The impact of causality on information-theoretic source and channel coding problems

Harikrishna R Palaiyanur

May 13, 2011

The impact of causality on information-theoretic source and channel coding problems

by

Harikrishna R. Palaiyanur

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

in

Engineering — Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Anant Sahai, Chair
Professor Kannan Ramchandran
Professor Jim Pitman

Spring 2011

The impact of causality on information-theoretic source and channel coding problems

Abstract

The impact of causality on information-theoretic source and channel coding problems

by

Harikrishna R. Palaiyanur

Doctor of Philosophy in Engineering — Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Anant Sahai, Chair

This thesis studies several problems in information theory where the notion of causality comes into play. Causality in information theory refers to the timing of when information is available to parties in a coding system.

The first part of the thesis studies the error exponent (or reliability function) for several communication problems over discrete memoryless channels. In particular, it studies an upper bound to the error exponent, or equivalently, a lower bound to the error probability of general codes, called the Haroutunian exponent. The Haroutunian exponent is the best known upper bound to the error exponent for two channel coding problems: fixed blocklength coding with feedback and fixed delay coding without feedback. For symmetric channels like the binary symmetric channel and the binary erasure channel, the Haroutunian exponent evaluates to the sphere-packing exponent, but for asymmetric channels like the Z-channel, the Haroutunian exponent is strictly larger than the sphere-packing exponent. The reason for the presumed looseness of the Haroutunian exponent is that it assumes, despite the inherent causality of feedback, a code might be able to predict future channel behavior based on past channel behavior and accordingly tune its input distribution. Intuitively, this kind of noncausal information should not be available to an encoder when the channel is memoryless. While we have not been able to tighten the Haroutunian exponent to the sphere-packing exponent for fixed blocklength codes with feedback, we describe some attempts made at bridging the gap. Additionally, we describe how to tighten the upper bound for two cases when the encoder is somehow limited: if the encoding strategy is constrained to use fixed type inputs regardless of output sequence and if there is a delay in the feedback path. The latter of these results leads to the insight that the Haroutunian exponent of a parallel channel constructed of independent uses of the original asymmetric channel approaches the sphere-packing exponent of the original channel after normalization. This fact can then be used to show that the error exponent for fixed delay codes is upper bounded by the sphere-packing exponent.

The second part of the thesis studies lossy compression of the arbitrarily varying sources introduced by Berger in his paper entitled 'The Source Coding Game'. An arbitrarily varying

source is a model for a source that samples other subsources under the control of an agent called a switcher. Motivated by compression of active vision sources, we seek upper and lower bounds for the rate-distortion function of an arbitrarily varying source when the switcher has noncausal knowledge about the realizations of the subsources it samples from. We find that when the subsources are memoryless, noncausal knowledge of subsource realizations is strictly better than information about past subsource realizations only.

To Amma, Appa and Sarah.

Contents

Acknowledgments

I do not have the required faculty with words to express how grateful I am to have been advised by Anant Sahai, but I will try anyway. I have never met someone who combines such great technical horsepower with raw, unbridled passion for research. Additionally, Anant possesses a remarkable intellectual curiosity about the workings of the world that has rubbed off on me to some extent. When I came to Berkeley, I believed from my undergraduate education that being able to solve problems was the most important skill needed to do research. Over my time as a graduate student, Anant taught me that being able to ask questions was as important, if not more important, for conducting research. As a student of Anant's, I received all-encompassing training to improve my presentation, communication and writing skills. In a good way, Anant never let me feel as if there was not room for improvement, and I thank him for that. There were many times in my graduate career when I would get stalled or frustrated by the slow pace of progress I was making. Anant was always patient and encouraging, especially during these periods of high stress and low productivity. There were so many meetings with him, after which I would feel reinvigorated to tackle problems that I had almost given up on. It was here at Berkeley, with the help of Anant, that I developed the capability to 'just keep moving forward'.

In the course of my time at Berkeley, I have also been fortunate to have some wonderful collaborators. Interactions with Baris Nakiboglu of MIT sparked some of the channel coding investigations in the thesis. Cheng Chang helped quite a bit in the early versions of the source coding material. I also worked with Pulkit Grover, Kris Woyach and Rahul Tandra on interesting topics not included in the thesis. I would like to thank Professors Kannan Ramchandran, Michael Gastpar and Jim Pitman for serving on the Qualifying Exam committee and their useful comments. Professor Pitman's Stat 205 A and B courses were extremely useful for developing the probabilistic maturity needed to do research in information theory. I would also like to thank Cheng Chang, Gireeja Ranade and Pulkit Grover for reading chapters of the thesis. I also appreciate the late Sergio Servetto for introducing me to research in information theory when I was at Cornell.

Obviously, I would never even be at an institution as fine as Berkeley if it were not for my family. As a child, it is hard to appreciate how one's parents make decisions and sacrifices that profoundly affect one's future. I thank my parents, Ravi and Sumathi, for their love and support throughout my life. I also thank them for instilling in me a love for learning and appreciation for education at an early age. The song 'Everybody's Free (To Wear Sunscreen)' by Baz Luhrmann has a lyric that has stuck with me. It says "your choices are half chance, so are everybody else's". I take two cautionary messages from this lyric: don't be so quick to judge others for their plights and don't accept that all fortunes gained in your lifetime were derived from your own efforts. One thing you don't get to choose in life is your parents, and I got some great ones. I also got an OK brother. Just kidding, Shyam, I love you too. My grandparents also showered me with love and encouragement. Chengalpattu[1] Thatha

[1] For some reason, I never dropped the childhood habit of adding a geographical location as an adjective

gave me a formidable nose to face the world with and I wish I knew him through more than just photos. KK Nagar Patti taught me the utility of discipline and consistency. KK Nagar Thatha's affection and enthusiasm for life and the simple pleasures was infectious. Finally, my lone living grandparent, Chengalpattu (now Ekkatuthangal) Patti has been a huge inspiration, and is someone I can talk to for hours. I am so happy to make her proud by becoming, in her words, the first Ph.D. in the family.

Life at Berkeley has not been all about work. Since its inception, the people of Wireless Foundations have made it a wonderful place to work, think, talk and laugh. These people include(d), in no particular order, Dan Hazen, Sameer Vermani (who taught me to cook a few dishes), Amin Gohari, John Secord, Bobak Nazer (my social hour partner and fellow cookie connoisseur), Krish Eswaran, Anand Sarwate (the mother bear of WiFo, in an endearing way), Alex Dimakis, Cheng Chang (my pool playing partner), Rahul Tandra (my diligent workout buddy), Mubaraq Mishra, Galen Reeves (my trusted cubiclemate), Guy Bresler (who gave me some amazing pep talks), Kris Woyach, Gireeja Ranade, Se-Yong Park, Jiening Zhan, Jonathan Tsao, Venky Ekambaram, Sameer Pawar, Salim El Rouayheb, Parv Venkitasubramaniam, Sahand Neghaban, Dapo Omidiran (my pickup basketball buddy) and Pulkit Grover (the meanest guy I ever met, just kidding). My friends from high school, Nikhil Kothari, Arjun Banker and Sravi Chennupati, who became Bay Area transplants as well, were always good for a fun overnight trip to the city. To anyone I forgot, I apologize, but I certainly appreciate your friendship.

I fell in love with basketball as an undergrad, but I became addicted to it at the RSF in Berkeley. One of the things that made me feel like a part of the EECS community was playing intramural basketball with people from across the department: Alessandro Abate (who graciously invited me to play on the team after I sent a blind email to the department), Alvise Boniventi, Nate Pletcher (a true Indiana ballplayer), Dapo Omidiran, Eric Battenberg, Galen Reeves, Marshall Miller, Simone Gambini, Matt Pierson, Alessandro Uccelli, Sandeep Mohan and even Professor Seth Sanders. Thanks for the fun and competition guys.

Last, but certainly not least, I want to thank my girlfriend, Sarah Kloss. Her orthogonal pursuits have been a constant reminder that the world is much bigger and more beautiful than what I know and experience. More importantly, her unconditional love and support are priceless.

Finally, I would like to gratefully acknowledge the National Science Foundation for financial support of the research in this thesis, first in the form of a Graduate Research Fellowship and second in the form of grants when I was a GSR.

to a relationship for my grandparents.

Chapter 1

Introduction

In a theory of information, it is not surprising that the idea of causality naturally arises. This is because the utility of knowledge is based on many factors: what is learned, how much it costs, who learns it (i.e., what actions can they carry out with the information) and when the knowledge is learned. The last factor, the timing of when knowledge is learned, is the subject of this thesis.

In the popular vernacular, the concept of causality is about cause and effect. In the social and physical sciences, effects are observed and correlated with hypothetical causes, and one of the goals of science is to show that a hypothesized cause is actually causative of the effect. A related meaning is the idea that before one event happens, a different one must occur first. This meaning of causality is related to the causality in the information theoretic problems in this thesis. We will study source and channel coding problems where actions of agents in the problems are taken with varying levels of knowledge of the realizations of certain random variables in the problems. The amount of knowledge available before these actions are taken will determine whether we think of the problems as having a causal or noncausal nature.

The first part of the thesis continues the study of error exponents for point-to-point channel coding. In point-to-point channel coding, as shown in Figure 1.1, an encoder wishes to communicate a message over a noisy communication medium called a channel to a decoder. The message is assumed to be uniformly random from a finite set and the encoder maps the message to input symbols for the channel. The channel randomly (noisily) maps the channel input symbol to a channel output symbol according to a known conditional probability distribution, $W(y|x)$. The encoder has a certain number of uses of the channel (called blocklength) to communicate the message to the decoder reliably (with low probability of error). Shannon, in his seminal paper [1] launching the field of information theory, showed that there is a critical quantity called the capacity of the channel, $C(W)$, that determines how much information can be communicated reliably across the channel. He showed the capacity can be calculated as

$$C(W) = \max_{P} I(P, W),$$

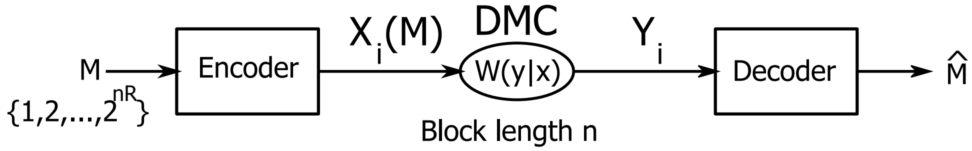

Figure 1.1: Fixed blocklength coding over a discrete memoryless channel $W(y|x)$. The rate of communication is R bits per channel use, and the blocklength is n, so the message is uniformly drawn from one of $\exp(nR)$ possibilities.

where P is a distribution on the input of the channel and $I(P, W)$ denotes the mutual information between the input and output of the channel W when the input distribution is P. The operational meaning of the capacity of the channel is that if the rate of communication (measured in bits communicated per use of the channel) is less than the capacity, there are codes that communicate the message with arbitrarily low probability of error in the limit of large blocklengths.

Because large blocklengths also correspond to large delay, one would also like to know how large a blocklength is needed to achieve a desired reliability while communicating at a given rate over the channel. The study of error exponents seeks to answer this question by analyzing the probability of error for optimal codes with a given blocklength and rate. It can be shown that for optimal codes,

$$\mathbb{P}(Error) \simeq e^{-nE(R)},$$

where n is the blocklength and $E(R)$ is the error exponent at rate R. One finds lower bounds to $E(R)$ by proving upper bounds to error probability for specific codes and upper bounds to $E(R)$ by finding lower bounds to error probability for general codes.

The first part of the thesis explores the noncausal interpretation of one upper bound, called the Haroutunian bound, to the error exponent for several variants of the communication problem just described. The original problem where the Haroutunian exponent appears is channel coding with casual output feedback. Shannon showed [2] that even if the encoder is given knowledge of the realizations of previous channel outputs before choosing a channel input at each time, the capacity of the (discrete, memoryless) channel is unchanged. Haroutunian [3] proved an upper bound to the error exponent for fixed blocklength codes with feedback that suffers from a technical weakness: it assumes that causal feedback allows the encoder to predict future channel behavior based on past channel behavior. The bound assumes that this knowledge can then be acted on by the encoder to improve error performance by optimizing its input distribution using this noncausal knowledge of the channel behavior it will face. This fact only makes the bound weak for asymmetric channels (channels for which the uniform input distribution is not optimal in some sense). The weakness of the bound lies in the fact that for memoryless channels, past channel behavior cannot be used to predict future channel behavior. Thus, the Haroutunian bound appears to grant the encoder with feedback a power it does not in reality possess. A more precise introduction to the problem is given in Chapter 2, but suffice to say for now that Chapters 2 and 3 are devoted to disallowing this possibliity that such noncausal knowledge could be learned by the code in several channel coding problems where the Haroutunian bound arises.

The second part of the thesis deals with lossy compression of arbitrarily varying sources (AVSs). Arbitrarily varying sources are used to model sources which output data that is actually a processed or dynamically sampled version of other 'subsources', as shown in Figure 1.2. An agent with some knowledge about the realizations and distributions of the subsources controls the processing or sampling operation. The concept of an arbitrarily varying source was introduced by Berger in his paper "The Source Coding Game" [4]. The source coding game is a game played between two players, a coder and a switcher. The coder is trying to lossily compress the output of the AVS to a specified distortion. The switcher is the agent in control of which subsource is sampled at each time, and in the game, is trying to make life difficult for the coder by forcing the coder to use as high a rate as possible. In this version of the game, the switcher is an adversary to the coder. Berger characterizes the rate-distortion function when the switcher is adversarial and has strictly causal knowledge of the realizations of memoryless subsources. That is, the switcher must set the switch position *before* learning of the subsource realizations at the current time. Berger asks the question of whether 'noncausal' knowledge of the subsource realizations can be used by the switcher to increase the rate-distortion function. It is this question that we answer in Chapter 4, both for the adversarial model of the switcher studied by Berger and a 'helpful' model we introduce, where the switcher is actually trying to help the coder use less rate.

In both parts of the thesis, we will model subsources and channels as discrete, memoryless systems. The impact of causality will then come through by how varying levels of knowledge for the agent in the two problems can shape the relevant controlled distributions. In the channel coding problem, the encoder is the agent in control of the output of the channel by using its input distribution as a 'control' input. In the source coding problem, the switcher is the agent in control of the output distribution of the AVS by using its knowledge of the subsource realizations and distributions to decide how to sample the subsource.

1.1 What could noncausal feedback do?

To get a glimpse of what can happen when we look at differing levels of causality in an information-theoretic problem, let us think of channel coding with feedback. The main weakness of the Haroutunian bound is that it assumes that causal output feedback can allow the encoder to tune its input distribution by predicting future channel behavior (because it is difficult to prove otherwise). Thus, the causal output feedback is giving some kind of 'precognitive' knowledge of future channel behavior. Why is this something to be afraid of to begin with, from the point of view of understanding fundamental behavior of optimal codes? One might guess that knowledge of the channel behavior at the encoder only is not

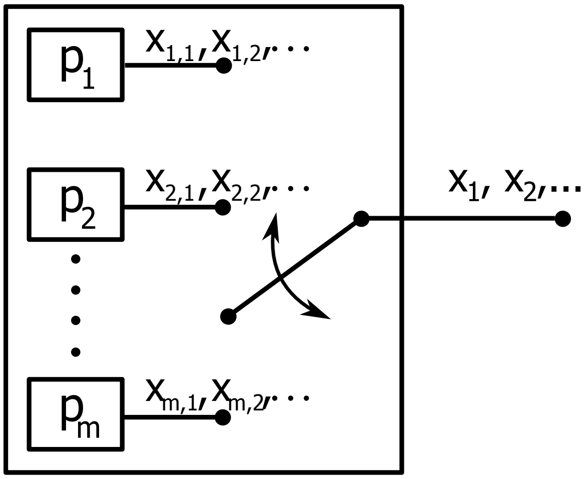
p_1
$x_{1,1}, x_{1,2}, \ldots$
p_2
$x_{2,1}, x_{2,2}, \ldots$
p_m
$x_{m,1}, x_{m,2}, \ldots$
$x_1, x_2, \ldots$

Figure 1.2: An arbitrarily varying source whose output symbol at each time is the output of one of a finite number of 'subsources'. In Berger's paper [4] and this thesis, the subsources are assumed to be discrete, memoryless and stationary. The switcher decides which subsources' symbol will be output based on the knowledge it has, which may be noncausal.

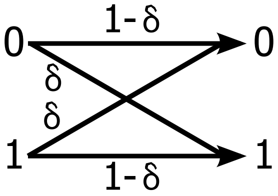

Figure 1.3: The binary symmetric channel with crossover probability δ. The output Y is the input X with probability $1-\delta$ and is $1-X$ with probability δ.

enough to improve performance because the decoder is still unaware of the channel noise the encoder faced, and therefore the lack of synchronicity might render the noncausal knowledge useless. To dispel this notion, we look into some models of feedback that appear aphysical and are noncausal, but do have applications in problems involving interference and storage on memories.

