum_{\mathbf{v} \in N(\mathbf{u})} \mathbf{1}_{\{\mathbf{v} \in B^n\}} \geq K \Big).$$

From the construction of the codebook, we know that for each string $\mathbf{v}$, $\mathbf{1}_{\{\mathbf{v} \in B^n\}}$ is Bernoulli with parameter p. Furthermore, by Lemma 23, we know that $|N(\mathbf{u})| \leq \exp(n[\kappa(G_U, Q_U^*) + \tilde{\delta}_n]) \triangleq J(Q_X)$. Therefore, by bounding the number of terms in the summation, letting D_i be a sequence of i.i.d. Bernoulli(p) random variables, we have

$$\begin{aligned}
&\Pr(\Delta(G_{U^*}^n \cap B^n(Q_X)) > K) \\
&\leq |T_{Q_{U^*}}| \Pr\Big(\sum_{i=1}^{J(Q_X)} D_i \geq K \Big).
\end{aligned}$$

Focusing on the probability, using the exponential form of Markov's inequality, one has for any $\theta > 0$

$$\begin{aligned}
\Pr\Big(\sum_{i=1}^{J(Q_X)} D_i \geq K \Big) &\leq \frac{\exp_e(J(Q_X) \ln(1 + p(e^\theta - 1)))}{\exp_e(\theta K)} \\
&\leq \frac{\exp_e(J(Q_X) p(e^\theta - 1))}{\exp_e(\theta K)} \\
&\leq \frac{\exp_e(J(Q_X) p e^\theta)}{\exp_e(\theta K)} \\
&\leq \exp_e(2^{n[\kappa_2(Q_X)+\delta_n+\tilde{\delta}_n]+\theta \log e} - \theta 2^{n[\kappa_2(Q_X)+\tilde{\delta}_n+\lambda_n]}). \qquad \text{(C.1)}
\end{aligned}$$

Choosing $\theta = 1$, we have

$$\Pr\Big(\sum_{i=1}^{J(Q_X)} D_i \geq K\Big) \leq \exp_e(2^{n[\kappa_2(Q_X)+\delta_n+\tilde{\delta}_n]}(2^{\log e} - (n+1)^2)).$$

For $n \geq 1$, $(e - (n+1)^2) < -1$, hence

$$\begin{aligned}\Pr(\Delta(G^n_{U^*} \cap B^n(Q_X)) > K) &\leq |T_{Q_{U^*}}| \exp_e(-2^{n[\kappa_2(Q_X)+\delta_n+\tilde{\delta}_n]}) \\ &\leq |T_{Q_{U^*}}| \exp_e(-2^{n\delta_n}) \\ &\leq |T_{Q_{U^*}}| \exp_e(-(n+1)^3),\end{aligned}$$

for all n sufficiently large. Since $|T_{Q_{U^*}}|$ is only exponential in n, the result holds.

□

On account of the previous lemma, we have a bound, which holds with high probability, on the degree of $G_{U^*} \cap B^n(Q_X)$. For each Q_{XYU}, we define the event $F(Q_{XYU})$ as follows

$$F(Q_{XYU}) \triangleq \{\Delta(B^n(Q_X) \cap G_{U^*}) > e^{n[\kappa_2^n(Q_X)+\lambda_n]}\}.$$

Lemma 59. *For all n sufficiently large and any type Q_{XYU}*

$$\Pr(F(Q_{XYU})) \leq \exp(-(n+1)^2).$$

Proof. The result follows directly from Lemma 58. □

In the remainder of this appendix κ_2^n and λ_n will be defined as in the statement of Lemma 58.

C.1.3 Error Analysis

Let

$$\mathcal{E}_1 = \{(\mathbf{x}, \mathbf{y}, \mathbf{u}) : \mathbf{u} \notin T_{Q^*_{U|X}}(\mathbf{x})\}$$

$$\mathcal{E}_2 = \{(\mathbf{x}, \mathbf{y}, \mathbf{u}) : \mathbf{u} \in T_{Q^*_{U|X}}(\mathbf{x}), d(\mathbf{x}, \phi_{Q_{\mathbf{x}}, Q_{\mathbf{y}}}(\mathbf{u}, \mathbf{y})) < \Delta$$
$$\exp(n[\kappa_2^n(Q_{\mathbf{x}}) + \lambda_n]) + 1 \geq \exp(nR)\}$$

$$\mathcal{E}_3 = \{(\mathbf{x}, \mathbf{y}, \mathbf{u}) : \mathbf{u} \in T_{Q^*_{U|X}}(\mathbf{x}), d(\mathbf{x}, \phi_{Q_{\mathbf{x}}, Q_{\mathbf{y}}}(\mathbf{u}, \mathbf{y})) < \Delta$$
$$\exp(n[\kappa_2^n(Q_{\mathbf{x}}) + \lambda_n]) + 1 < \exp(nR)\}$$

$$\mathcal{E}_4 = \{(\mathbf{x}, \mathbf{y}, \mathbf{u}) : \mathbf{u} \in T_{Q^*_{U|X}}(\mathbf{x}), d(\mathbf{x}, \phi_{Q_{\mathbf{x}}, Q_{\mathbf{y}}}(\mathbf{u}, \mathbf{y})) \geq \Delta\}$$