How does one begin to think of precognitive knowledge of channel behavior? For example, take the binary symmetric channel (BSC), shown in Figure 1.3, which flips its input with probability $\delta < 1/2$. The capacity of the BSC (without feedback or with causal output feedback) is $1 - h_b(\delta)$, where $h_b(\delta)$ is the binary entropy function $h_b(\delta) = -\delta \log \delta - (1 - \delta)\log(1-\delta)$. One model for the BSC is that the output Y is the modulo-2 sum of the input X an an independent Bernoulli random variable Z with parameter δ, $Y_i = X_i \oplus Z_i$. In this model, if the encoder knows the value of Z_i in a procognitive way before inputting X_i, clearly the capacity is 1 bit per channel use, which is larger than $1 - h_b(\delta)$. But this is not the only model for noise in the BSC. Consider a 'noisy packet drop' model for the BSC. Let $\widetilde{Z}_i$ be a Bernoulli process with parameter 2δ and $\widetilde{Y}_i$ an independent Bernoulli process with parameter 1/2. Then, let the output of the channel be

$$Y_i = X_i(1 - \widetilde{Z}_i) + \widetilde{Y}_i\widetilde{Z}_i.$$

This channel outputs the input X_i with probability $1 - 2\delta$ and disconnects the output from the input and outputs a random bit with probability 2δ. The effective channel, without knowledge of the $\widetilde{Z}_i$, is a BSC with crossover probability δ. It can be shown, however, that even if the encoder knows the value of the $\widetilde{Z}_i$ (but not the $\widetilde{Y}_i$), the capacity under this model is still $1 - h_b(\delta)$. So we have two models of channel noise that appear the same to the coding system without noncausal knowledge, but are different with noncausal knowledge. It is unclear how to generalize these two models to anything other than additive noise channels, so let us consider a model that is analogous to the switching model for source coding developed by Berger.

In channel coding, the agent capable of acting on feedback information is the encoder. In a sense, the goal of the encoder is to use the knowledge it has of the conditional distributions of the channel output given the channel inputs to 'control' the output of the channel and convey the message. One can then think of the channel output as being the switched output of $|\mathcal{X}|$ memoryless 'subchannel' outputs with distributions $\{W(\cdot|x) : x \in \mathcal{X}\}$, as shown in Figure 1.4. With causal output feedback, where the encoder learns Y_i immediately before deciding on the switch position X_{i+1}, the capacity of the channel is unchanged, as shown by Shannon [2].

What might a more general model that includes 'precognitive' feedback look like? In the traditional setup, there is only one channel output Y_i whose conditional distribution is $W(Y_i|X_i)$. We propose that precognitive feedback give the encoder advance knowledge of the realizations of the $|\mathcal{X}|$ subchannels. We will think of two cases of precognitive feedback: barely precognitive feedback if X_i is decided with knowledge of the realizations of the $|\mathcal{X}|$ subchannels up to and including time i, and fully precognitive feedback if X_i is decided with knowledge of the realizations of the $|\mathcal{X}|$ subchannels over the entire blocklength of n time steps. We are simply interested in knowing if this advance knowledge can be used by the encoder to increase capacity. Further, if it does increase capacity, does fully precognitive feedback increase capacity even more than barely precognitive feedback?

Luckily for us, these questions have already been answered if we transform the problem from one where the encoder has feedback to one where the encoder has advance knowledge of a channel 'state'. At time i, we let the state s_i be the realizations of the $|\mathcal{X}|$ subchannels, $s_i = (y_i[x] : x \in \mathcal{X})$. That is, the channel state tells the encoder exactly what the output of the channel will be if the input x is chosen for each input symbol. The probability the channel is in each state will be

$$P_S(s) = \mathbb{P}(s = (y[x] : x \in \mathcal{X})) = \prod_{x \in \mathcal{X}} W(y[x]|x),$$

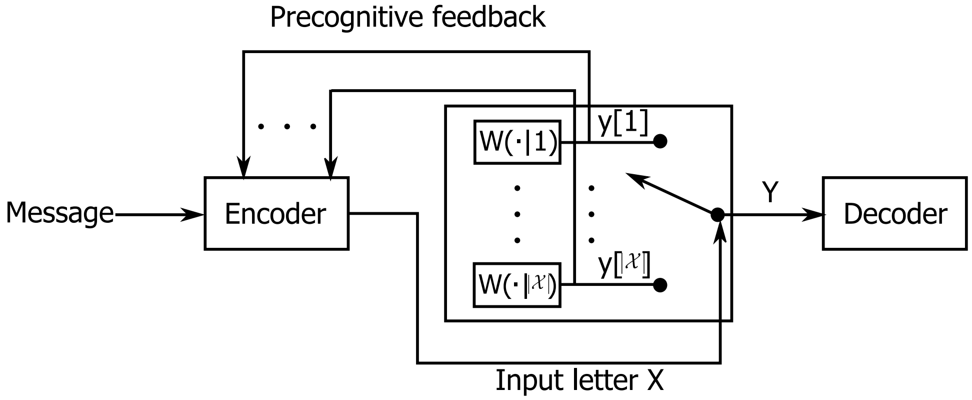

Figure 1.4: A way of thinking of channels analogous to the arbitrarily varying source. The channel is composed of $|\mathcal{X}|$ subchannels, each of which produce symbols from the output alphabet IID according to distributions $\{W(\cdot|x)\}$. The encoder then chooses which symbol is the output of the channel by selecting the input symbol. Without feedback, or with causal feedback, the channel appears to the encoder and decoder to be a DMC with transition probabilities $W(y|x)$. This model allows for a notion of 'precognitive feedback', where the encoder would be aware of the outputs of each of the subchannels before making a decision on the input symbol.

where for simplicity, we have made the assumption that the subchannels are independent of each other. We note that other models for a DMC can be recovered by making the subchannels correlated with each other (but independent over time). For example, the additive noise model for the BSC has fully correlated subchannels, where either both inputs are flipped or neither input is flipped.

Without any feedback or only causal knowledge of the channel state, the encoder appears to be facing a channel with conditional distribution W, because its knowledge of the output of the channel for a given input symbol is just the conditional distribution. With precognitive feedback, it knows the state and therefore the output of the channel deterministically given the input. Barely precognitive feedback then means that the encoder decides X_i with knowledge of $(s_1, \ldots, s_i)$ and fully precognitive feedback means the encoder chooses X_i with knowledge of $(s_1, \ldots, s_n)$. The capacity for these 'state knowledge' problems was determined by Shannon [5] in the barely precognitive case and Gelfand and Pinsker [6] in the fully precognitive case. These are also called channel side-information problems because the encoder gets side-information about the state of the channel before deciding on the input letter. However, in this precognitive feedback instance, the side-information is very special because given the side-information of the channel state, the output is a deterministic function of the input. The key insight of [5] is that, rather than thinking of inputs as being letters from $\mathcal{X}$, we think of inputs as being 'strategies', i.e., functions mapping an observed state to a

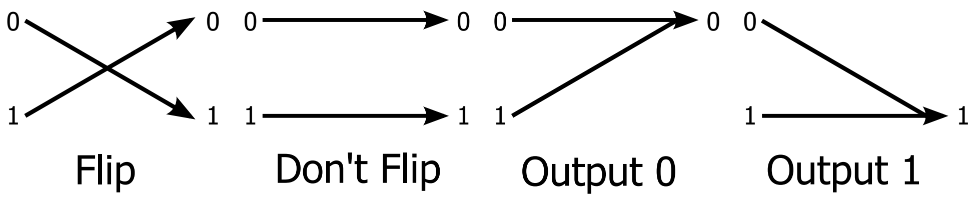

Figure 1.5: The possible states the BSC can be in for the model in Figure 1.4.

channel input letter,

$$t : \mathcal{S} \to \mathcal{X}.$$

In the case of barely precognitive feedback, the capacity as determined by Shannon is

$$C_{bp} = \max_{P_T} I(T;Y),$$

where the joint distribution is

$$\mathbb{P}(s,t,y) = P_S(s)P_T(t)1(y[t(s)] = y),$$

and $I(T;Y)$ is the mutual information between the random variables T and Y. The notation $y[x]$ denotes the component of s corresponding to the input x, which has marginal distribution $W(\cdot|x)$. In the fully precognitive case, Gelfand and Pinsker showed that the capacity is

$$C_{fp} = \max_{P_{T|S}} I(T;Y) - I(T;S),$$

where the joint distribution is

$$\mathbb{P}(s,t,y) = P_S(s)P_{T|S}(t|s)1(y[t(s)] = y)$$

and the optimization is over conditional distributions of the strategy T, conditioned on the state S.

Let us evaluate these capacities for two simple binary input, binary output channels: the binary symmetric channel (BSC) and the Z-channel. For the BSC, at each time, the channel can be in one of four states: flip both inputs with probability δ^2, don't flip the inputs with probability $(1-\delta)^2$, output only 0 with probability $\delta(1-\delta)$, or output only 1 with probability $\delta(1-\delta)$ as shown in Figure 1.5.

When the state of the channel is 'flip both inputs' or 'don't flip the inputs', the encoder with advance knowledge of the state can make the output of the channel be either 0 or 1. In the other states, the encoder is constrained to output the symbol that occurs as the realization of both subchannels. It is inconsequential what a strategy does in these constrained states

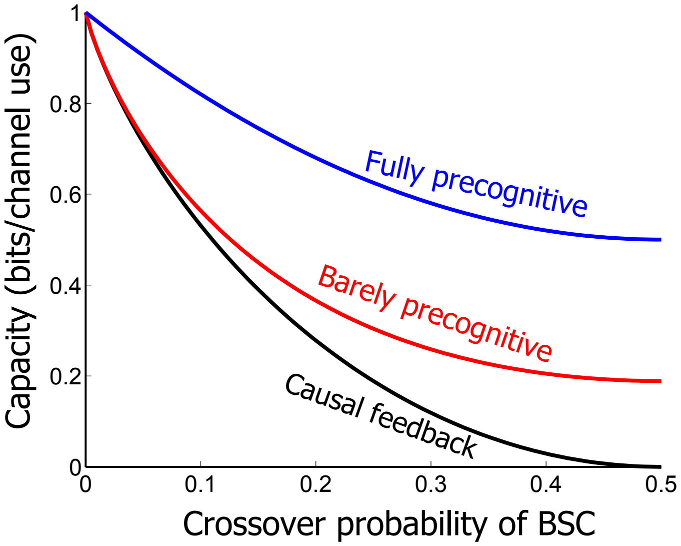

Figure 1.6: Capacity of the BSC with no feedback, barely precognitive feedback and fully precognitive feedback as a function of the crossover probability $\delta \in [0, 1/2]$. Fully precognitive feedback increases capacity over barely precognitive feedback, which in turn is larger than capacity without feedback.

as it cannot affect the output of the channel. It can be shown that, with barely precognitive feedback, the capacity of the channel is the capacity of a BSC with crossover probability $\delta(1-\delta)$, which is less than δ. Thus

$$C_{bp} = 1 - h_b(\delta(1-\delta)) > 1 - h_b(\delta).$$

In a paper that studies storage in memories with known defects, Heegard and El Gamal [7] show that with fully precognitive feedback, the encoder can communicate a bit for every time instant that the state allows the output of the channel to be chosen, even though the decoder does not know when these time instants occur. Thus,

$$C_{fp} = 1 - 2\delta(1-\delta) > C_{bp} > 1 - h_b(\delta).$$

Hence, we find that noncausal knowledge of channel behavior can indeed be used to increase capacity. In the example of the BSC, the difference between barely precognitive feedback and fully precognitive feedback is also an increase in capacity, as shown in Figure 1.6.

As another example, consider the Z-channel, a channel which faithfully transmits 0's as 0's, but flips 1's to 0's with probability δ. Without feedback, or with causal output feedback, the capacity of the Z-channel with crossover probability δ is

$$C_Z(\delta) = h_b\left(p^*(\delta)(1-\delta)\right) - p^*(\delta)h_b(\delta),$$

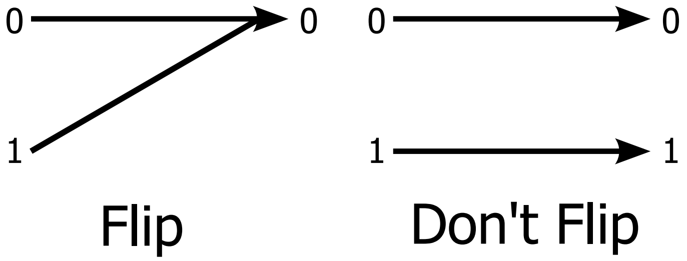

Figure 1.7: The states that a Z-channel can be in. A 0 is always transmitted as a 0, but a 1 can either be transmitted as a 1 or flipped to a 0.

where

$$p^*(\delta) = \left[(1-\delta) \left(1 + \exp\left(\frac{h_b(\delta)}{1-\delta} \right) \right) \right]^{-1}.$$

With precognitive feedback, the encoder is made to know in advance whether the channel will flip an input of 1 to 0 or not, as shown in Figure 1.7. With barely precognitive feedback, it can be checked that the capacity is unchanged, i.e., $C_{bp} = C_Z(\delta)$. With fully precognitive feedback, the capacity actually increases to $C_{fp} = 1 - \delta$ (bits per channel use), the fraction of time the channel output can be freely chosen to be either 0 or 1. Hence, for the Z-channel, knowledge of the channel state immediately before deciding the input does not increase capacity (as shown in Figure 1.8), but knowledge of all future channel behavior does increase capacity. These examples show that:

1. Noncausal knowledge can have a large impact on the fundamental limits of performance in information theoretic problems.

2. The degree of noncausality can also have an impact for some problems, but not for others, even when the underlying randomness is memoryless.

Jafar [8] has shown that for encoders with channel side-information, even one bit of side-information about the channel state given to the encoder before choosing the input can increase capacity by an unbounded amount. So we see that in terms of increasing capacity, the future can start now (as in the BSC example) or later (as in the Z-channel example) and knowledge of the future can increase capacity by an unbounded amount.

The purpose of these examples is to foreshadow that causality can indeed have a big on information-theoretic problems. We will see this quite clearly in the source coding problem by the impact on the rate-distortion function. For the Haroutunian exponent, we want to rule out that a code with causal feedback might be able to improve its error performance by tuning its input distribution to future channel behavior. To avoid confusion, we note that the Haroutunian exponent does not assume that the encoder knows future channel behavior and thus can increase capacity. Rather, it can be interpreted as assuming that when the

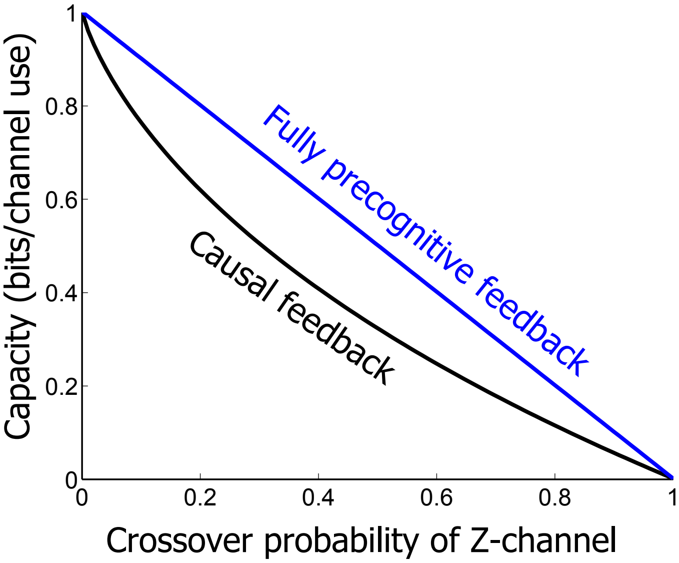

Figure 1.8: Capacity of the Z-channel with no feedback, barely precognitive feedback and fully precognitive feedback as a function of the crossover probability. Barely precognitive feedback does not increase the capacity, but fully precognitive feedback does.

channel behaves atypically, the encoder knows this in advance and can change the input distribution according to an optimal one (in a sense to be described in Chapter 2). The fact that the optimal input distribution changes with the channel is only true for asymmetric channels, where the uniform input distribution is not optimal.

1.2 Contributions

While we detail the contributions of the thesis precisely in the chapters where the results are given, here we briefly highlight those results. Chapter 2 studies the Haroutunian exponent as the upper bound to the error exponent for fixed blocklength codes with causal output feedback. The goal, as has been the case since the bound was first proved, is to show that the sphere-packing exponent, which is tighter than the Haroutunian exponent for asymmetric channels, is an upper bound to the error exponent. The sphere-packing exponent is also an upper bound for fixed blocklength codes without feedback, and is tight at rates near capacity, so this problem is really about showing that causal output feedback does not improve the asymptotic reliability of fixed blocklength codes.

Unfortunately, we have not succeeded in proving that the sphere-packing bound holds for fixed blocklength codes with feedback. Chapter 2 chronicles some of the attempts made and what technical obstacles were faced. We then show that tightening the error exponent from the Haroutunian exponent is possible in two restricted cases. First, if the encoder is constrained to use a fixed input distribution regardless of the feedback information it receives, the sphere-packing bound holds. Second, if there is a delay in the feedback path, the sphere-packing bound holds in the limit as the delay gets large. This result essentially makes the feedback even more causal and shows that the more stale the feedback information, the more we can prove that feedback cannot be used to optimally change the channel input distribution according to future channel behavior. We then reinterpret the delayed feedback result as applying to parallel channels with instantaneous feedback rather than the original channel with delayed feedback. This reinterpretation leads to the surprising result that the Haroutunian exponent of the parallel channel is not the addition of the Haroutunian exponent of the original channels for asymmetric channels W. In fact, after normalization, the Haroutunian exponent of the parallel channel approaches the sphere-packing exponent of W in the limit as the parallelization of the channels gets large. While this insight is not useful for proving stronger results for block codes with feedback, it can be applied to another problem where the Haroutunian exponent appears in Chapter 3. In fixed delay coding, bits arrive at the encoder in a streaming fashion as opposed to the message being known fully in advance for fixed blocklength codes. Each bit is then to be decoded within a fixed delay of the time it arrives at the encoder. The error exponent for fixed delay codes is taken with respect to delay, analogous to the blocklength for fixed blocklength codes. The Haroutunian exponent appears as the best known upper bound to the error exponent for fixed delay codes without feedback. While there is no feedback, the previously employed proof technique could not disallow the possibility that the encoder might be able to predict when the channel behavior is bad enough so that reliably communicating any given bit is hopeless. Using the insight for parallel channels developed in Chapter 2, we are able to tighten the upper bound to the sphere-packing exponent for fixed delay codes.

In Chapter 4, we are motivated by the field of computer vision and Berger's paper to study the arbitrarily varying source with new models of knowledge and intentions for the switcher. Extending the work of Berger, we first study the rate-distortion function of the AVS when the switcher is adversarial and has knowledge of the realizations of the memoryless subsources either immediately before selecting the switch position or fully in advance. This rate-distortion function is characterized completely as the maximum of the IID rate-distortion function over distributions the switcher can simulate at the output of the AVS. In this problem, the level of noncausality does not matter: the rate-distortion function increases when going from causal knowledge to barely noncausal, but no further increase occurs when the switcher receives fully noncausal knowledge of subsource realizations. We then characterize the rate-distortion function if the switcher is adversarial and has noisy and noncausal access to the subsource realizations. The next step is to then consider what happens if the switcher is actually helpful, the opposite of adversarial. If the switcher is helpful and has fully noncausal knowledge of subsource realizations, we fully characterize the rate-distortion function as the IID rate-distortion function for an associated source. With other levels of causality in knowledge, we give upper bounds on the rate-distortion function for helpful switching. Finally, to show that brute force computation of the $R(D)$ function of an AVS can be provably done in a finite amount of time, we prove a technical lemma about the uniform continuity of the IID rate-distortion function that may be of independent interest.

While the statements here may seem vague, we invite the reader to proceed to the chapter introductions for a more precise description of the contents of the thesis.

Chapter 2

The Haroutunian exponent for block coding with feedback

2.1 Introduction

This chapter explores a particular problem that arises in the classical information-theoretic study of block coding for discrete memoryless channels (DMCs). A DMC is a communication medium with a finite input space $\mathcal{X}$ and finite output space $\mathcal{Y}$ for which, at any time, if the input to the channel is $x \in \mathcal{X}$, the output of the channel is $y \in \mathcal{Y}$ with probability $W(y|x)$, where W is a probability transition matrix. Block coding refers to the assumption that the entire message to be communicated is known before transmission is commenced for a fixed (pre-determined) amount of time[1]. The duration of transmission for a (message) block, in number of channel uses, is called the blocklength.

For this problem, Shannon [1] showed that the capacity of the channel, $C(W)$, is a fundamental quantity that determines quite precisely the number of bits per channel use that can be reliably communicated in the limit of long blocklengths. In this first-order notion of reliable communication, all that is required is that the probability of incorrect decoding of the message goes to 0 with longer and longer blocklengths. Soon after the first rigorous proof

[1] In the anytime coding chapter, we study communication systems where the message to be communicated is causally revealed to the transmitter. There are also notions of variable-length codes ([9], [10], [11]) that allow for the transmission duration to depend on the quality of the channel. In this thesis, we do not consider variable-length codes, only fixed-length codes.

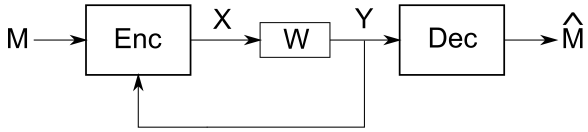

Figure 2.1: Fixed-length block coding with feedback is the problem considered in this chapter.

by Feinstein [12] that $C(W)$ was the highest rate one could reliably communicate messages over W, it was also proved by Feinstein [13] that the reliability, or probability of error, decays exponentially to 0 with the blocklength for any rate below capacity. This result set afire the study of the *reliability function* for communication over DMCs, a second-order notion of the quality of the channel W for communication purposes.

Around the same time as the reliability function was being investigated, there was also heavy interest in understanding the fundamental information-theoretic limits on communication when the additional resource of *feedback* is available. In many practical communication situations, the receiver may be able to send information to the transmitter while the transmitter is attempting to send a message to the receiver. In order to simplify the practical aspects of the problem (i.e., the feedback is noisy, rate-limited and/or delayed), efforts were concentrated on understanding what can happen when the transmitter is made aware of the channel outputs (received symbols) at the same time that the receiver is made aware of them. This is the case of *perfect feedback*; perfect because the received symbol is available to the transmitter both noiselessly and without delay.

Shannon himself [2] showed that for DMCs, feedback does not increase the capacity of the channel. The natural question at this point was

Does feedback increase the reliability function of a DMC?

The answer to this question is not as simple as the answer that Shannon provided for capacity. It turns out that if the channel is symmetric (e.g. the binary symmetric channel or binary erasure channel), in a sense to be defined later, feedback does not increase the reliability function much. However, if the channel is asymmetric, the answer to this question is still unknown. This leaves us with one of two possibilities: either feedback does not improve reliability for asymmetric channels and this is difficult to prove, or feedback can improve reliability only by exploiting asymmetry of the channel.

In order to get to the point where we can more meaningfully discuss the contents of this chapter, a brief review of the literature on the reliability function (also known as the error exponent) is in order.

2.1.1 A brief history of error exponents for block coding over DMCs

If we let $P_e(n, R)$ denote[2] the lowest error probability of all block codes without feedback of blocklength n and rate at least R that can be used for communicating over W, the error

[2] The notation, for now, suppresses the dependence of the error probability on the channel W.

exponent or reliability function is defined as[3]

$$E(R) \triangleq \lim_{n\to\infty} -\frac{1}{n}\log P_e(n,R).$$

As noted earlier, Feinstein [13] showed that $E(R)$ is positive for all $R < C(W)$. If we let $P_{e,fb}(n,R)$ denote the lowest error probability for all block codes with feedback of blocklength n and rate at least R that can be used for communicating over W, the error exponent or reliability function with feedback is defined as

$$E_{fb}(R) \triangleq \lim_{n\to\infty} -\frac{1}{n}\log P_{e,fb}(n,R).$$

Because a code with feedback can simply choose to ignore the feedback, it trivially follows that $P_{e,fb}(n,R) \le P_e(n,R)$, and therefore,

$$E(R) \le E_{fb}(R).$$

We will say that feedback does not improve reliability if $E(R) = E_{fb}(R)$, even though it might be possible that $P_e(n,R)$ can be strictly larger then $P_{e,fb}(n,R)$ in this case. More precisely, what we mean is that feedback does not improve the reliability function.

As an aside, the notion of an error exponent or reliability function is only useful if $-(1/n)\log P_e(n,R)$ approaches $E(R)$ fairly quickly (and similarly for $E_{fb}(R)$). The results discussed in this chapter typically have a 'convergence rate' of $O(\sqrt{\log(n)/n})$, meaning that if $E_l(R)$ is a lower bound to $E(R)$ and $E_u(R)$ is an upper bound to $E(R)$, we can say that

$$\exp\left(-n\left[E_u(R) + O\left(\sqrt{\frac{\log n}{n}}\right)\right]\right) \le P_e(n,R) \le \exp\left(-n\left[E_l(R) - O\left(\sqrt{\frac{\log n}{n}}\right)\right]\right).$$

Error exponents without feedback

In order to determine $E(R)$, researchers set out to find upper and lower bounds to $P_e(n,R)$ (yielding lower and upper bounds to $E(R)$ respectively). One of the first ways to find lower bounds to the error exponent[4] was to analyze the performance of *random codes*. Upper

[3] Throughout this chapter and Chapter 3, the log and exp functions are taken to the base e, but in plots the units may be to the base 2 (as noted in each plot). Also, we note that the limit is presumed to exist, but it is not known for sure at some rates if it does. In seeking bounds to the error exponent, bounds to the lim inf and lim sup are sought.

[4] The term error exponent refers both to the reliability function $E(R)$ and to exponents that serve as upper or lower bounds to $E(R)$. When we say 'the error exponent', we are referring to the reliability function, but when we say 'an error exponent' or 'error exponents', we mean upper and lower bounds to the reliability function. Hopefully, this distinction is clear from context.

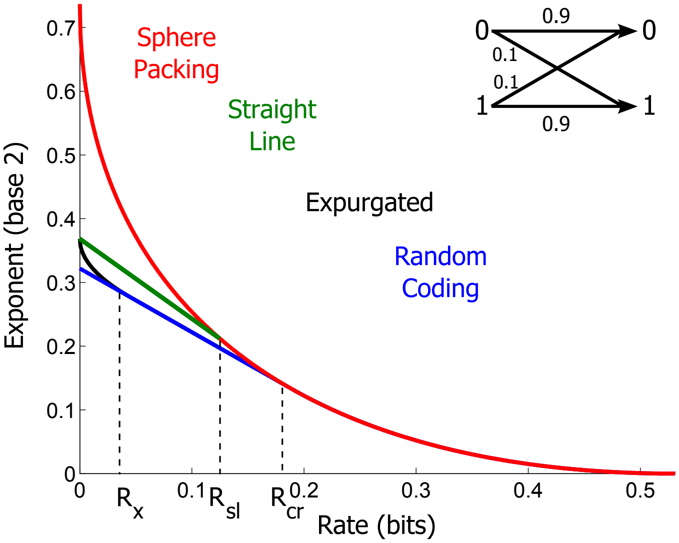

Figure 2.2: Error exponents without feedback for a BSC with crossover probability 0.1. The capacity for this channel in bits is 0.53 bits per channel use. The random coding exponent $E_r(R)$, expurgated exponent $E_{ex}(R)$, straight line exponent $E_{sl}(R)$ and sphere-packing exponent $E_{sp}(R)$ are shown (they are individually denoted by color, but can be identified also by the ordering $E_r(R) \leq E_{ex}(R) \leq E_{sl}(R) \leq E_{sp}(R)$). For this channel, $R_{cr} = 0.1881$ bits per channel use, $R_x = 0.0452$ bits per channel use and $E_{ex}(0) = 0.3685$, which is much less than $E_{sp}(0) = 0.737$. Thus, the straight line bound is significantly tighter for rates between 0 and $R_{sl} = 0.13$ bits per channel use.

bounds to the error probability of random codes yielded the random coding exponent, $E_r(R)$, where

$$\begin{aligned} E(R) &\geq E_r(R) \\ E_r(R) &\triangleq \max_{\rho \in [0,1]} \max_P E_0(\rho, P) - \rho R \qquad (2.1) \\ E_0(\rho, P) &\triangleq -\log \sum_y \left[\sum_x P(x) W(y|x)^{\frac{1}{1+\rho}} \right]^{1+\rho}, \qquad (2.2) \end{aligned}$$

with P denoting a distribution (probability mass function) on $\mathcal{X}$. Elias [14] showed that $E(R) \geq E_r(R)$ if W is a binary symmetric channel (BSC) or binary erasure channel (BEC), while Fano [15] showed that $E(R) \geq E_r(R)$ for a general DMC. Gallager [16] gave a simple derivation of the random coding exponent bound and derived most of the useful properties of the function $E_r(R)$.

Since any realization of a random code can be bad because the codewords drawn could be identical to each other or otherwise 'too close', Gallager ([16], [17]) improved on random codes by expurgating these 'bad' codewords and showed that the resulting expurgated exponent, $E_{ex}(R)$, is a lower bound to $E(R)$, where

$$\begin{aligned} E_{ex}(R) &\triangleq \sup_{\rho \geq 1} \max_P E_x(\rho, P) - \rho R \\ E_x(\rho, P) &\triangleq -\rho \log \sum_{x, x' \in \mathcal{X}} P(x) P(x') \left[\sum_y \sqrt{W(y|x) W(y|x')} \right]^{1/\rho}. \end{aligned}$$

In general, there is an $R_x \in [0, C(W)]$ for which,

$$\begin{aligned} E_r(R) &= E_{ex}(R), \text{ if } R \in [R_x, C(W)] \\ E_r(R) &< E_{ex}(R), \text{ if } R \in [0, R_x), \end{aligned}$$

so $E_{ex}(R)$ improves on $E_r(R)$ only for low rates.

For upper bounding the reliability function, the most fundamentally important upper bound is the sphere-packing exponent[5], $E_{sp}(R)$,

$$E_{sp}(R) \triangleq \sup_{\rho \geq 0} \max_P E_0(\rho, P) - \rho R,$$

where $E_0(\rho, P)$ is defined in (2.2). The sphere-packing bound was first shown to hold by Elias [14] for rates close to capacity if W is a BSC or a BEC, and discovered by Fano [15] for general DMCs. The first rigorous proof of the sphere-packing bound for general DMCs was given by Shannon, Gallager and Berlekamp [18]. It was later independently recognized

[5] For a quick reminder of important notation, please flip ahead to Tables 2.1 and 2.2 in Section 2.2.

by Haroutunian [19] and Blahut [20] that the sphere-packing exponent could be expressed another way, as

$$E_{sp}(R) = \max_{P} \min_{V:I(P,V)\leq R} D(V||W|P), \tag{2.3}$$

$$= \max_{P} E_{sp}(R,P) \tag{2.4}$$

where $I(P,V)$ denotes the mutual information across the channel V when the input distribution is P and $D(V||W|P)$ denotes the conditional divergence between the two channels V and W when the input distribution is P. The form of the sphere-packing exponent above lends itself to a simple interpretation. One can prove the sphere-packing bound by the following process. Since the code is a block code with codewords fixed ahead of time, there is a subcode of rate nearly R whose codewords are all of some type P. Then, for this subcode and a test channel V that have a low mutual information, the error probability must be high (actually convergent to 1). The exponent that governs[6] the probability of the channel W 'acting like' channel V when the input type is P is $D(V||W|P)$.

From the so-called parametric-ρ form of $E_{sp}(R)$ and $E_r(R)$, and the properties of $E_0(\rho,P)$ derived by Gallager in [16], it can be shown that there is a critical rate, $R_{cr} \in [0, C(W)]$, such that $E_r(R) = E_{sp}(R)$ if R is at least R_{cr}. Therefore, for $R \geq R_{cr}$, the reliability function is pinned down to be $E(R) = E_r(R) = E_{sp}(R)$. For $R < R_{cr}$, we see that the reliability function is sandwiched between the random coding exponent and the sphere-packing exponent, $E_r(R) \leq E(R) \leq E_{sp}(R)$.

Another interesting fact about the error exponent for block codes without feedback can be deduced by studying the parametric ρ-form of $E_{sp}(R)$ for rates below R_{cr} ([17], Problem 5.20). If we relax the notion of decoding to allow a fixed number, L, of decoded messages (called list decoding), one can show that random codes with maximum-likelihood decoding to lists of size L achieves the following error exponent:

$$E_{r,L}(R) = \max_{\rho\in[0,L]} \max_{P} E_0(\rho,P) - \rho R.$$

By the properties of $\max_P E_0(\rho,P)$ as a function of ρ, it also follows that $E_{r,L}(R) = E_{sp}(R)$ for $R \geq R_{cr,L}$ and $R_{cr} = R_{cr,1} \geq R_{cr,2} \geq R_{cr,3} \geq \ldots$. Further, one can show that for each $R > 0$, there is an L' such that if $L \geq L'$, $E_{r,L}(R) = E_{sp}(R)$. Therefore, the sphere-packing bound is achievable with random codes and list decoding for a large enough (but finite and not growing with blocklength) list size. Hence, the gap between $E_{r,1}(R) = E_r(R)$ and the sphere-packing exponent is caused by the decoder being uncertain of the message up to just a few bits.