and

$$\mathcal{D}_1 = \{Q_{XYU} : Q_{U|X} \neq Q^*_{U|X}(Q_X))\}$$

$$\mathcal{D}_2 = \{Q_{XYU} : \exp(n[\kappa_2^n(Q_X) + \lambda_n]) + 1 \geq \exp(nR)$$
$$Q_{U|X} = Q^*_{U|X}(Q_X), \mathbb{E}_Q[d(X, \phi_{Q_X, Q_Y}(U, Y)) < \Delta\}$$

$$\mathcal{D}_3 = \{Q_{XYU} : \exp(n[\kappa_2^n(Q_X) + \lambda_n]) + 1 < \exp(nR)$$
$$Q_{U|X} = Q^*_{U|X}(Q_X), \mathbb{E}_Q[d(X, \phi_{Q_X, Q_Y}(U, Y)) < \Delta\}$$

$$\mathcal{D}_4 = \{Q_{XYU} : Q_{U|X} = Q^*_{U|X}(Q_X), \mathbb{E}_Q[d(X, \phi_{Q_X, Q_Y}(U, Y)) \geq \Delta\}.$$

The sets defined above and the following Lemmas allow us to bound the error probability for our improved scheme.

Lemma 60. *For all strings* $\mathbf{x}, \mathbf{y}$*, let*

$$S(\mathbf{x}|\mathbf{y}) = \{\tilde{\mathbf{x}} | H(\tilde{\mathbf{x}}|\mathbf{y}) \leq H(\mathbf{x}|\mathbf{y}), Q_{\tilde{\mathbf{x}}} = Q_{\mathbf{x}}\}.$$

Then

$$|S(\mathbf{x}|\mathbf{y})| \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(nH(\mathbf{x}|\mathbf{y})).$$

Proof.

$$\begin{aligned}
|S(\mathbf{x}|\mathbf{y})| &\leq |\{\tilde{\mathbf{x}}|H(\tilde{\mathbf{x}}|\mathbf{y}) \leq H(\mathbf{x}|\mathbf{y})\}| \\
&= \sum_{V:V\in\mathcal{C}^n(Q_{\mathbf{y}},\mathcal{X})} \sum_{\tilde{\mathbf{x}}\in T_V(\mathbf{y}):H(\tilde{\mathbf{x}}|\mathbf{y})\leq H(\mathbf{x}|\mathbf{y})} 1 \\
&= \sum_{\substack{V:V\in\mathcal{C}^n(Q_{\mathbf{y}},\mathcal{X}) \\ H(V|Q_{\mathbf{y}})\leq H(\mathbf{x}|\mathbf{y})}} |T_V(\mathbf{y})| \\
&\leq \sum_{\substack{V:V\in\mathcal{C}^n(Q_{\mathbf{y}},\mathcal{X}) \\ H(V|Q_{\mathbf{y}})\leq H(\mathbf{x}|\mathbf{y})}} \exp(nH(\mathbf{x}|\mathbf{y})) \\
&\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(nH(\mathbf{x}|\mathbf{y}))
\end{aligned}$$

□

Lemma 61. *Let $X^n, Y^n, U^n = U^*$ be generated according to our scheme, then for all n sufficiently large and all $(\mathbf{x}, \mathbf{y}, \mathbf{u}) \in \mathcal{E}_1$*

$$\Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{\mathbf{xyu}})) \leq \exp(-(n+1)^2).$$

Proof.

$$\begin{aligned}
&\Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{\mathbf{xyu}})) \\
&= \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}) \\
&\quad \times \Pr(F^c(Q_{\mathbf{xyu}})|X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}) \\
&\leq \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u})
\end{aligned}$$

Let A denote the event that there does not exist a $\mathbf{u} \in B^n(Q_{\mathbf{x}})$ such that $\mathbf{u} \in T_{Q_{U^*|X}}(\mathbf{x})$. For $(\mathbf{x}, \mathbf{y}, \mathbf{u}) \in \mathcal{E}_1$, the event $\{X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}\}$ implies that

the event A has occurred. Hence

$$\begin{aligned}
&\Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}) \\
&= \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, A) \\
&\leq \Pr(X^n = \mathbf{x}) \Pr(A | X^n = \mathbf{x}) \\
&\leq \Pr(A | X^n = \mathbf{x}).
\end{aligned}$$

Recalling p was the probability that each codeword is added to the codebook. We have

$$\begin{aligned}
\Pr(A | X^n = \mathbf{x}) &= \Pr(\forall \mathbf{u} \in T_{Q_{U^*|X}} : \mathbf{u} \notin B^n(Q_{\mathbf{x}})) \\
&= (1-p)^{|T_{Q_{U^*|X}(\mathbf{x})}|} \\
&\leq \exp(-p|T_{Q_{U^*|X}(\mathbf{x})}|).
\end{aligned}$$

For $\mathbf{x} \in T_{Q_X}$ we have the lower bound,

$$|T^n_{Q_{U^*|X}}(\mathbf{x})| \geq (n+1)^{-|\mathcal{X}||\mathcal{U}|} \exp(nH(Q_{U^*|X}|Q_X))$$

substituting this and the value of p we obtain

$$\begin{aligned}
\Pr(A | X^n = \mathbf{x}) &\leq \exp\Big(- \exp\Big(n \Big[3 \frac{|\mathcal{U}|\mathcal{X}|}{n} \log(n+1) - \frac{|\mathcal{U}|\mathcal{X}|}{n} \log(n+1) \Big] \Big) \Big) \\
&\leq \exp(-(n+1)^2).
\end{aligned}$$

□

Lemma 62. *Let* $\mathbf{x}, \mathbf{y}, \mathbf{u} \in \mathcal{E}_1^c$*, then*

$$\begin{aligned}
&\Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{\mathbf{xyu}})) \\
&\leq P^n_{XY}(\mathbf{x}, \mathbf{y}) \exp(-n[H(Q^*_{U|X}(Q_{\mathbf{x}})|Q_{\mathbf{x}}) - \delta_n]),
\end{aligned}$$

where

$$\delta_n = 3 \frac{|\mathcal{U}|\mathcal{X}|}{n} \log(n+1).$$

Proof. Proceeding as in proof of Lemma 61, we have

$$\begin{aligned}
&\Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{\mathbf{xyu}})) \\
&\leq \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}) \\
&= \Pr(X^n = \mathbf{x}, Y^n = \mathbf{u}) \Pr(U^n = \mathbf{u} | X^n = \mathbf{x}, Y^n = \mathbf{y}).
\end{aligned}$$

Conditional on $\{X^n = \mathbf{x}\}$, the event $\{U^n = \mathbf{u}\}$ is equivalent to $\{\mathbf{u} \in B^n(Q_{\mathbf{x}})\} \cap \{\mathbf{u}$ was chosen among all $\tilde{\mathbf{u}} \in B^n(Q_{\mathbf{x}})$ with $\tilde{\mathbf{u}} \in T_{Q^*_{U|X}}(\mathbf{x})\}$. Bounding the latter probability by 1, we have

$$\begin{aligned}
&\Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{\mathbf{xyu}})) \\
&\leq P^n_{XY}(\mathbf{x}, \mathbf{y}) \exp(-n[H(Q^*_{U|X}|Q_{\mathbf{x}}) - 3\frac{|\mathcal{U}|\mathcal{X}|}{n} \log(n+1)])
\end{aligned}$$

□

Lemma 63. *For any $Q_{XYU} \in \mathcal{D}^c_1$ and any P_{XY}*

$$\begin{aligned}
&\sum_{(\mathbf{x},\mathbf{y},\mathbf{u}) \in T_{Q_{XYU}}} \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{XYU})) \\
&\leq \exp(-n[D(Q_{XYU}||P_{XY}Q^*_{U|X}(Q_X)) - \delta_n]),
\end{aligned}$$

where δ_n is the same as in the statement of Lemma 62.

Proof. Using the bound of Lemma 62 and the following identity for $(\mathbf{x}, \mathbf{y}) \in T_{Q_{XY}}$,

$$P^n_{XY}(\mathbf{x}, \mathbf{y}) = \exp(-n[D(Q_{XY}||P_{XY}) + H(Q_{XY})]),$$

we have

$$\begin{aligned}
&\sum_{(\mathbf{x},\mathbf{y},\mathbf{u})\in T_{Q_{XYU}}} \Pr(X^n=\mathbf{x}, Y^n=\mathbf{y}, U^n=\mathbf{u}, F^c(Q_{XYU})) \\
&\leq \sum_{(\mathbf{x},\mathbf{y},\mathbf{u})\in T_{Q_{XYU}}} \exp(-n[D(Q_{XY}||P_{XY}) + H(Q_{XY}) \\
&\quad + H(Q_{U|X}|Q_X) - \delta_n]) \\
&\leq \exp(-n[D(Q_{XY}||P_{XY}) - H(Q_{U|XY}|Q_{XY}) \\
&\quad + H(Q_{U|X}|Q_X) - \delta_n]). \qquad \text{(C.2)}
\end{aligned}$$

Applying the identity

$$\begin{aligned}
&D(Q_{XY}||P_{XY}) - H(Q_{U|XY}|Q_{XY}) + H(Q_{U|X}|Q_X) \\
&= D(Q_{XYU}||P_{XY}Q_{U|X})
\end{aligned}$$

in (C.2) gives the result. □

Lemma 64. *For n sufficiently large and $(\mathbf{x},\mathbf{y},\mathbf{u}) \in \mathcal{E}_2$*

$$\begin{aligned}
&\Pr(d(X^n,\hat{X}^n) > \Delta | X^n=\mathbf{x}, Y^n=\mathbf{y}, U^n=\mathbf{u}, F^c(Q_{\mathbf{xyu}})) \\
&\leq \frac{\exp(-n[R - I_{Q_{\mathbf{xyu}}}(X;U) - I_{Q_{\mathbf{xyu}}}(U;Y) - \delta_n]^+)}{1-\exp_e(-(n+1)^2)},
\end{aligned}$$

where $\delta_n \to 0$ as $n \to \infty$.