[6] This intuition is reminiscent of the analysis of the optimal asymptotic error probability for hypothesis testing between two distributions. In the limit as the number of observations go to infinity, the exponent of the error probability (say of deciding on P' when the true distribution is P) is the divergence $D(P'||P)$. This result is known as Stein's Lemma (Theorem 12.8.1 of [21]) and can be interpreted as forcing an error by making a random variable with distribution P behave like distribution P'.

Around the same time that the sphere-packing bound was being rigorously proved, a related upper bound to $E(R)$ called the straight-line bound, $E_{sl}(R)$, was derived by Shannon, Gallager and Berlekamp ([18], [22]). The straight-line bound takes any non-increasing upper bound to the reliability function, call it $E_u(R)$, and shows that for all $0 \leq R_1 \leq R_2 \leq C(W)$ and $\lambda \in [0, 1]$,

$$E\left(\lambda R_1 + (1-\lambda)R_2\right) \leq \lambda E_u(R_1) + (1-\lambda)E_{sp}(R_2). \tag{2.5}$$

The straight-line bound is generally used with

$$\begin{aligned} E_u(R) &= \max_P - \sum_{x,x' \in \mathcal{X}} P(x)P(x') \log \sum_y \sqrt{W(y|x)W(y|x')} \\ &= E_{ex}(0). \end{aligned}$$

The result that $E_{ex}(0)$ is an upper bound to the error exponent for zero-rate communication (i.e., the reliability function for codes with a fixed number of messages, as the number of messages goes to infinity) is also derived in [22]. Then, because it can be shown that $E_{ex}(0) < E_{sp}(0)$, and $E_{sp}(R)$ is convex-∪ in R, there is some $R_{sl} \in [0, C(W)]$ for which $R_2 = R_{sl}$ and $R_1 = 0$ gives the best upper bound to $E(R)$ from (2.5), so

$$E_{sl}(R) = \begin{cases} \frac{R_{sl}-R}{R_{sl}} E_{ex}(0) + \frac{R}{R_{sl}} E_{sp}(R_{sl}), & R \in [0, R_{sl}] \\ E_{sp}(R), & R > R_{sl}. \end{cases}$$

Note that because $E(R) = E_{sp}(R)$ for $R \in [R_{cr}, C(W)]$, $R_{sl} < R_{cr}$. Unfortunately, the straight-line bound does not have an intuitive interpretation like the sphere-packing bound (other than the obvious geometric interpretation).

For block codes without feedback, since 1968, the state of affairs has been that $E(R) = E_r(R) = E_{sp}(R)$ if $R \geq R_{cr}$ and $E_{ex}(R) \leq E(R) \leq E_{sl}(R)$ for $R \in [0, R_{cr}]$ with $E_{sl}(0) = E_{ex}(0)$. The four exponents discussed in this section are plotted in Figure 2.2 for a BSC with crossover probability 0.1.

Error exponents with feedback

While our focus is on upper bounds to $E_{fb}(R)$, we will briefly discuss lower bounds to $E_{fb}(R)$ that are different from the lower bounds to $E(R)$. Clearly, because a code with feedback can choose to ignore the feedback, $E(R) \leq E_{fb}(R)$.

An important coding scheme with feedback is called *posterior matching*. When posterior matching is used, the transmitter calculates the posterior probabilities of each message based on the received symbols and groups them in a particular way to determine the next input symbol. Zigangirov [23] showed that the error exponent of the posterior matching scheme with feedback, $E_{pm}(R)$, when used over BSCs has the property that $E_{pm}(R) = E_{sp}(R)$ for $R \geq R_{cr,fb}$, where $R_{cr,fb} < R_{cr}$. Therefore, posterior matching has a better error exponent than the best known schemes for codes without feedback: random coding and expurgation. Also, it was shown that $E_{pm}(0) > E_{ex}(0)$ for the BSC, so feedback at least improves the reliability function for very low rates. Further extensions of this result followed by Dyachkov [24] (showing how to perform posterior matching for arbitrary DMCs and analyzing its performance for a larger class of symmetric channels), Burnashev (adapting his two-phase approach for variable-length codes with feedback [9] to fixed-length codes with feedback for the BSC [25]) and Nakiboglu (extending the schemes of Dyachkov and Burnashev and improving their analyses).

Meanwhile, for upper bounds to $E_{fb}(R)$, Dobrushin [26] had shown that $E_{fb}(R) \leq E_{sp}(R)$ if the channel is symmetric at both the input and output[7]. This result showed that $E_{fb}(R) = E(R)$ for $R \geq R_{cr}$ for symmetric channels, so it seemed that feedback did not increase reliability (at least for 'high' rates).

Intuitively, the Dobrushin's result says that when dealing with 'additive noise' channels like the BSC, the important factor in determining the quality of the channel is the 'variance' or 'power' of the noise and not really what is happening at the input. From a coding theory perspective, the sphere-packing bound is looking at points in the input space and puts noise spheres around the 'codewords'. When the channel is augmented with feedback, the noise-spheres still determine the minimum distance needed between codewords and hence the sphere-packing bound still holds provided the channel looks like an 'additive noise' channel.

At this point, the steady march of progress in this area took an interesting turn. In 1970, Haroutunian presented ([27], [28]) an upper bound to $E_{fb}(R)$ valid for all DMCs, which is presently called the Haroutunian exponent. The Haroutunian exponent is denoted $E_h(R)$ and defined as

$$E_h(R) \triangleq \min_{V:C(V)\leq R} \max_{P} D(V||W|P). \tag{2.6}$$

He also showed that $E_h(R) = E_{sp}(R)$ if W is output-symmetric[8], recovering the upper bound proved by Dobrushin, but in general $E_h(R) > E_{sp}(R)$ for asymmetric channels. Somewhat disturbed by the gap between $E_h(R)$ and $E_{sp}(R)$ for asymmetric channels, Haroutunian waited five years to submit his result to a journal [3], at which time he wrote

> The result derived here was included in [27] and [28]. The author, however, was in no hurry to have it published in full. The whole point lies in that in the widely adopted hypothesis, the lower bound of error probability for channels with feedback

[7] A DMC W is symmetric at both the input and the output if the rows are permutations of each other and the columns are also permutations of each other. A BSC is symmetric at both the input and the output, but a BEC is not.

[8] A channel is output-symmetric if the output set can be partitioned into subsets such that in each subset the matrix of transition probabilities (from all inputs to this subset of the output) has the property that each row is a permutation of every other row and each column is a permutation of every other column. Output-symmetric channels include BSCs and BECs.

> or with no feedback is the bound for packing of spheres.... The author has tried for some time to improve the proof so that the hypothesis is valid for any discrete channels, although up until now without success. A related question also arises: is it possible to construct block codes with feedback possessing an error probability for unsymmetric channels which would be exponentially lower than the bound for packaging a sphere?

There are two points in the above paraphrasing of Haroutunian that are important to note here. First is the point that it is widely believed that $E_{fb}(R) \leq E_{sp}(R)$ for all DMCs. That is, even with feedback, 'almost everyone' believes that one cannot beat the sphere-packing bound. The reason for this intuition is that the output feedback is available only causally. The major difference between the Haroutunian exponent (2.6) and the sphere-packing exponent (2.3) is that the order of the max and min is interchanged. The sphere-packing bound knows the strategy of the code (P) and chooses a test channel (V) that will cause error with high probability. The Haroutunian bound, on the other hand, fixes a test channel (V) that will cause error with high probability and the code chooses a strategy (P) that makes the error less likely. However, because feedback is in reality, only available causally, the code should not be able to 'predict' what channel will occur and pick its input distribution accordingly.

The second point, which is much more provocative is that perhaps the sphere-packing bound does not hold for asymmetric channels when feedback is available. If this were the case, then it leaves open the possibility that $E_{fb}(R) > E_{sp}(R) = E_r(R)$ for $R > R_{cr}$ if the channel is asymmetric. What makes this idea so provocative (and likely far-fetched) is that the channel W is not handed down from Nature with no possibility of change. Rather, W is a probabilistic model of a *designed* communication system that can involve/require modulation, time and phase synchronization, equalization, etc. For the simple case of BPSK modulation, it is generally *assumed* that the modulation and demodulation operations should be symmetric irrespective of whether the symbol to be sent is 0 or 1. However, it could easily be designed that the energy used to send a 1 be less than the energy used to send a 0 for example. This redesigned BPSK modulation might have the dual advantage of lowering power consumption and increasing reliability. While we don't believe this to be the case, it is an important reason to verify the sphere-packing bound is still an upper bound on the reliability function for all DMCs with feedback[9].

For completeness, we should also mention that the straight-line bound of (2.5) also holds for codes with feedback provided the second error exponent (which is the sphere-packing exponent in codes without feedback) applies to codes with feedback using list decoding. The Haroutunian bound applies to codes with feedback using list decoding, but $E_{ex}(0)$ is no longer a proven upper bound to $E_{fb}(0)$, so the lowest left endpoint of the straight-line bound

[9] Additionally, W may be asymmetric even though it was intended to be symmetric due to imperfections in the physical components in the communication system.

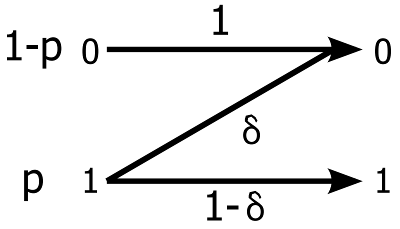

Figure 2.3: A Z-Channel is a binary input, binary output channel with a one-sided crossover probability, denoted here by δ. A 0 is always perfectly received, while a 1 is received as a 0 with probability δ.

(to our knowledge) that has been proved is $E_h(0)$. This straight line is always looser than $E_h(R)$ because $E_h(R)$ is convex.

2.1.2 The simplest family of asymmetric channels: the Z-channel

For now, let us investigate a bit more closely the bound of Haroutunian:

$$E_{fb}(R) \leq E_h(R) = \min_{V:C(V)\leq R} \max_{P} D(V||W|P). \tag{2.7}$$

The first thing to note is that because $D(V||W|P)$ is linear in P, it is maximized by a P that places all its mass on a single $x \in \mathcal{X}$, so

$$E_h(R) = \min_{V:C(V)\leq R} \max_{x} D(V(\cdot|x)||W(\cdot|x)),$$

where $D(V(\cdot|x)||W(\cdot|x))$ denotes the divergence between the distributions on $\mathcal{Y}$: $V(\cdot|x)$ and $W(\cdot|x)$. In order to prove that the Haroutunian exponent upper bounds $E(R)$ for block codes with feedback, one takes a test channel V with capacity lower than the rate of the code. The weak or strong converse can be used to show that the error probability under the 'test' channel V is high. Then, the probability that an error occurs under channel W is governed by $D(V||W|P)$ where P can be thought of as the input distribution *during the error event* for channel V.

Unfortunately, because the code has feedback[10], the input distribution for the error event under channel V need not be the same as the distribution under channel W, and hence the max in (2.7) is taken as a worst-case bound. The conditional divergence is linear in the input distribution however, so the resulting optimizing input distribution places all its mass on one letter. Of course, no good code could do such a thing without dooming itself to error, but the bound of (2.7) essentially assumes that because the code has feedback, it can somehow realize that the channel is behaving like V and use this maximizing letter repeatedly.

[10] When a code has feedback, the input symbols depend on past output symbols, so the input distribution can depend nontrivially on the probabilistic description of the channel.

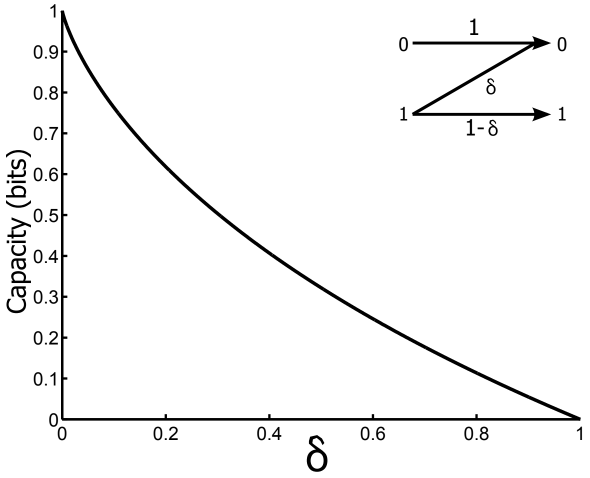

Figure 2.4: Capacity of the Z-Channel for all crossover probabilities $\delta \in [0,1]$. The capacity is a monotone strictly decreasing function of δ, and also convex-∪ in δ.

Let us now see what this all means for the simplest asymmetric channel: the Z-channel, shown in Figure 2.3. A Z-channel with crossover probability $\delta \in [0,1]$ sends a 0 to a 0 with probability 1 and flips a 1 with probability δ. The capacity of the Z-channel as a function of δ, denoted $C_Z(\delta)$, is shown in Figure 2.4.

The most fundamental property of asymmetric channels that separate them from symmetric channels is that the capacity achieving distribution depends quite a bit (and varies) with the channel W. Figure 2.5 shows the capacity achieving distribution (the probability of inputting 1) for a Z-channel with crossover probability δ. The capacity achieving probability ranges from 0.5 at $\delta = 0$ (a noiseless one-bit pipe) to approximately 0.36 as $\delta \to 1$. Contrast this to the BSC, for which the capacity achieving distribution inputs 1 with probability 1/2 for all crossover probabilities. Perhaps even more striking is Figure 2.6, which plots the *sphere-packing* optimizing probability of inputting 1 for the Z-channel with crossover probability $\delta = 0.5$ as a function of the rate. That is, the optimizing P in (2.3) changes for a fixed channel as a function of the rate, from 1 to the capacity achieving $P(1)$. Again, for the BSC or BEC, the sphere-packing optimizing P is uniform on the input for all rates.

Fix a δ and assume that W is a Z-channel with crossover probability δ. Although it is somewhat repetitive, for clarity, we want to interpret the Haroutunian bound for the Z-channel. Now, if a test channel V is not a Z-channel (meaning that $V(1|0) > 0$) and $P(0) > 0$, $D(V||W|P) = \infty$, so the only test channels that are feasible in the optimization of 2.7 are other Z-channels. If V and W are Z-channels with crossover probabilities β and δ respectively,

$$\begin{aligned} D(V||W|P) &= P(1)D(V(\cdot|1)||W(\cdot|1)) \\ &= P(1)D_b(\beta||\delta), \end{aligned}$$

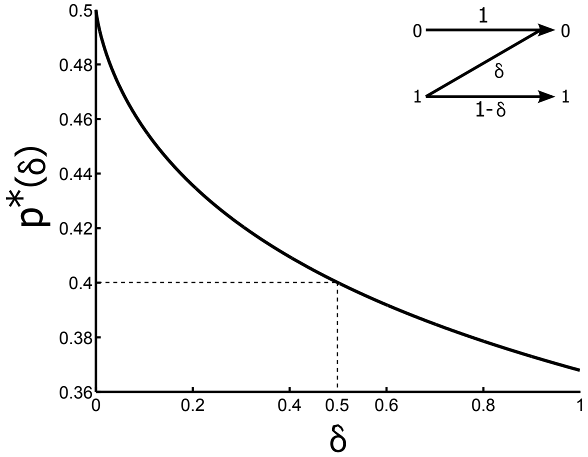

Figure 2.5: The capacity achieving distribution $p^*(\delta)$ for a Z-channel with crossover probability $\delta \in [0, 1)$. Note how $p^*(0) = 1/2$ and for increasing δ, $p^*(\delta)$ is decreasing, requiring a different capacity achieving distribution for each channel in the family, as opposed to symmetric channels like the BSC or BEC.

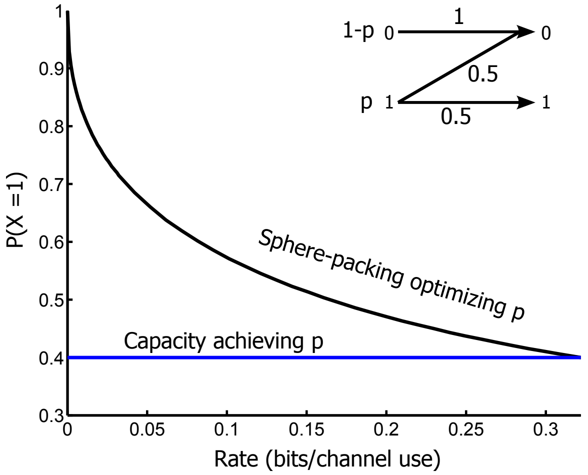

Figure 2.6: The sphere-packing optimizing p as a function of R for a Z-channel with crossover probability $\delta = 0.5$. Note that $p^*_{sp}(R, \delta)$ ranges from 1 down to $p^*(\delta)$ and has a non-zero slope when it reaches $p^*(\delta)$.