Proof. Let L be the event that the decoder decodes the wrong codeword, i.e.

$$\begin{aligned}
L \triangleq \{&\exists \tilde{\mathbf{u}} \neq U(X^n) : H(\tilde{\mathbf{u}}|\mathbf{y}) \leq H(U(X^n)|\mathbf{y}), \tilde{\mathbf{u}} \in B^n(Q_{X^n}), \\
&b_{Q_{X^n}}(U(X^n)) = b_{Q_{X^n}}(\tilde{\mathbf{u}})\}
\end{aligned}$$

and note that $\{d(X^n,\hat{X}^n) > \Delta\} \cap \mathcal{E}_2 \subseteq L$. We can bound the conditional proba-

bility of L as follows

$$\begin{aligned}
&\Pr(L|X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{\mathbf{xyu}})) \\
&= \frac{\Pr(L, F^c(Q_{\mathbf{xyu}})|X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u})}{\Pr(F^c(Q_{\mathbf{xyu}})|X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u})} \\
&\leq \frac{\Pr(L|X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u})}{\Pr(\Delta(B^n(Q_{\mathbf{x}}) \cap Q_{U^*}) \leq e^{n[\kappa_2^n(Q_{\mathbf{x}}) + \lambda_n]})}.
\end{aligned}$$

We now bound the numerator. Recalling the definition of $S(\mathbf{u}|\mathbf{y})$ from Lemma 60 and invoking the union bound gives

$$\begin{aligned}
&\Pr(L|X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}) \\
&\leq \sum_{\tilde{\mathbf{u}} \in S(\mathbf{u}|\mathbf{y})} \Pr(\tilde{\mathbf{u}} \in B^n(Q_{\mathbf{x}}), b_{Q_{\mathbf{x}}}(\mathbf{u}) = b_{Q_{\mathbf{x}}}(\tilde{\mathbf{u}})),
\end{aligned}$$

and substituting the various bounds gives

$$\exp(-n[R - I_{Q_{\mathbf{xyu}}}(X;U) + I_{Q_{\mathbf{xyu}}}(U;Y) - \delta_n]^+),$$

where $\delta_n = 4\frac{|\mathcal{U}||\mathcal{X}|}{n}\log(n+1)$. To handle the denominator, by Lemma 58 the complementary event goes to zero super exponentially as $n \to \infty$. □

Lemma 65. *Let $\delta_n, \tilde{\delta}_n, \tilde{\tilde{\delta}}_n, \tilde{\tilde{\tilde{\delta}}}_n$ be positive sequences converging to 0 as $n \to \infty$,*

$$\eta^n(R, P_{XY}, Q_{XYU}, \phi) = \begin{cases} D(Q_{XYU}||P_{XY}Q_{U|X}) - \delta_n & \text{if } \mathbb{E}_Q[d(X, \phi(Y,U))] \geq \Delta \\ D(Q_{XYU}||P_{XY}Q_{U|X}) - \delta_n + [R & \text{if } \mathbb{E}_Q[d(X, \phi(Y,U))] < \Delta \\ \quad -I_Q(X;U) + I_Q(Y;U) - \tilde{\delta}_n]^+ - \tilde{\tilde{\delta}}_n & \text{and } \kappa_2^n(Q_X) + \lambda_n \geq R - \tilde{\tilde{\tilde{\delta}}}_n \\ \infty & \text{otherwise,} \end{cases}$$

$$\beta^n(R, \Delta, P_{XY}, d) = \min_{Q_X} \max_{Q_{U|X}} \min_{Q_Y} \max_{\phi} \min_{Q_{XYU}} \eta^n(R, P_{XY}, Q_{XYU}, \phi)$$

$$\eta(R, P_{XY}, Q_{XYU}, \phi) = \begin{cases} D(Q_{XYU}||P_{XY}Q_{U|X}) & \text{if } \mathbb{E}_Q[d(X, \phi(Y,U))] \geq \Delta \\ D(Q_{XYU}||P_{XY}Q_{U|X}) + & \text{if } \mathbb{E}_Q[d(X, \phi(Y,U))] < \Delta \\ \quad \{R - I_Q(X;U) + I_Q(Y;U)\}^+ & \text{and } \kappa_2(Q_X) \geq R \\ \infty & \text{otherwise} \end{cases}$$

and $\beta(R, \Delta, P_{XY}, d) = \inf_{Q_X} \sup_{Q_{U|X}} \inf_{Q_Y} \sup_{\phi} \inf_{Q_{XYU}} \eta(R, P_{XY}, Q_{XYU}, \phi)$.

Then

$$\liminf_{n\to\infty} \beta^n(R, \Delta, P_{XY}, d) \geq \beta(R, \Delta, P_{XY}, d)$$

(Note in β^n the maximizations are over types/conditional types and in β over distributions.)