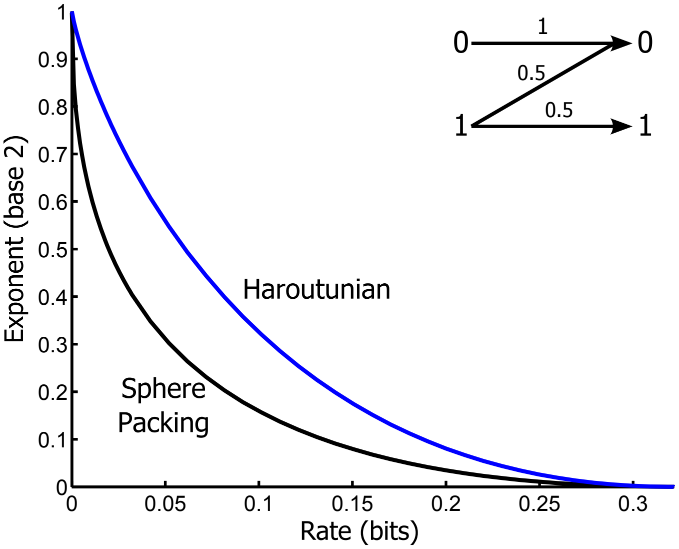

Figure 2.7: The sphere-packing exponent, $E_{sp}(R)$, and the Haroutunian exponent, $E_h(R)$, for a Z-channel with crossover probability 1/2. The capacity of the channel is 0.32 bits per channel use.

where $D_b(\beta||\delta) = \beta \log \frac{\beta}{\delta} + (1-\beta)\log\frac{1-\beta}{1-\delta}$ is the binary divergence between β and δ. The Haroutunian exponent for the Z-channel therefore evaluates to

$$\begin{aligned} E_h(R) &= \min_{\beta: C_Z(\beta) \leq R} \max_P P(1) D_b(\beta||\delta) \\ &= \min_{\beta: C_Z(\beta) \leq R} D_b(\beta||\delta), \end{aligned}$$

where the maximizing $P(1)$ is 1 because divergence is non-negative. Now, presumably $R < C_Z(\delta)$ (otherwise reliable communication is not possible). Therefore, any β for which $C_Z(\beta) \leq R$ will be larger than δ. Because $D_b(\beta||\delta)$ is increasing in β if $\beta \geq \delta$, and $C_Z(\beta)$ has an inverse function, it follows that

$$E_h(R) = D_b\left(C_Z^{-1}(R)||\delta\right).$$

This means that for the Haroutunian bound, the best 'test' channel to bound the error probability with is the Z-channel with rate (slightly less than) R. The probability that W 'behaves like' V is exponential in n with exponent $P(1)D_b(C_Z^{-1}(R)||\delta)$, where $P(1)$ is the probability the code with feedback inputs the symbol 1 *during the error event*. The unsatisfactory part of the Haroutunian bound is that it assumes (because it is difficult to prove otherwise) that the code only inputs 1 during the error event. The reason this assumption is unsatisfactory is the causal nature of feedback. By definition, the input symbol to the channel is decided by the transmitter before the output symbol is revealed

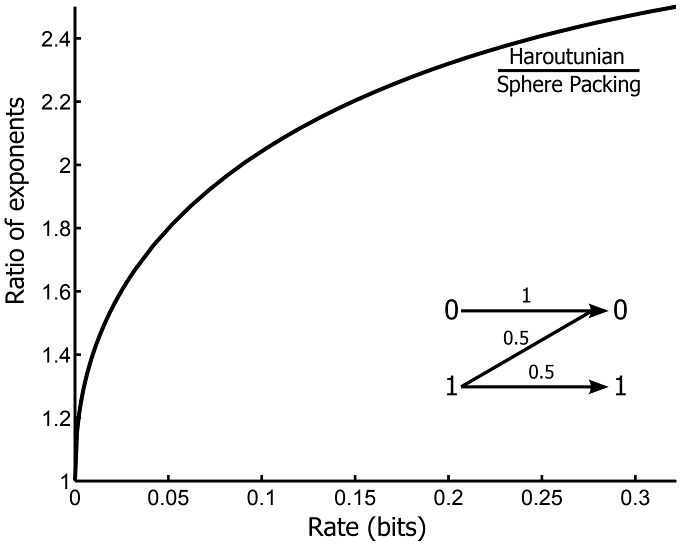

Figure 2.8: The ratio of the Haroutunian exponent to the sphere-packing exponent, $E_h(R)/E_{sp}(R)$, for a Z-channel with crossover probability 1/2. The ratio tends to a value greater than 2 as the rate approaches capacity.

to the receiver and fed back to the transmitter. It stands to reason, therefore, that the transmitter does not know that the error event will definitely occur until at least partway through the block because there are output sequences that lead to error as well as those that do not, which share the same common initial sequence. By memorylessness of the channel and causality of feedback, the transmitter cannot only input the symbol 1 without 'abandoning' those output sequences that do not lead to error.

Figure 2.7 plots the sphere-packing and Haroutunian exponents for a Z-channel with crossover probability 1/2. They are equal at rates 0 and $C(W)$ but there is a sizable gap for all rates in between. Figure 2.8 shows the ratio of the two exponents. Interestingly, the ratio of the exponents is always larger than 2 (for the Z-channel) as the rate approaches capacity.

2.1.3 An alternate view of the rate-reliability tradeoff

The reliability function characterizes the relationship between rate, blocklength and error probability for optimal codes by fixing a rate and blocklength and asking the (approximate) error probability of the optimal code of that rate and blocklength. This error probability turns out to be exponential in blocklength, so one can invert this relationship to get bounds on the required blocklength to achieve a given rate and desired error probability. There is an alternate view of the tradeoff between these fundamental performance parameters. This view looks at the maximum achievable rate for a given blocklength n, and allowable error probability ϵ. For block codes without feedback, let this quantity be defined as

$$R^*(n, \epsilon),$$

while the same quantity for block codes with feedback is denoted $R_f^*(n, \epsilon)$. We know from the channel coding theorem and converse (with feedback), that for a DMC W,

$$\lim_{n\to\infty} R^*(n, \epsilon) = \lim_{n\to\infty} R_f^*(n, \epsilon) = C(W).$$

Polyanskiy, et. al. [29] have built on prior work and shown that

$$R^*(n, \epsilon) = C(W) - \sqrt{\frac{\sigma_W^2}{n}} Q^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right),$$

where σ_W^2 is a channel dependent constant called the *channel dispersion* and Q^{-1} is the inverse of the standard Gaussian Q function. This view of the rate-reliability tradeoff is derived from the central-limit theorem perspective of the limiting distribution for mutual information, as opposed to the large-deviations perspective that gives rise to error exponents. For symmetric channels, they have also shown that [11]

$$R_f^*(n, \epsilon) = C(W) - \sqrt{\frac{\sigma_W^2}{n}} Q^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right) \tag{2.8}$$

even though feedback is available. Therefore, for symmetric channels, feedback does not significantly improve the rate-reliability tradeoff from this perspective either. As noted in [11], this should not be surprising because the sphere-packing exponent is the governing error exponent for block codes with and without feedback. Further, the behavior of the sphere-packing exponent around capacity is given by

$$E_{sp}(R) \simeq \frac{(C(W) - R)^2}{2\sigma_W^2},$$

a fact possibly due to moment generating functions being at the heart of proofs of both the central-limit theorem and large-deviations theorems. Unfortunately, it is not known if for asymmetric channels like the Z-channel, whether the approximation of (2.8) still holds, leaving the door open for an improvement in the rate-reliability tradeoff from this perspective with feedback for asymmetric channels. Again, this may incidentally be due to the fact that the Haroutunian bound has significantly different behavior around capacity than the sphere-packing bound, as seen in Fig. 3.3. The hope is that if one can even prove an upper bound to $E_{fb}(R)$ that has the same behavior around capacity as $E_{sp}(R)$, then feedback can be shown to be useless for improving rate for a fixed reliability over asymmetric channels from this alternative point of view.

2.1.4 Contributions

Unfortunately, we have not been able to show that the sphere-packing bound holds with feedback (i.e., $E_{fb}(R) \leq E_{sp}(R)$) for general DMCs. The main contribution of this chapter is a documentation of the progress made in our understanding of this difficult problem. This progress was made both by succeeding in proving partial results towards sphere-packing in special cases as well as by failing to get to there in general through what looked to be several promising methods. A minor contribution is a description of why the proof in a paper by Sheverdyaev [30] claiming that $E_{fb}(R) \leq E_{sp}(R)$ for general DMCs has serious flaws, which can be read in Appendix A.7.

First, the failures (presented in Section 2.4) are described. Upon seeing Haroutunian's exponent and its change-of-measure approach to error exponents, the first exploratory attempt at the problem might be to try Fano's inequality. At first glance, the restriction on capacity in (2.6) is useful because we know that if the capacity of V is too small, the error probability will be bounded away from 0. We do not require the capacity of V to be too small to reach this conclusion however. By Fano's inequality, we need only that the mutual information across the channel is too low. If P_V is the input distribution for a given code with feedback when[11] the channel is V and $I(P_V, V) \leq R - \epsilon$ (for some small $\epsilon > 0$), the error probability under channel V is bounded away from 0 (even if the capacity of V is larger than the rate). The exponent of the error probability for a given code can then be shown to be upper bounded by

$$\min_{V: I(P_V, V) \leq R-\epsilon} D(V||W|P_{I,V}), \tag{2.9}$$

where $P_{I,V}$ is the distribution of the input restricted to the error causing output sequences for each message. Through our interactions with Baris Nakiboglu at MIT, we knew that (somewhat surprisingly)

$$\min_{V: I(P_V, V) \leq R-\epsilon} D(V||W|P_V) \leq E_{sp}(R - \epsilon), \tag{2.10}$$

but such a conclusion is difficult to reach in (2.9) because nothing is known about $P_{I,V}$. For example, without more information, $P_{I,V}$ could place all its mass on the x that maximizes $D(V(\cdot|x)||W(\cdot|x))$ for each V. Further, without more information, the only V for which we definitively know that $I(P_V, V) \leq R - \epsilon$ are those V for which $C(V) \leq R - \epsilon$. So without information to refute these last two points, the exponent of (2.9) reduces to

$$\begin{aligned} \min_{V: I(P_V, V) \leq R-\epsilon} D(V||W|P_{I,V}) &\leq \min_{V: C(V) \leq R-\epsilon} \max_x D(V(\cdot|x)||W(\cdot|x)) \\ &= E_h(R - \epsilon), \end{aligned}$$

[11] If the code has feedback, the input distribution depends on the probability measure on the output sequences, which in turn depends on the channel.

which is of course the Haroutunian exponent.

At this point, one might think that in order to make use of (2.10), we should try to show that $P_{I,V} \simeq P_V$. One way of showing that $P_{I,V} \simeq P_V$ is to show that under channel V, the error probability is very close to 1. If $C(V) \leq R - \epsilon$, this conclusion is called the strong converse, but we want to show that the error probability is close to 1 even if $I(P_V, V) \leq R - \epsilon$. This conclusion actually turns out to be untrue, and we give a basic counterexample showing that this kind of 'refined' strong converse does not hold.

Following this development, we stepped back and thought about what makes codes with feedback different from codes without feedback. The only reason that codes with feedback could conceivably beat the sphere-packing bound is that they might learn that the channel is behaving atypically and change their input distribution accordingly. Causality of feedback and memorylessness of the channel should imply that the Haroutunian bound is unattainable, but perhaps doing better than sphere-packing is not out of the question. The last failure attempted to combat this reasoning by allowing the test channel V to also adapt according to the strategy of the code. Essentially, we let the test channel have memory, i.e., at each time, the test channel depends on the output sequence received so far. For each time instant in the block and received sequence at that time, the received channel is chosen optimally (i.e., the sphere-packing test channel) depending on the posterior input distribution. Due to some technical difficulties we had with convergence of random variables, we were not able to show that this choice of test channel yields the sphere-packing bound, but we conjecture that it does.

Finally, we worked backwards. We assumed that the sphere-packing bound holds with feedback and attempted to reduce this fact to a simple condition on codes with feedback and their encoding trees. This approach does not appear to be very illuminating, but provides an alternative statement to aim to prove, and is described in Section 2.6.

In terms of partial results, our first one came out of bridging from codes without feedback to codes with feedback. The natural candidate for a code that uses feedback but does not change its "strategy" during the block is a code composed of what we term 'fixed-type encoding trees'. In a fixed-type encoding tree (an encoding tree is the encoding function for one message in a code), the type (or composition or empirical distribution) of the input sequence is the same for every received sequence. This restriction does not preclude the code from using the feedback in a non-trivial way. We show that in this special case, the code appears to not be using feedback at all. What is meant by this is that, for any conditional type that relates input sequences to output sequences, the number of output sequences with that conditional type is exactly the same as when the code does not have feedback (but instead uses a codeword of the same input type). This can be used to prove that the sphere-packing bound holds by the usual combinatorial approach. What is interesting about this result is not that the sphere-packing bound continues to hold (which can be shown by change-of-measure and Fano's inequality), it is that the sizes of the conditional shells are *exactly* the same as they would be if the code did not use feedback. This insight is described in Section 2.5.

The second partial result moves in the direction of making the feedback model more realistic. When proving upper bounds to $E_{fb}(R)$, we assume that the feedback is noiseless and delay-free, as this is the best that one could hope for in practice. In reality, however, for the same reasons that noise and delay impair the forward channel W, they also impair the output feedback. Our partial result in Section 2.7 shows that if the output feedback is delayed by T symbols, the error exponent is upper bounded by

$$E_{fb,T}(R) \leq E_{sp}\left(R - O\left(\frac{\log T}{T}\right)\right) + O\left(\frac{\log T}{T}\right). \tag{2.11}$$

In the limit as T tends to infinity, we see that the sphere-packing bound must hold. This result shows that feedback information about the very far past is useless for improving the error exponent. It also complements recent results of Baris Nakiboglu and Giacomo Como (which are commented on in Section 2.9), who showed that the sphere-packing bound holds for codes with feedback if the encoders are restricted to hold information about the very near past only.

The result of (2.11) naturally leads into the question of what the Haroutunian exponent behaves like for parallel channels. The reason is that, instead of thinking of using the channel W one symbol at time with a delay of T symbols in the feedback path, we can think of using W T symbols at a time with a delay of 1 supersymbol (a block of T symbols from $\mathcal{Y}$), with each use of W being independent of the others. If we let $W^{(T)}$ denote this T-wise parallel channel, we show in Section 2.8 that

$$E_h(TR; W^{(T)}) \leq TE_{sp}\left(R - \frac{|\mathcal{X}|}{T}\log(T+1); W\right),$$

where the left hand side denotes the Haroutunian exponent for the parallel channel at rate TR and the right hand side denotes T times the sphere-packing exponent for W evaluated at rate $R - O((\log T)/T)$. Thus, in the limit as T gets large, the normalized Haroutunian exponent for the parallel channel approaches the sphere-packing exponent of W. This is a rather surprising discovery when compared to the analogous statements about capacity and sphere-packing exponents for the parallel channel. Namely, the capacity of $W^{(T)}$ is $TC(W)$ and the sphere-packing exponent of $W^{(T)}$ at rate TR is $TE_{sp}(R; W)$. As described in Section 2.8, the parallel channel $W^{(T)}$ starts to look more and more symmetric as T gets large. While this development is not of further use for block coding with feedback, it is used in Chapter 3 to show that the Haroutunian exponent can be tightened to sphere-packing for fixed delay codes.

To conclude the chapter, we discuss the current state of affairs for the error exponent with feedback for general DMCs. As a final introductory remark, we note that much of the work in this chapter came out of discussions with Baris Nakiboglu and Giacomo Como at MIT on the paper of Sheverdyaev [30]. We use two results of theirs in this chapter and cite them accordingly. The result on delayed feedback appears in [31].

2.2 Definitions and notation

For those familiar with channel coding in information theory, the notation used in this section is introduced in Tables 2.1 and 2.2. For a further description of notation, please see Appendix A.1.

2.3 The sphere-packing and Haroutunian bounds

To understand the motivation for our failed approaches to proving that $E_{fb}(R) \le E_{sp}(R)$, it is important to know how the sphere-packing bound is proved for codes without feedback and how the Haroutunian bound is proved for codes with feedback. We should mention here that much of the work in this thesis uses the method of types [32] to prove and get intuition about discrete, memoryless systems. Knowledge of the method of types will make it much easier to understand the material.