Proof. One sees that $\kappa_2^n(Q_X) + \lambda_n = \kappa_2(Q_X) + o(n)$ is upper semicontinuous in Q_X, with this established the proof then follows a similar proof for the Wyner-Ziv error exponent in the previous appendix. □

Proof of Theorem 2. Define

$$\mathcal{E} = \{d(X^n, \hat{X}^n) > \Delta\},$$

then for our scheme we have

$$\begin{aligned} P_e = & \sum_{\mathbf{x},\mathbf{y},\mathbf{u}} \Pr(\mathcal{E}|X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F(Q_{\mathbf{xyu}})) \\ & \times \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F(Q_{\mathbf{xyu}})) \\ & + \sum_{\mathbf{x},\mathbf{y},\mathbf{u}} \Pr(\mathcal{E}|X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{\mathbf{xyu}})) \\ & \times \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{\mathbf{xyu}})). \end{aligned}$$

By definition, when F occurs the encoder sends an error symbol, which we assume leads to the distortion constraint being violated. Using this observation, and rewriting the above equation, first summing over types then over sequences gives

$$\begin{aligned} P_e \leq & \sum_{Q_{XYU}} \sum_{\mathbf{x},\mathbf{y},\mathbf{u} \in T_{Q_{XYU}}} \Big[\Pr(\mathcal{E}|X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{XYU})) \\ & \times \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{XYU})) \Big] \\ & + \sum_{Q_{XYU}} |T_{Q_{XYU}}| \Pr(F(Q_{XYU})). \end{aligned}$$

On account of the fact that $\Pr(F(Q_{XYU}))$ goes to zero super exponentially for any choice of Q_{XYU} and the fact that there are only exponentially many sequences and polynomially many types, the final summand can be safely ignored for the error exponent calculation. We use $a \preceq b$ to mean that

$$\limsup_{n\to\infty} \frac{1}{n} \log a \leq \limsup_{n\to\infty} \frac{1}{n} \log b.$$

Let

$$P(\mathbf{x}, \mathbf{y}, \mathbf{u}) = \Pr(X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{\mathbf{xyu}}))$$

and

$$P(\mathcal{E}|\mathbf{x}, \mathbf{y}, \mathbf{u}) = \Pr(\mathcal{E}|X^n = \mathbf{x}, Y^n = \mathbf{y}, U^n = \mathbf{u}, F^c(Q_{\mathbf{xyu}})).$$

We now group the summation according to the sets outlined at the start of this section. This gives

$$\begin{aligned} P_e \preceq \sum_{Q_X}\sum_{Q_Y}\Big[& \sum_{Q_{XYU}\in\mathcal{D}_1}\sum_{\mathbf{x},\mathbf{y},\mathbf{u}\in T_{Q_{XYU}}} P(\mathbf{x},\mathbf{y},\mathbf{u})P(\mathcal{E}|\mathbf{x},\mathbf{y},\mathbf{u}) \\ & + \sum_{Q_{XYU}\in\mathcal{D}_2}\sum_{\mathbf{x},\mathbf{y},\mathbf{u}\in T_{Q_{XYU}}} P(\mathbf{x},\mathbf{y},\mathbf{u})P(\mathcal{E}|\mathbf{x},\mathbf{y},\mathbf{u}) \\ & + \sum_{Q_{XYU}\in\mathcal{D}_3}\sum_{\mathbf{x},\mathbf{y},\mathbf{u}\in T_{Q_{XYU}}} P(\mathbf{x},\mathbf{y},\mathbf{u})P(\mathcal{E}|\mathbf{x},\mathbf{y},\mathbf{u}) \\ & + \sum_{Q_{XYU}\in\mathcal{D}_4}\sum_{\mathbf{x},\mathbf{y},\mathbf{u}\in T_{Q_{XYU}}} P(\mathbf{x},\mathbf{y},\mathbf{u})P(\mathcal{E}|\mathbf{x},\mathbf{y},\mathbf{u})\Big] \end{aligned}$$

where in the inner summations over Q_{XYU} on the sets $\mathcal{D}_i$, the types of Q_X and Q_Y are fixed to be those set by the outer summations. On the set $\mathcal{D}_1$, Lemma 61 implies the quantity $P(\mathbf{x},\mathbf{y},\mathbf{u})$ decays super exponentially. Since there are only polynomially many types and exponentially many sequences this term can therefore be safely ignored. On the set $\mathcal{D}_3$, conditional on the event $F^c(Q_{\mathbf{xyu}})$, the codeword can be decoded without error, and hence there is no error. Using the result of Lemmas 63 and 64 we therefore have

$$\begin{aligned} P_e \preceq \sum_{Q_X}\sum_{Q_Y}\Big[& \sum_{Q_{XYU}\in\mathcal{D}_2} \exp(-n[D(Q_{XYU}||P_{XY}Q_{U|X}) - \delta_n \\ & + [R - I_Q(X;U) + I_Q(Y;U) - \tilde{\delta}_n]^+ - \tilde{\tilde{\delta}}_n]) \\ & + \sum_{Q_{XYU}\in\mathcal{D}_4} \exp(-n[D(Q_{XYU}||P_{XY}Q_{U|X}) - \delta_n])\Big] \end{aligned}$$

where $\tilde{\tilde{\delta}}_n = -\frac{1}{n}\log(1-\exp_e(-(n+1)^2)$. Bounding the summands by their max-

imum value gives

$$\begin{aligned} P_e \preceq\ & |\mathcal{P}^n(\mathcal{X})| \max_{Q_X} |\mathcal{P}^n(\mathcal{Y})| \max_{Q_Y} |\mathcal{P}^n(\mathcal{X}\times\mathcal{Y}\times\mathcal{U})| \\ & \times \Big[\max_{Q_{XYU}\in\mathcal{D}_2} \exp(-n[D(Q_{XYU}||P_{XY}Q_{U|X}) - \delta_n \\ & \quad + [R - I_Q(X;U) + I_Q(Y;U) - \tilde{\delta}_n]^+ - \tilde{\tilde{\delta}}_n]) \\ & + \max_{Q_{XYU}\in\mathcal{D}_4} \exp(-n[D(Q_{XYU}||P_{XY}Q_{U|X}) - \delta_n])\Big] \end{aligned} \tag{C.3}$$

Let

$$\tilde{\tilde{\delta}}_n(Q_X) = \frac{1}{n}\log(\exp(n[\kappa_2^n(Q_X) + \lambda_n]) + 1) - (\kappa_2^n(Q_X) + \lambda_n)$$

and let $\tilde{\tilde{\delta}}_n$ be the maximum over $Q_X \in \mathcal{P}^n(\mathcal{X})$ of $\tilde{\tilde{\delta}}_n(Q_X)$; it follows that $\tilde{\tilde{\delta}}_n \to 0$. Adopting the definitions from the statement of Lemma 65 and using $a + b \leq 2\max(a,b)$ to combine the two sums of (C.3) gives

$$\begin{aligned} P_e \preceq\ & 2|\mathcal{P}^n(\mathcal{X})||\mathcal{P}^n(\mathcal{Y})||\mathcal{P}^n(\mathcal{X}\times\mathcal{Y}\times\mathcal{U})| \\ & \times \max_{Q_X}\max_{Q_Y}\max_{Q_{XYU}:Q_{U|X}=Q^*_{U|X}(Q_X)} \exp(-n[\eta^n(R, P_{XY}, Q_{XYU}, \phi)]) \end{aligned}$$

Finally, we can optimize over $Q^*_{U|X}$ and ϕ, and move the optimizations in the exponent to give

$$\begin{aligned} P_e \preceq\ & 2|\mathcal{P}^n(\mathcal{X})||\mathcal{P}^n(\mathcal{Y})||\mathcal{P}^n(\mathcal{X}\times\mathcal{Y}\times\mathcal{U})| \\ & \times \exp(-n[\min_{Q_X}\max_{Q_{U|X}}\min_{Q_Y}\max_{\phi}\min_{Q_{XYU}} \eta^n(R, P_{XY}, Q_{XYU}, \phi)]). \end{aligned}$$

Taking the log, dividing by $-n$ and then taking the $\liminf_{n\to\infty}$ of both sides, invoking Lemma 65 on the righthand side gives the result. □

BIBLIOGRAPHY

[1] J. Neyman and E. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference: Part I," *Biometrika*, vol. 20A, no. 1/2, pp. 175–240, Jul 1928.

[2] D. Williams, *Probability with Martingales*. Cambridge University Press, 1991.

[3] R. H. Baayen, *Word Frequency Distributions*. Kluwer Academic Press, 2001.

[4] G. K. Zipf, *Selected Studies of the Principle of Relative Frequency in Language*. Harvard university Press, 1932.

[5] B. Mandelbrot, "Information theory and psycholinguistics: a theory of word frequencies," in *Readings in Mathematical Social Science*, P. Lazarfeld and N. Henry, Eds. MIT Press.

[6] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Stat.*, Jan 1952.

[7] ——, "Large-sample theory: Parametric case," *Ann. Math. Stat.*, Jan 1956.

[8] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Stat.*, Apr 1965.

[9] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 401 – 408, Mar 1989.

[10] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inf. Theory*, vol. 34, no. 2, pp. 278 – 286, Mar 1988.

[11] M. Feder and N. Merhav, "Universal composite hypothesis testing: a competitive minimax approach," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, pp. 1504 – 1517, 2002.

[12] A. Barron, "Uniformly powerful goodness of fit tests," *Ann. Stat.*, vol. 17, no. 1, pp. 107–124, Mar 1989.

[13] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4750 – 4755, Oct 2008.

[14] M. S. Ermakov, "Asymptotic minimaxity of chi-square tests," *Theory Probab. Appl.*, vol. 42, no. 4, pp. 589–610, 1998.

[15] L. Holst, "Asymptotic normality and efficiency for certain goodness-of-fit tests," *Biometrika*, vol. 59, no. 1, pp. 137–145, Apr 1972.