Theorem 1 (Sphere-packing for block codes without feedback)**.** *Fix a $\delta > 0$. There is a finite $n_{SP}(W,\delta)$ such that for $n \ge n_{SP}(W,\delta)$, any block code of length n and rate R without feedback has*

$$-\frac{1}{n}\log P_e(W) \le E_{sp}(R-\delta)+\delta.$$

Proof: There are two conceptually different proofs that we will consider. The first is a combinatorial proof by the method of types and the second is a proof motivated by the strong converse and involves the use of a 'change-of-measure'. In both proofs, however, a 'pre-processing' step must be performed. Given any code without feedback of blocklength n, there is a subcode whose codewords are all of the same type and the rate of this subcode is at least $R - \frac{|\mathcal{X}|}{n}\log(n+1)$. So fix this subcode and the common type of its codewords P. We give here an outline of the two proofs.

- **Method of types** For a full proof, see Appendix A.2. Given P, we choose the $V \in \mathcal{V}_n(P)$ that is 'closest' to the sphere-packing optimizer, that is

  $$V \in \arg\min_{U\in\mathcal{V}_n(P):I(P,V)\le R-2\delta} D(U||W|P).$$

  Consider the V-shells around each $x^n(m), m \in \mathcal{M}$. These V-shells have cardinality at least $\exp(nH(V|P))/(n+1)^{|\mathcal{X}||\mathcal{Y}|}$, and for every $m \in \mathcal{M}$, $T_V(x^n(m)) \subset T_{PV}$, with

  $$|T_{PV}| \le \exp(nH(PV)) \ll \exp\left(n\left(R - \frac{|\mathcal{X}|}{n}\log(n+1) + H(V|P)\right)\right)$$

  because we chose V so that $I(P,V) \le R-2\delta$ and we can assume for large enough n, $\frac{|\mathcal{X}|}{n}\log(n+1) \le \delta$. Therefore, there must be significant overlap in the V-shells around

Notation	Description
log, exp	Logarithm and exponential to the base e unless otherwise specified
$\mathcal{X}$	Finite channel input alphabet
$\mathcal{Y}$	Finite channel output alphabet
W	'True' channel
V	'Test' channel
κ_V	$\max_{x,y:V(y\|x)>0} -\log V(y\|x)$
τ_V	$\min_{x,y} V(y\|x)$
$\mathcal{W}$	Set of channel transition matrices with input/output alphabets $\mathcal{X}$ and $\mathcal{Y}$
$\mathcal{P}$	Set of distributions on $\mathcal{X}$
$\mathcal{Q}$	Set of distributions on $\mathcal{Y}$
$\mathcal{P}_n$	Set of types of length n for alphabet $\mathcal{X}$
$\mathcal{V}_n(P)$	Set of conditional types of length n for input type P
$\mathcal{Q}_n$	Set of types of length n for alphabet $\mathcal{Y}$
x^n	Vector notation for $x^n = (x_1, \ldots, x_n)$ (non-random)
X^n	Vector notation for $X^n = (X_1, \ldots, X_n)$ (capital letters for random variables)
T_P	Type class of vectors with type P for some $P \in \mathcal{P}_n$
$T_V(x^n)$	Conditional V-shell of vector x^n
$H(P)$	Entropy of distribution P
$H(V\|P)$	Conditional entropy of output of channel V when input has distribution P
PV	Distribution on $\mathcal{Y}$ when input distribution is P and channel is V
(P, V)	Joint distribution in $\mathcal{X} \times \mathcal{Y}$ when input distribution is P and channel is V
$I(P, V)$	Mutual information between input and output of channel V when input has distribution P
$I(X; Y)$	Mutual information between random variables X and Y
$C(V)$	Capacity of channel V
$D(P\|\|\widetilde{P})$	Divergence between distributions P and $\widetilde{P}$
$D(V\|\|W\|P)$	Conditional divergence between V and W when input distribution is P
$\|\|P - P'\|\|_1$	$\mathcal{L}_1$ distance between P and P'
$h_b(\delta)$	Binary entropy $-\delta \log \delta - (1-\delta)\log(1-\delta)$
$D_b(\delta\|\|\beta)$	Binary divergence between δ and β
$E_h(R)$	Haroutunian exponent at rate R
$E_{sp}(R)$	Sphere-packing exponent at rate R

Table 2.1: Definitions for distributions, channels, types and functions of distributions.

Notation	Description
n	Blocklength
$\mathcal{M}$	Message set
$\lvert\mathcal{M}\rvert$	Message set size
R	Rate of code, $\frac{1}{n}\log\lvert\mathcal{M}\rvert$
$\phi(m)$	Codeword for message m (for block code without feedback)
$x^n(m)$	Codeword for message m (for block code without feedback)
$x_i(m)$	i-th input letter for message m (for block code without feedback)
$x^n(m, y^n)$	Codeword for message m when received output is y^n (for block code with feedback)
$x_i(m, y^{i-1})$	i-th input letter for message m when received output is y^{i-1} (for block code with feedback)
$P(m, y^n)$	Type of $x^n(m, y^n)$
$V(m, y^n)$	Conditional type that y^n is in if input is $x^n(m, y^n)$
$B(m, P, U)$	y^n for which $P(m, y^n) = P$ and $V(m, y^n) = U$
$\mathcal{D}_m$	Decoding region for message m (all block codes)
$P_e(V)$	Average error probability under channel V
$P_c(V)$	Average correct reception probability under channel V
P_V	Average input distribution under channel V (block code with feedback)
$P_{I,V}$	Average input distribution under error event for channel V

Table 2.2: Block coding definitions.

each of the messages. The overlap implies that when those sequences lying in overlapping V-shells are output, an error occurs with high probability because each sequence can only be decoded to one message. For each message, the probability of the output being in the V-shell around its codeword is approximately $\exp(-nD(V||W|P))$. The exponent of the error probability is thus approximately

$$\min_{V \in \mathcal{V}_n(P): I(P,V) \leq R-2\delta} D(V||W|P).$$

The set of conditional types for P, $\mathcal{V}_n(P)$, gets 'close' to all of $\mathcal{W}$ as n gets large, so for large n,

$$\begin{aligned} \min_{V \in \mathcal{V}_n(P): I(P,V) \leq R-2\delta} D(V||W|P) &\simeq \min_{V \in \mathcal{W}: I(P,V) \leq R-2\delta} D(V||W|P) \\ &\triangleq E_{sp}(R-2\delta, P). \end{aligned}$$

Because $E_{sp}(R) = \max_P E_{sp}(R, P)$, the proof is complete.

- **Change of measure** For a full proof, see Appendix A.3. Given $P \in \mathcal{P}_n$, directly choose the sphere-packing optimizing V. That is, let

  $$V \in \arg \min_{U \in \mathcal{W}: I(P,U) \leq R-2\delta} D(V||W|P).$$

  Because all the codewords are of type P, and $I(P,V) \leq R - 2\delta$, one can prove a strong converse for this subcode[12] and show that if n is large, $P_e(V) \simeq 1$. We want to show that $P_e(W)$ is lower bounded by the sphere-packing bound, so we perform a *change of measure* from V to W. The change of measure says that, approximately up to subexponential terms,

  $$\begin{aligned} \frac{P_e(W)}{P_e(V)} &\geq \exp\left(-nD(V||W|P)\right) \\ &= \exp(-nE_{sp}(R-2\delta, P)) \\ &\geq \exp(-nE_{sp}(R-2\delta)). \end{aligned}$$

  Because $P_e(V) \simeq 1$, we get that the sphere-packing bound holds.

Now, moving to codes with feedback, the type of the input over the blocklength depends on both the message *and the output sequence.* The first idea for extending the sphere-packing proofs to codes with feedback might be to group message and output sequence pairs (m, y^n) according to the type of the input sequence $P(m, y^n)$, and concentrate on the P with the largest representation. Unfortunately, there is no guarantee that if we look at

[12] The usual strong converse says that if $C(V) < R$, $P_e(V)$ tends to 1 as $n \to \infty$.

a sphere-packing optimizing V and its conditional shell $B(m, P, V)$ around each message, $|B(m, P, V)|$ is large enough to cause errors with high probability. This is the impediment that prevented the original proof of the sphere-packing bound from going through for codes with feedback, as noted by Shannon, Gallager and Berlekamp [18]. Therefore, Haroutunian looks at channels V that he knows will cause errors: those that have a capacity lower than the rate.

Theorem 2 (Haroutunian bound for block codes with feedback)**.** *For any $\delta > 0$, there exists a finite $n_h(W, R, \delta)$ such that any fixed-length code with feedback of rate R and length $n \geq n_h(W, R, \delta)$ has*

$$-\frac{1}{n} \log P_e(W) \leq E_h(R - \delta) + \delta.$$

Proof: For a full proof, see Appendix A.4. This proof also proceeds by a strong converse and then a change of measure. Fix the Haroutunian optimizing V,

$$V \in \arg \min_{U \in \mathcal{W}: C(U) \leq R-\delta} \left\{ \max_{P \in \mathcal{P}} D(U||W|P) \right\}.$$

Because $C(V) \leq R - \delta$, the strong converse with feedback says that $P_e(V) \simeq 1$. A change of measure shows that, up to subexponential terms,

$$\frac{P_e(W)}{P_e(V)} \geq \exp(-nD(V||W|P_{I,V})),$$

where $P_{I,V}$ is the distribution of the input ‘during the error event’. This distribution could be anything, so to get a bound for all codes, we take the maximum over $P_{I,V}$ to get Haroutunian’s exponent.

2.4 Failed Approaches to proving the sphere-packing bound for codes with feedback

After being interested in the reliability function with feedback problem for some time, we were made aware by Baris Nakiboglu at MIT of a paper by Sheverdyaev [30] that claimed to prove that $E_{fb}(R) \leq E_{sp}(R)$ for all DMCs that have no 0’s. After examining the paper, we found two major flaws (which are described in more detail in Appendix A.7). One flaw was a claim (without proof and stated to be obvious) that

$$\inf_{V: I(P_V, V) \leq R} D(V||W|P_V) \leq \max_{P} \min_{V: I(P,V) \leq R} D(V||W|P) \tag{2.12}$$

$$= E_{sp}(R), \tag{2.13}$$

where P_V is the average input distribution of a given code when the channel is V. It is not at all obvious how to show (2.12) is true (or whether it is true at all). The problem is the lack of information available about the distribution P_V as a function of V. In particular, what if P_V is the capacity achieving distribution for V with $C(V) > R$ and is a point mass on the Haroutunian exponent maximizing x for V with $C(V) \leq R$? The left hand side of (2.12) would then evaluate to $E_h(R)$. Of course, such a P_V is not possible because we know that the input distribution must be continuous as a function of the channel. Using this information, Nakiboglu [33] was able to prove the following lemma.

Lemma 1. *Let P_V be any continuous function from $\mathcal{W}$ to $\mathcal{P}$, and let for any $V \in \mathcal{W}$, $\kappa_V = \max_{x,y:V(y|x)>0} \log(1/V(y|x))$. Then, for all $R > 0, \epsilon \geq 0$,*

$$\inf_{V:I(P_V,V)\leq R} D(V||W|P_V) + \epsilon \max\{\kappa_V, \kappa_W\} \leq E_{sp}(R) + \epsilon(\kappa_W + \log|\mathcal{Y}|).$$

The proof of the lemma can be found in Appendix A.5 as it is not available in the literature and Nakiboglu currently has no plans to publish it. The proof involves considering a special family of parametrized test channels and using the intermediate value theorem. The fact that $\max\{\kappa_W, \kappa_V\} \leq \kappa_W + \log|\mathcal{Y}|$ for this family of channels is useful in bounding for change of measure.

Lemma 1 closes one gap in the proof of Sheverdyaev, but there is a second flaw in the paper involving Taylor expansions (as detailed in Appendix A.7). This flaw can be exposed by showing that a claim derived by Sheverdyaev (after taking Taylor expansions and using 'uniformly bounded' constants for the error terms in the expansion) can be refuted by a counterexample. It is our feeling that one philosophical reason why the Sheverdyaev proof does not work is because it does not use causality of feedback in a serious way. We believe that any proof of sphere-packing with feedback must crucially use causality *and* memorylessness of the channel.

Imre Csiszar informed us at ISIT 2010 that it was known within the community of researchers interested in this problem that there are flaws in the proof of [30]. He also pointed us to a manuscript of Augustin [34] that was never published in a peer-reviewed format which claims[13] to show that $E_{fb}(R) \leq E_{sp}(R)$ for all channels. We cannot make a definitive statement about the correctness of Augustin's proof because of the quality of the translation (from German) as well as lack of detail in the proof. See Appendix A.8 for further comments.

We now give three of our attempts to prove $E_{fb}(R) \leq E_{sp}(R)$ for general DMCs, starting with the simplest idea of using Fano's inequality to induce errors.

[13] Csiszar noted that he had never reviewed the proof due to the length of the manuscript and did not know if it was correct.

2.4.1 A first attempt via Fano's inequality

For any $V \in \mathcal{W}$, and a given block code with feedback, let the **input distribution** under channel V be

$$\forall\, x \in \mathcal{X},\ P_V(x) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{y^n} \mathbb{P}_V(Y^n = y^n | M = m) \frac{1}{n} \sum_{i=1}^{n} 1(x_i(m, y^{i-1}) = x) \tag{2.14}$$

and let the **incorrect input distribution** under channel V be

$$\forall\, x \in \mathcal{X},\ P_{I,V}(x) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{y^n \notin \mathcal{D}_m} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)} \frac{1}{n} \sum_{i=1}^{n} 1(x_i(m, y^{i-1}) = x). \tag{2.15}$$

To recap, we have just defined two input distributions of the code with feedback as a function of the channel V it faces. The reason new definitions are required is because the input distribution of a code with feedback critically depends on the channel behavior. The input distribution of the code can be used to select a channel that we know (by Fano's inequality) will induce errors because the mutual information of the input distribution (under V) over channel V is too low. When, we change the measure from V to W, we will see that the input distribution that shows up in the divergence is the incorrect input distribution.

Lemma 2 (Not quite sphere-packing via Fano's inequality)**.** *Fix an $R > 0$ and $\delta \in (1/n, R)$. For any block code with feedback of rate R and length n,*

$$-\frac{1}{n} \log P_e(W) \leq \inf_{V : I(P_V, V) \leq R - \delta} D(V||W|P_{I,V}) + \frac{2 \max\{\kappa_V, \kappa_W\}}{\frac{1}{R}\left(\delta - \frac{1}{n}\right)} \beta(n, |\mathcal{X}|, |\mathcal{Y}|) + \frac{1}{n} \log\left(\frac{1}{R}\left(\delta - \frac{1}{n}\right)\right),$$

where

$$\beta(n, |\mathcal{X}|, |\mathcal{Y}|) = \inf_{\epsilon > 0} \epsilon + (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n \frac{\epsilon^2}{2}\right) = O\left(\sqrt{\frac{|\mathcal{X}||\mathcal{Y}| \log n}{n}}\right).$$

This lemma is proved by using Fano's inequality for channel V and then performing a change of measure from V to W. The full proof is in Appendix A.6.1. The important term above (that does not decay to 0 as $n \to \infty$) is the divergence term.

2.4.2 A refined strong converse

It looks as though the approach given by Fano's inequality will not yield the sphere-packing bound because we do not know enough about $P_{I,V}$ as a function of V. In fact, the best we can say without further information is that $P_{I,V}$ might be a point mass for a given V (it cannot simultaneously be a point mass that changes for each V because it must at least be continuous). Therefore, we cannot apply Lemma 1 to get a relation between the exponent in Lemma 2 and the sphere-packing exponent. Seeing this, we posit a property of codes with feedback that, if it held, would ensure that the sphere-packing bound would hold. Unfortunately, we will show that the condition we posited does not hold and altering it to one that likely does hold would not allow us to prove that the sphere-packing holds.

Definition 1. *We say the* **refined strong converse** *holds for block codes with feedback if $\forall \delta > 0, V \in \mathcal{W}, n \geq 1$, there exists a function $\gamma_{RSC}(n, \delta, R, \kappa_V)$ such that for any length n block code with feedback of rate R, if $I(P_V, V) \leq R - \delta$, then*

$$P_c(V) \leq \gamma_{RSC}(n, \delta, R, \kappa_V)$$

and if δ, R, V are fixed,

$$\lim_{n \to \infty} \gamma_{RSC}(n, \delta, R, \kappa_V) \to 0.$$

This is called the r*efined strong converse because the standard strong converse requires that the probability of error goes to 1 if $C(V) < R$, while this refined condition requires only that $I(P_V, V) < R$, where P_V is the average input distribution under channel V. Also, we require that the dependence on V be through κ_V (and $|\mathcal{X}|$ and $|\mathcal{Y}|$, which are suppressed in the notation), with a smaller κ_V leading to a smaller upper bound on the error probability.*

Lemma 3. *If the refined strong converse holds, then for any $R > 0$, $\delta \in (0, R)$, there is a finite $n_{RSC}(W, R, \delta)$ such that for any block code with feedback of rate R and length $n \geq n_{RSC}(W, R, \delta)$,*

$$-\frac{1}{n} \log P_e(W) \leq E_{sp}(R - \delta) + \delta.$$

Hence, if the refined strong converse holds, so does the sphere-packing bound.