[16] M. Quine and J. Robinson, "Efficiencies of chi-square and likelihood ratio goodness-of-fit tests," *Ann. Stat.*, vol. 13, no. 2, pp. 727–742, Jun 1985.

[17] W. Kallenberg, "On moderate and large deviations in multinomial distributions," *Ann. Stat.*, vol. 13, no. 4, pp. 1554–1580, Dec 1985.

[18] T. R. Read and N. A. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer-Verlag, 1988.

[19] P. Harremoës and I. Vajda, "On Bahadur efficiency of power divergence statistics," 2010, submitted to IEEE Trans. Inf. Theory.

[20] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Probability estimation in the rare-events regime," *IEEE Trans. Inf. Theory*, to appear.

[21] N. Santhanam, A. Orlitsky, and K.Viswanathan, "New tricks for old dogs: Large alphabet probability estimation," in *Information Theory Workshop, 2007. ITW '07. IEEE*, 2007, pp. 638 – 643.

[22] J. Acharya, H. Das, A. Orlitsky, S. Pan, and N. P. Santhanam, "Classification using pattern probability estimators," in *IEEE International Symposium on Information Theory*, 2010, pp. 1493–1497.

[23] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98*, ser. Lecture Notes in Computer Science, C. Ndellec and C. Rouveirol, Eds. Springer Berlin / Heidelberg, 1998, vol. 1398, pp. 137–142, 10.1007/BFb0026683.

[24] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[25] C. E. Shannon, "A mathematical theory of communication," *Bell Sys. Tech. J.*, vol. 27, pp. 379–423,623–656, 1948.

[26] J. Ziv, "Coding of sources with unknown statistics–I: Probability of encoding error," *IEEE Trans. Inf. Theory*, vol. 18, no. 3, pp. 384 – 389, 1972.

[27] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic Press, 1981.

[28] A. Orlitsky and N. P. Santhanam, "Speaking of infinity," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2215 – 2230, 2004.

[29] W. Szpankowski and M. J. Weinberger, "Minimax redundancy for large alphabets," in *Proc. IEEE Int. Symp. Inf. Theory 2010*, 2010, pp. 1488 – 1492.

[30] L. D. Davisson, "Universal noiseless coding," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 783 – 795, 1973.

[31] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 6, pp. 674 – 682, 1978.

[32] B. Ryabko, J. Astola, and A. Gammerman, "Adaptive coding and prediction of sources with large and infinite alphabets," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3808 – 3813, Aug 2008.

[33] T. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 752 – 772, 1993.

[34] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 4, pp. 471 – 480, July 1973.

[35] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken: John Wiley & Sons, 2006.

[36] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, January 2005.

[37] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3037–3063, September 2005.

[38] M. Westover and J. O'Sullivan, "Achievable rates for pattern recognition," *IEEE Trans. Inf. Theory*, vol. 54, no. 1, pp. 299 – 320, Jan 2008.

[39] A. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 294 – 300, May 1975.

[40] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Trans. Inf. Theory*, vol. 21, no. 6, pp. 629 – 637, Jan 1975.

[41] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder II: General sources," *Inform. Contr.*, vol. 38, pp. 60–80, July 1978.

[42] A. Sahai and S. Mitter, "The necessity and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link-part I: Scalar systems," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3369–95, Aug. 2006.

[43] C. Cheng and A. Sahai, "Trade-off of lossless source coding error exponents," in *IEEE International Symposium on Information Theory*, 2008, pp. 1528–32.

[44] E. Arutyunyan and R. S. Marutyan, "E-optimal coding rates of a randomly varying source with side information at the decoder," *Problemy Peredachi Informatsii*, vol. 25, no. 4, pp. 24–34, 1989.

[45] E. Arutyunyan, Private Communication, 2007.

[46] K. Eswaran and M. Gastpar, "Achievable error exponents in multiterminal source coding," in *Proceedings of the 40th Conference on Information Sciences and Systems (CISS 2006)*, Princeton, NJ, March 2006.

[47] S. Jayaraman, "On error exponents in multiterminal source coding and hypothesis testing," Master's thesis, Cornell University, 1995.

[48] T. S. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2300 – 2324, Oct 1998.

[49] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source cod-

ing and channel coding and its extension to the side information case," *IEEE Trans. Inf. Theory*, vol. 49, no. 5, pp. 1181–203, May 2003.

[50] T. Liu, P. Moulin, and R. Koetter, "On error exponents of modulo lattice additive noise channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 454 – 471, Feb 2006.

[51] P. Moulin and Y. Wang, "Capacity and random-coding exponents for channel coding with side information," *IEEE Trans. Inf. Theory*, vol. 53, no. 4, pp. 1326–47, April 2007.

[52] R. G. Gallager, "Source coding with side information and universal coding," 1976, M.I.T. LIDS-P-937.

[53] I. Csiszár and J. Körner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 5 – 12, Jan 1981.

[54] B. Kelly and A. B. Wagner, "Improved Slepian-Wolf exponents via Witsenhausen's rate," in *IEEE International Symposium on Information Theory*, 2009, pp. 874–878.

[55] I. Csiszár and J. Körner, "Towards a general theory of source networks," *IEEE Trans. Inf. Theory*, vol. 26, no. 2, pp. 155–165, March 1980.

[56] I. Csiszár, "Linear codes for sources and source networks: Error exponents, universal coding," *IEEE Trans. Inf. Theory*, vol. 28, no. 4, pp. 585–592, July 1982.

[57] Y. Oohama and T. S. Han, "Universal coding for the Slepian-Wolf data compression system and the strong converse theorem," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1908–1919, November 1994.

[58] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. 20, no. 2, pp. 197–199, March 1974.

[59] S. Ihara and M. Kubo, "Error exponent for coding of memoryless Gaussian sources with a fidelity criterion," *IEICE Trans. Fundamentals*, vol. E83-A, pp. 1891–1897, 2000.

[60] H. Witsenhausen, "The zero-error side information problem and chromatic numbers (corresp.)," *IEEE Trans. Inf. Theory*, vol. 22, no. 5, pp. 592 – 593, Jan 1976.

[61] J. Körner and A. Orlitsky, "Zero-error information theory," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2207 – 2229, Oct 1998.

[62] J. Körner and G. Longo, "Two-step encoding for finite sources," *IEEE Trans. Inf. Theory*, vol. 19, no. 6, pp. 778 – 782, Jan 1973.

[63] G. Simonyi, "On Witsenhausen's zero-error rate for multiple sources," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3258 – 3260, Dec 2003.

[64] J. Körner, "Coding of an information source having ambiguous alphabet and the entropy of graphs," in *Trans. 6th Prague Conf. Information Theory, etc.* Prague, Czechoslovakia: Academia, 1973, pp. 411–425.

[65] E. V. Khmaladze, "Statistical analysis of large number of rare events," Centre for Mathematics and Computer Science, Netherlands, Tech. Rep. MS-R8804, 1988.

[66] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, 1986.

[67] Y. Ritov and P. Bickel, "Achieving information bounds in non and semi-parametric models," *Ann. Stat.*, Jan 1990.

[68] B. Efron and C. Stein, "The jackknife estimate of variance," *Ann. Stat.*, vol. 9, no. 3, pp. 586 – 596, May 1981.

[69] J. M. Steele, "An Efron-Stein inequality for nonsymmetric statistics," *Ann. Stat.*, vol. 14, no. 2, pp. 753 – 758, Jun 1986.

[70] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1602 – 1609, 2000.

[71] A. van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 1998.

[72] E. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, 3rd ed. Springer, 2005.

[73] L. Le Cam and G. L. Yang, *Asymptotics in Statistics: Some Basic Concepts*. Springer-Verlag, 1990.

[74] A. Cardoso-Cachopo, "Datasets for single-label text categorization," http://web.ist.utl.pt/ acardoso/datasets/, 2007.

[75] C. McDiarmid, "On the method of bounded differences," in *In: Surveys and Combinatorics 1989*. Cambridge University Press, 1989, pp. 148–188.

[76] S. Jayaraman and T. Berger, "An error exponent for lossy source coding with side information at the decoder," in *IEEE International Symposium on Information Theory*, 1995, p. 263.

[77] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, January 1976.

[78] A. Orlitsky and J. Roche, "Coding for computing," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 903 – 917, Mar 2001.

[79] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder. II. general sources," *Inform. Contr.*, vol. 38, no. 1, pp. 60–80, Jul. 1978.

[80] J. Justesen and T. Høholdt, "Maxentropic markov chains (corresp.)," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 665 – 667, 1984.

[81] K. Marton, "On the Shannon capacity of probabilistic graphs," *J. Combin. Theory Ser. B*, pp. 183 – 195, Jan 1993.

[82] I. Csiszár, J. Körner, L. Lovász, K. Marton, and G. Simonyi, "Entropy splitting for antiblocking corners and perfect graphs," *Combinatorica*, vol. 10, no. 1, pp. 27–40, Mar 1990.

[83] R. Diestel, *Graph Theory*. Springer, 2000.

[84] J. Moon and L. Moser, "On cliques in graphs," *Israel J. Math.*, vol. 3, no. 1, pp. 23 – 28, 1965.

[85] P. Koulgi, E. Tuncel, S. Regunathan, and K. Rose, "On zero-error source coding with decoder side information," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 99 – 111, Jan 2003.

[86] V. Anantharam, "Error exponents in a source coding problem of Körner," *J. Combin. Inform. System Sci.*, vol. 20, no. 1 – 4, pp. 141 – 152, 1995.

[87] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: A survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79 – 127, Jan 2006.

[88] T. Chalker, A. Godbole, P. Hitczenko, and J. Radcliff, "On the size of a random sphere of influence graph," *Adv. Appl. Prob.*, Jan 1999.

[89] A. Godbole and P. Hitczenko, "Beyond the method of bounded differences," in *Microsurveys in Probability*, 1998, pp. 1528–32.

[90] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.

[91] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[92] E. Arikan and N. Merhav, "Guessing subject to distortion," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1041 – 1056, May 1998.