For the proof of this lemma, see Appendix A.6.2. Essentially, it uses the refined strong converse assumption to show that if $I(P_V, V) \leq R - \delta$, $||P_{I,V} - P_V||_1 \simeq 0$ and so

$$\inf_{V: I(P_V, V) \leq R - \delta} D(V||W|P_{I,V}) \simeq \inf_{V: I(P_V, V) \leq R - \delta} D(V||W|P_V) \leq E_{sp}(R),$$

where the inequality is from Lemma 1.

While Lemma 3 is correct, the assumption that the refined strong converse holds is not. In fact, the refined strong converse does not even hold for codes without feedback.

Example 1. *Consider a binary code without feedback with $1/8$ its codewords in the type $(1/2, 1/2)$ and the rest being all zero codewords. So $1/8$ of the code can be considered a good code, while $7/8$ is a bad code that communicates nothing. The average input distribution of the code is $P_V = (15/16, 1/16)$ for all $V \in \mathcal{W}$ (where $\mathcal{W}$ is the set of binary input, binary output channels). Suppose the rate of this code is $R = 1 - h_b(1/3) \simeq 0.082$ bits per symbol.*

Now let V be a binary symmetric channel with crossover probability $1/4$, so $C(V) = 1 - h_b(1/4) \simeq 0.19$ bits per symbol. Now, $I(P_V, V) = 0.046 < R = 0.082$. However, the subcode composed of type $(1/2, 1/2)$ codewords can be a very good code and its rate is about 0.082 for large n. The channel V has capacity $0.19 > 0.082$, so this good subcode can have a very low error probability (say about 0). Hence, the probability of error for large n is upper bounded by approximately $7/8$ and does not approach closer to 1 in the limit. Therefore the refined strong converse does not hold.

With this example in mind, what is likely to be true is a weaker refined strong converse, a condition that we define below.

Definition 2. *We say the* **refined strong converse with message selection** *holds if for $n \geq n_{MS}(R, \delta, |\mathcal{X}|, |\mathcal{Y}|), \delta > 0, R > 0, V \in \mathcal{W}$, there is a $\gamma_{MS}(n, \delta, R, \kappa_V)$ such that for all length n, rate R codes with feedback, if $I(P_V, V) \leq R - \delta$, there is a subcode (a subset of messages) of rate at least $R - \frac{|\mathcal{X}||\mathcal{Y}|}{n} \log(n+1)$ with*

$$P_c(V) \leq \gamma_{MS}(n, \delta, R, \kappa_V)$$

and $\gamma_{MS}(n, \delta, R, \kappa_V) \to 0$ as $n \to \infty$ with the other parameters fixed.

The refined strong converse with message selection certainly holds for codes without feedback. In fact, it is how the sphere-packing bound for codes without feedback is proved. Additionally, intuition suggests that it is also true for codes with feedback. Unfortunately, if we assume that the refined strong converse with message selection holds for codes with feedback, it does not immediately follow that the sphere-packing bound holds as in Lemma 3. The reason is that Lemma 1 requires that the input distribution P_V be continuously varying with V. When the strong converse only holds for a subset of messages, it is possible that the subset can depend on the test channel V. Therefore P_V will vary discontinuously with V if that subset changes (as the subset must change discontinuously with V, being a discrete object). It is unclear if an additional argument on top of this can be made to somehow 'smooth out' the discontinuity with V by adding in terms from different messages according to how far the messages are from lying in the 'bad' subcode guaranteed by Definition 2.

2.4.3 A test channel with memory

As mentioned earlier, another approach we took was to step back and give ourselves a wider choice of test channels to force errors for a code with feedback. We will do so by allowing the test channel to depend on the received sequence (essentially using a test channel with memory). For a given block code with feedback, let $\widetilde{V}$ denote a probability measure on $\mathcal{M} \times \mathcal{X}^n \times \mathcal{Y}^n$ with

$$\mathbb{P}_{\widetilde{V}}(m, x^n, y^n) = \frac{1}{|\mathcal{M}|} \prod_{i=1}^{n} 1(x_i(m, y^{i-1}) = x_i) \widetilde{V}_{y^{i-1}}(y_i|x_i),$$

where $\widetilde{V}_{y^i} \in \mathcal{W}$ for $i = 0, \ldots, n-1$, $y^i \in \mathcal{Y}^i$. Let $P_{y^i} \in \mathcal{P}$ denote the input distribution under measure $\widetilde{V}$ after the channel output y^i has been received. That is,

$$\begin{aligned} \forall\ x \in \mathcal{X},\ P_{y^i}(x) &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{\mathbb{P}_{\widetilde{V}}(Y^i = y^i | M = m)}{\mathbb{P}_{\widetilde{V}}(Y^i = y^i)} 1(x_{i+1}(m, y^i) = x) \\ &= \sum_{m \in \mathcal{M}} \mathbb{P}_{\widetilde{V}}(M = m | Y^i = y^i) 1(x_{i+1}(m, y^i) = x), \end{aligned} \tag{2.16}$$

where

$$\mathbb{P}_{\widetilde{V}}(Y^i = y^i) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_{\widetilde{V}}(Y^i = y^i | M = m).$$

Note that the P_{y^i} are defined recursively. That is, first define for $y^0 = \emptyset$,

$$P_{\emptyset}(x) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} 1(x_i(m) = x).$$

Then, set a $\widetilde{V}_{\emptyset} \in \mathcal{W}$. The choice of $\mathcal{V}_{\emptyset}$ now induces the measure $\mathbb{P}_{\widetilde{V}}(Y^1 = y^1 | M = m)$ for each m through the code, which in turn induces the input distribution P_{y^1} for each $y^1 \in \mathcal{Y}^1$ through (2.16). Now, for each $y^1 \in \mathcal{Y}^1$, set a $\widetilde{V}_{y^1} \in \mathcal{W}$. Thereafter, an input distribution P_{y^2} is induced for each $y^2 \in \mathcal{Y}^2$, and so on. Hence, if just choose $\widetilde{V}_{y^i}$ for each y^i, $i = 0, \ldots, n-1$, the measure $\widetilde{V}$ is well defined. The channel $\widetilde{V}_{y^i}$ depends on the received sequence y^i and hence the test channel $\widetilde{V}$ is a channel with memory. Additionally, like the test channel used in the proof of the sphere-packing bound, it should depend on the code to give a good lower bound to the probability of error. This attempt at least tries to use causality of feedback in a nontrivial way in order to prove a lower bound for the error probability. For each time and with each received sequence, we are attempting to see what happens if the best (in terms of divergence), bad (in terms of mutual information) channel shows up.

The intuition for attempting this approach is that a code with feedback can potentially change its coding strategy according to how 'good' or 'bad' the channel is. Indeed, this is exactly the reason why Haroutunian's bound contains the maximization over input distributions inside the minimization over channels. However, we know that even though the code can adaptively change its input, it cannot 'predict' whether the channel will be good or bad.

At best, it can only react to whether the channel has been good or bad up to the present time because of the causal nature of feedback. Therefore, allowing the channel to depend on the past received symbols might allow us to tailor the error inducing channel to the part of the encoding tree we are in. With that in mind, we will make the following choice for $\widetilde{V}_{y^i}$:

$$\widetilde{V}_{y^i} \in \arg\min_{V \in \mathcal{W}} \left\{ D\left(V || W | P_{y^i}\right) : I\left(P_{y^i}, V\right) \leq R - 2\delta \right\}, \forall\ i = 0, \ldots, n-1,\ y^i \in \mathcal{Y}^i \tag{2.17}$$

for some small $\delta > 0$.

At this point, we will do two things with this test channel with memory. First, we will show that if two statements about the code used over this test channel with memory are true, then the sphere-packing bound holds. Second, we will consider the technical obstacles in showing that those two technical statements are true for the choice of $\widetilde{V}$ in (2.17).

First, we will show that the sphere-packing bound must hold for channel W provided two statements can be made with high probability. This approach is adapted from the information-spectrum literature [35].

Lemma 4 (Information Spectrum Converse). *Let $\widetilde{V}$ be a measure*[14] *on $\mathcal{M} \times \mathcal{Y}^n$. Fix a $\delta > 0$. Let*[15]

$$A \triangleq \left\{ (m, y^n) : \frac{1}{n} \log \frac{\mathbb{P}_{\widetilde{V}}(Y^n = y^n | M = m)}{\mathbb{P}_{\widetilde{V}}(Y^n = y^n)} \leq R - \delta \right\} \tag{2.18}$$

$$B \triangleq \left\{ (m, y^n) : \frac{1}{n} \log \frac{\mathbb{P}_{\widetilde{V}}(Y^n = y^n | M = m)}{\mathbb{P}_{W}(Y^n = y^n | M = m)} \leq E_{sp}(R - 2\delta) + \delta \right\} \tag{2.19}$$

$$E \triangleq \left\{ (m, y^n) : y^n \notin \mathcal{D}_m \right\}.$$

Then,

$$P_e(W) \geq \exp\left(-n\left[E_{sp}(R - 2\delta) + \delta\right]\right) \left[1 - \exp(-n\delta) - \mathbb{P}_{\widetilde{V}}(B^c) - \mathbb{P}_{\widetilde{V}}(A^c)\right].$$

Hence,

$$P_e(W) \geq \exp(-n[E_{sp}(R - 2\delta) + 2\delta])$$

for large n provided

$$\mathbb{P}_{\widetilde{V}}\left(\frac{1}{n} \log \frac{\mathbb{P}_{\widetilde{V}}(Y^n | M)}{\mathbb{P}_{\widetilde{V}}(Y^n)} \leq R - \delta \right) \to 1,\ n \to \infty \tag{2.20}$$

$$\mathbb{P}_{\widetilde{V}}\left(\frac{1}{n} \log \frac{\mathbb{P}_{\widetilde{V}}(Y^n | M)}{\mathbb{P}_{W}(Y^n | M)} \leq E_{sp}(R - 2\delta) + \delta \right) \to 1,\ n \to \infty. \tag{2.21}$$

[14] Of course, this measure is induced from one on $\mathcal{M} \times \mathcal{X}^n \times \mathcal{Y}^n$.

[15] Recall that capital letters are used to denote random variables while lower case vectors denote nonrandom realizations.

A proof of the lemma can be found in Appendix A.6.3. Our task now would be to show that with the choice of $\widetilde{V}$ in (2.17), statements (2.20) and (2.21) hold. Unfortunately, this task is harder than in the case when $\widetilde{V}$ does not have memory and the code does not have feedback. First, note that

$$
\begin{aligned}
Z &\triangleq \frac{1}{n}\log\frac{\mathbb{P}_{\widetilde{V}}(Y^n|M)}{\mathbb{P}_{\widetilde{V}}(Y^n)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathbb{P}_{\widetilde{V}}(Y_i|M,Y^{i-1})}{\mathbb{P}_{\widetilde{V}}(Y_i|Y^{i-1})} \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\frac{\widetilde{V}_{Y^{i-1}}(Y_i|X_i(M,Y^{i-1}))}{\sum_x P_{Y^{i-1}}(x)\widetilde{V}_{Y^{i-1}}(Y_i|x)} \\
&= \frac{1}{n}\sum_{i=1}^{n} Z_i \\
Z_i &\triangleq \log\frac{\mathbb{P}_{\widetilde{V}}(Y_i|M,Y^{i-1})}{\mathbb{P}_{\widetilde{V}}(Y_i|Y^{i-1})}.
\end{aligned}
\tag{2.22}
$$

Similarly,

$$
\begin{aligned}
\widetilde{Z} &\triangleq \frac{1}{n}\log\frac{\mathbb{P}_{\widetilde{V}}(Y^n|M)}{\mathbb{P}_{W}(Y^n|M)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathbb{P}_{\widetilde{V}}(Y_i|M,Y^{i-1})}{\mathbb{P}_{W}(Y_i|M,Y^{i-1})} \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\frac{\widetilde{V}_{Y^{i-1}}(Y_i|X_i(M,Y^{i-1}))}{W(Y_i|X_i(M,Y^{i-1}))} \\
&= \frac{1}{n}\sum_{i=1}^{n} \widetilde{Z}_i \\
\widetilde{Z}_i &\triangleq \log\frac{\mathbb{P}_{\widetilde{V}}(Y_i|M,Y^{i-1})}{\mathbb{P}_{W}(Y_i|M,Y^{i-1})}.
\end{aligned}
\tag{2.23}
$$

$\{Z_i\}_{i=1}^n$ and $\{\widetilde{Z}_i\}_{i=1}^n$ are random variables whose normalized sums are equal to Z and $\widetilde{Z}$ respectively. Now, we have two claims that amount to nothing more than algebraic bookkeeping (the proofs can be found in Appendix A.6.3).

Proposition 1. *With the choice of $\widetilde{V}$ in (2.17) and the definitions of Z_i and $\widetilde{Z}_i$ in (2.22)* *and (2.23) respectively, we have for $i = 1, \ldots, n$,*

$$\mathbb{E}_{\widetilde{V}}[Z_i] = \sum_{y^{i-1} \in \mathcal{Y}^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) I(P_{y^{i-1}}, \widetilde{V}_{y^{i-1}}) \leq R - 2\delta \tag{2.24}$$

$$\mathbb{E}_{\widetilde{V}}[\widetilde{Z}_i] = \sum_{y^{i-1} \in \mathcal{Y}^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) E_{sp}(R - 2\delta, P_{y^{i-1}}) \leq E_{sp}(R - 2\delta). \tag{2.25}$$

Therefore, by linearity of expectation,

$$\mathbb{E}_{\widetilde{V}}[Z] \leq R - 2\delta < R - \delta$$

$$\mathbb{E}_{\widetilde{V}}\left[\widetilde{Z}\right] \leq E_{sp}(R - 2\delta) < E_{sp}(R - 2\delta) + \delta.$$

Proposition 1 shows the rationale behind setting $\widetilde{V}$ as in (2.17). The random variables we want to converge with high probability in (2.20) and (2.21) have expectations on the proper side of $R - \delta$ and $E_{sp}(R - 2\delta) + \delta$ respectively at least. Now, this is a good start, but at this point all standard ways of proving convergence in probability of a normalized sum of random variables fail because of one crucial point. The point is that the expectations in (2.24) and (2.25) are averaged over both M and Y^n. Necessarily, any 'off-the-shelf' proof of convergence will require the Z_i to be independent or weakly dependent in some sense, otherwise convergence does not hold in general. Unfortunately, while that may be the case if M is held random while Z_i is averaged over, it cannot be the case for general codes once we fix $M = m$. That is, the course of $\frac{1}{n}\sum_{i=1}^{n} Z_i$ strongly depends on M and information about M may be revealed in Z.

For example, consider forming a martingale sequence as follows. Let the filtration be defined as $\mathcal{F}_i = \sigma(M, Y^i)$, and let

$$T_i \triangleq \frac{1}{n}(Z_i - \mathbb{E}_{\widetilde{V}}\left[Z_i | \mathcal{F}_{i-1}\right])$$

$$T = \sum_{i=1}^{n} T_i.$$

Then, T is a martingale because the T_i are zero mean martingale differences by construction. So with appropriate boundedness restrictions on T_i, we know that $T \to 0$ with high probability (w.h.p.). Notably, however, this only implies that

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \to \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\widetilde{V}}\left[Z_i | \mathcal{F}_{i-1}\right] \;\; w.h.p.$$

But,

$$\begin{aligned}
\mathbb{E}_{\widetilde{V}}\left[Z_i|\mathcal{F}_{i-1}\right] &= \mathbb{E}_{\widetilde{V}}\left[Z_i|M,Y^{i-1}\right] \\
&= \mathbb{E}_{\widetilde{V}}\left[\log \frac{\mathbb{P}_{\widetilde{V}}(Y_i|M,Y^{i-1})}{\mathbb{P}_{\widetilde{V}}(Y_i|Y^{i-1})}\middle| M,Y^{i-1}\right] \\
&= \sum_y \mathbb{P}_{\widetilde{V}}(Y_i=y|M,Y^{i-1}) \log \frac{\mathbb{P}_{\widetilde{V}}(Y_i=y|M,Y^{i-1})}{\mathbb{P}_{\widetilde{V}}(Y_i=y|Y^{i-1})} \\
&= \sum_y \widetilde{V}_{Y^{i-1}}(y|X_i(M,Y^{i-1})) \log \frac{\widetilde{V}_{Y^{i-1}}(y|X_i(M,Y^{i-1}))}{(P_{Y^{i-1}}\widetilde{V}_{Y^{i-1}})(y)} \\
&\neq I\left(P_{Y^{i-1}},\widetilde{V}_{Y^{i-1}}\right) \\
&= \sum_x P_{Y^{i-1}}(x) \sum_y \widetilde{V}_{Y^{i-1}}(y|x)) \log \frac{\widetilde{V}_{Y^{i-1}}(y|x))}{(P_{Y^{i-1}}\widetilde{V}_{Y^{i-1}})(y)}.
\end{aligned}$$

We see that $\mathbb{E}_{\widetilde{V}}[Z_i] \neq I(P_{Y^{i-1}},\widetilde{V}_{Y^{i-1}})$ and therefore, it is difficult to say much about what Z converges to except that the average of what it converges to over M, Y^n is less than $R-\delta$. The same problem arises when we attempt to analyze the convergence of $\widetilde{Z}$.

This argument, of course, does not preclude one from using more information about the Z and $\widetilde{Z}$ processes to obtain convergence results in probability. Indeed, it is our opinion that the event when Z is below $R-\delta$ and $\widetilde{Z}$ is below $E_{sp}(R-2\delta)+\delta$ has non-negligible probability for large n.

Conjecture 1. *With the choice of $\widetilde{V}$ as in (2.17), it is true that for some $g(\delta)>0$,*

$$\mathbb{P}_{\widetilde{V}}\left(\frac{1}{n}\log\frac{\mathbb{P}_{\widetilde{V}}(Y^n|M)}{\mathbb{P}_{\widetilde{V}}(Y^n)} > R-\delta\right) \leq \frac{1}{2}-g(\delta),\ n\to\infty$$
$$\mathbb{P}_{\widetilde{V}}\left(\frac{1}{n}\log\frac{\mathbb{P}_{\widetilde{V}}(Y^n|M)}{\mathbb{P}_{W}(Y^n|M)} > E_{sp}(R-2\delta)+\delta\right) \leq \frac{1}{2}-g(\delta),\ n\to\infty.$$

Thus, while there may not be true convergence for the two events, there should at least be non-negligible overlap of the complementary events.

2.4.4 The back-story bound

One of the unsatisfying elements of the Haroutunian exponent is that it leaves open the possibility that one should *try* to design the channel W to be asymmetric in order to beat the sphere-packing exponent with feedback. However, we know that for output-symmetric channels, the sphere-packing bounds holds (because $E_h(R) = E_{sp}(R)$ for those channels). This fact suggests a possible strategy to show that $E_h(R)$ is not achievable for some asymmetric channels. The idea is, given a symmetric channel such as a BEC, the decoder can add noise to the output symbols to simulate an asymmetric channel. The upper bound on the error exponent for the symmetric channel then also is an upper bound for the asymmetric channel because the contraposition leads to a contradiction.

Suppose that the error exponent for the simulated asymmetric channel is indeed larger than the error exponent for the symmetric channel. Then we can code over the symmetric channel by adding noise at the decoder and simulating the asymmetric channel. We thus have a contradiction if the error exponent for the asymmetric channel is larger than the error exponent for the symmetric channel. We call this the back-story bound. Unfortunately, in the examples we have tried, the resulting bound is weaker than the Haroutunian bound. Appendix B.2 shows how a Z-channel simulated by a BEC followed by collapsing the erasure and 0 symbols does not give a better bound than the Haroutunian exponent for the Z-channel.

2.5 Sphere-packing holds when the fixed type encoding tree condition holds

Thus far, we have attempted to show that $E_{fb}(R) \leq E_{sp}(R)$ by looking at feedback codes at a high level. In Section 2.4.3, we attempted to use memoryless of the channel to choose the test channel optimally at each time (and received sequence). Calculating bounds on the relevant probabilities turn out to be difficult, however. We turn our attention now to some analyses that look at codes with feedback in more detail.

In this section, we will show that the sphere-packing bound holds for a restricted class of codes with feedback. The restricted class of codes is said to have encoding trees that satisfy the 'fixed type encoding tree' condition. In these codes, the input codewords all have the same type regardless of the received sequence. The fact that sphere-packing holds for this restricted class of codes with feedback can be proved analogously to Haroutunian's bound for codes without feedback, by first proving a strong converse and then analyzing the error probability after a change-of-measure. We will, however, prove it in a combinatorial manner, analogous to Theorem 1. This approach will illuminate a somewhat surprising fact: that if one is restricted to using a fixed type P (with feedback) for a given message, *at the output, it looks as if feedback is not being used at all.* The precise meaning of this statement will be given later.

Definition 3 (Fixed type encoding tree condition)**.** *An encoding tree of an arbitrary message $m \in \mathcal{M}$ for a fixed-length code with feedback is said to satisfy the* **fixed type encoding tree condition** *if there exists a $P \in \mathcal{P}_n$ such that for all $y^n \in \mathcal{Y}^n$, $P(m, y^n) = P$. Recall that $P(m, y^n)$ denotes the type of the input codeword for message m along a received sequence y^n, i.e., the type of $(x_1(m), x_2(m, y^1), \ldots, x_n(m, y^{n-1}))$.*

Note that an encoding tree satisfying the fixed type encoding tree condition need not give up the use of feedback. That is, the channel input codeword can still depend non-trivially on the received sequence. For example consider the encoding tree in Figure 2.9. The channel input codewords along different received sequences are not necessarily the same.

Proposition 2. *Suppose an encoding tree for a message $m \in \mathcal{M}$ satisfies the fixed type encoding tree condition with type $P \in \mathcal{P}_n$. Then, for all $V \in \mathcal{V}_n(P)$,*

$$\begin{aligned} |B(m,P,V)| &= |T_V(x^n)| \\ &\geq \frac{\exp(nH(V|P))}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}}, \end{aligned}$$

where $x^n \in T_P$ is arbitrary because $|T_V(x^n)|$ depends only on the type of x^n.

This proposition says that if the types of the input vectors are the same for all y^n, then the conditional V-shells have the same size as for a code without feedback. This may be somewhat surprising, because in some sense, even though the encoding tree can use feedback nontrivially, it appears at the output (through the conditional relationship between input and output) as if feedback had not been used at all. Before proving the proposition rigorously, we give some intuition for why this result might hold. First, note that

$$\begin{aligned} B(m,P,V) &= \{y^n : P(m,y^n) = P, V(m,y^n) = V\} \\ &= \left\{ y^n : \forall\ x,y, \sum_{i=1}^n 1(x_i(m,y^{i-1}) = x, y_i = y) = nP(x)V(y|x) \right\} \\ &= \left\{ y^n : \forall\ x,y, \sum_{i=1}^n 1(x_i(m,y^{i-1}) = x, y_i = y) \leq nP(x)V(y|x) \right\}. \end{aligned}$$

The last line follows because if

$$\sum_{i=1}^n 1(x_i(m,y^{i-1}) = x, y_i = y) \neq nP(x)V(y|x),$$

then there is at least one x,y with $\sum_{i=1}^n 1(x_i(m,y^{i-1}) = x, y_i = y) > nP(x)V(y|x)$ because

$$\sum_{x,y}\sum_{i=1}^n 1(x_i(m,y^{i-1}) = x, y_i = y) = \sum_{x,y} nP(x)V(y|x) = n$$

and all terms in the sum are non-negative.

Consider the encoding tree for a message in a code with feedback as shown in Figure 2.9. We wish to count sequences y^n that end up in $B(m,P,V)$. Since $P(m,y^n) = P$ for all $y^n \in \mathcal{Y}^n$, we need to verify that $V(m,y^n) = V$ for a given y^n to be in $B(m,P,V)$. Consider

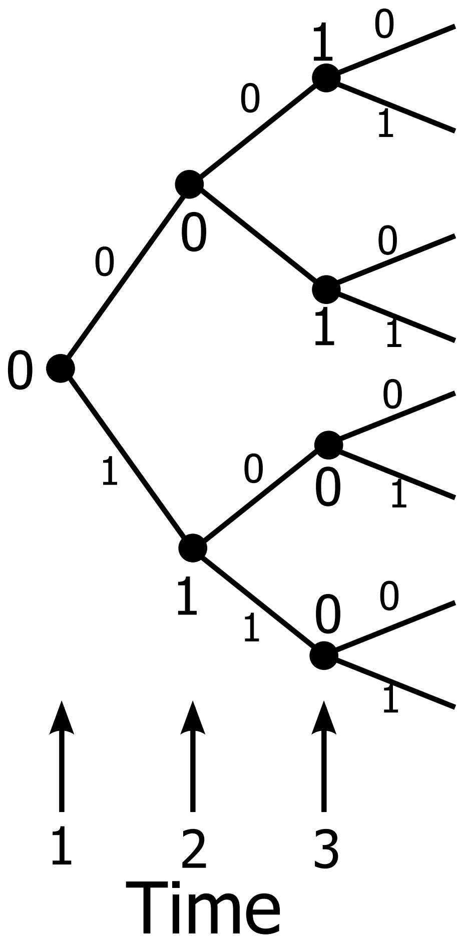

Figure 2.9: A nontrivial encoding tree for a code with feedback, $n = 3, \mathcal{X} = \mathcal{Y} = \{0, 1\}$. A symbol at a node denotes a channel input and a symbol on an edge denotes a channel output. The type of $x^3(m, y^3)$ is $(2/3, 1/3)$ for all $y^3 \in \mathcal{Y}^3$, so this encoding tree satisfies the fixed type encoding tree condition. It uses feedback nontrivially, however, as $x^3(m, (0, 0, 0)) = (0, 0, 1) \neq x^3(m, (1, 1, 1)) = (0, 1, 0)$.

marching along a received sequence from left to right in Figure 2.9. If $y^n \notin B(m, P, V)$, there is some minimal k such that

$$\forall\ x, y,\ \sum_{i=1}^{k} 1(x_i(m, y^{i-1}) = x, y_i = y) \leq nP(x)V(y|x)$$

and

$$\exists\ \widetilde{x}, \widetilde{y}\ \text{ such that } \sum_{i=1}^{k+1} 1(x_i(m, y^{i-1}) = \widetilde{x}, y_i = \widetilde{y}) > nP(\widetilde{x})V(\widetilde{y}|\widetilde{x}).$$

In other words, we are marching along y^n from left to right and at time 0, $y^0 = \emptyset$ still has children in $B(m, P, V)$, y^1 still has children in $B(m, P, V), \ldots,$ y^k still has children in $B(m, P, V)$, but y^{k+1} has no children in $B(m, P, V)$. On the graph of the encoding tree, with nodes corresponding to channel inputs and edges corresponding to channel outputs, we can visualize this by pruning the tree at the edge corresponding to y^{k+1}. The rule for whether or not to prune an edge does not change whether a code has feedback or not, and in some sense the pruning of an edge in one part of the tree does not affect any other part of the tree that is not a child. Therefore, one might intuitively deduce that $B(m, P, V)$ should have at least the same size as $T_V(x^n)$ for some x^n in T_P.

The proposition is proved rigorously (in Appendix A.9.1) by showing that given any $x^n \in T_P$, there must exist a one-to-one mapping τ from $\mathcal{Y}^n$ to $\mathcal{Y}^n$ that has the following property. For all $y^n \in \mathcal{Y}^n$, if $y^n \in T_V(x^n)$ for some $V \in \mathcal{V}_n(P)$, then $\tau(y^n) \in B(m, P, V)$. Therefore, it is not possible via feedback to move sequences into 'good' V-shells (those with high mutual information) from 'bad' ones (those with low mutual information) if the fixed type condition holds. Using Proposition 2, one can prove that the sphere-packing bound holds analogously to the method of types proof of sphere-packing for codes without feedback.

Theorem 3 (Sphere-packing holds if fixed type encoding tree condition holds)**.** *Fix a $\delta > 0, R > 0$. There exists a finite $n_{FT}(W, R, \delta)$ such that for any fixed-length code with feedback of length $n \geq n_{FT}(W, R, \delta)$ and rate R with encoding trees for all messages satisfying the fixed type encoding tree condition,*

$$-\frac{1}{n} \log P_e(W) \leq E_{sp}(R - \delta) + \delta.$$

For a proof, see Appendix A.9.2.

2.6 What needs to be proved for sphere-packing to hold for fixed-length codes with feedback?

This section addresses the question of what is needed to prove the sphere-packing bound for fixed-length block codes with feedback. There are many potential answers to that question, but here we give one take on what sets apart the case with feedback from the case without feedback. To that end, we will define two 'assertions'. The first assertion, call it A, is that the sphere-packing bound holds for codes with feedback. The second assertion, call it B, roughly says that encoding trees for codes with feedback are not so different from input codewords for codes without feedback in the way that output sequences are related back to them.

In this section, we show that if B holds, then so must A. It would be tempting to also claim that if A holds, then so does B, as we would then have an equivalent statement about encoding trees that would both imply and be implied by the assertion that sphere-packing holds for codes with feedback. Rather, what we can show holds is a much weaker condition, call it C, on all the encoding trees in an arbitrary code of length n and rate R. We will show that A implies C and C implies A, and hence B also implies C. All in all, we will have

$$B \Rightarrow A \Leftrightarrow C.$$

Definition 4 (SP holds with feedback). *We say that sphere-packing holds for fixed-length codes with feedback, or* **SP holds with feedback** *for short, if for all $\delta > 0$, $R > 0$ and $W \in \mathcal{W}$, there is a finite $n_{SP,fb}(\delta, R, W)$ such that any fixed-length code with feedback of rate at least R and length $n \geq n_{SP,fb}(\delta, R, W)$ has*

$$-\frac{1}{n}\log P_e(W) \leq E_{sp}(R-\delta) + \delta.$$

Definition 5 (SP Encoding Tree Condition). *We say that the* **sphere-packing encoding tree condition** *holds, if for all $\delta > 0, R > 0, W \in \mathcal{W}$, there is a finite $n_{ET,fb}(\delta, R, W)$ such that the following is true. Given any encoding tree for a fixed-length code with feedback of length $n \geq n_{ET,fb}(\delta, R, W)$, there exists a $P \in \mathcal{P}_n$, $V \in \mathcal{V}_n(P)$ (dependent on the encoding tree) such that*

$$|B(m, P, V)| \geq \exp(n(H(V|P) - \delta)) \tag{2.26}$$

$$D(V||W|P) \leq E_{sp}(R - 2\delta) + 2\delta \tag{2.27}$$

$$I(P, V) \leq R - 2\delta, \tag{2.28}$$

where the m in $B(m, P, V)$ is an indicator of the encoding tree in a code.

Whether the sphere-packing encoding tree condition is a true assertion is unproven currently. The impetus behind its definition is to look at the proof of the sphere-packing theorem for block codes without feedback. In both the case with and without feedback, showing the existence of a $P \in \mathcal{P}_n$ and channel V with $I(P, V) \leq R - \delta$ and $D(V||W|P) \leq E_{sp}(R-\delta)+\delta$ is straightforward for large n. In the case without feedback, we have good estimates for the size of the 'typical' output under channel V for a given message with codeword type P, namely it is exponential in $H(V|P)$. This allows us to prove the part of the statement

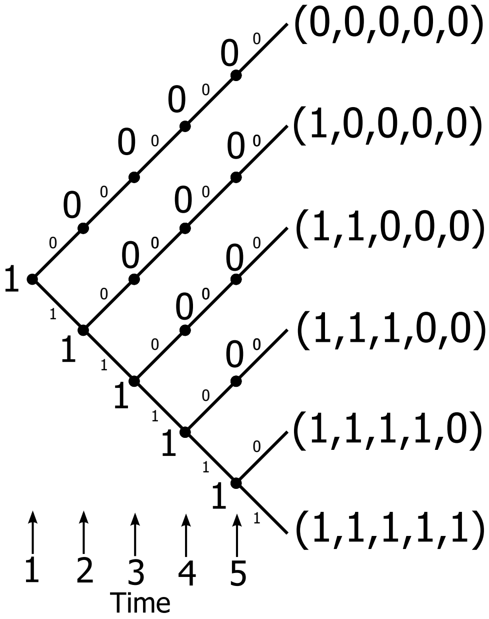

Figure 2.10: The encoding tree of Example 2 that inputs $X_1 = 1$ and $X_{i+1} = Y_i$ for $1 \leq i \leq 4$, with $n = 5$. Nodes in the tree represent input symbols, while edges represent channel output symbols. Only paths in the tree that occur with non-zero probability when W is a Z-channel are shown.

in (2.26) after selection of a P, V that satisfy (2.27) and (2.28). In the case of codes with feedback, however, the portion of the statement in (2.26) certainly does not hold for an arbitrary P, V because we do not know much about the types of the input codewords along different received sequences. We look at a simple example to explore the difficulty in proving the sphere-packing encoding tree coding in general.

Example 2. *Suppose $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and W is a Z-channel with crossover probability 0.5. Consider the encoding tree for which (for arbitrary n),*

$$X_1 = 1$$
$$X_{i+1} = Y_i, i \geq 1.$$

For this example (the encoding tree is shown in Figure 2.10), there are only $n + 1$ strings y^n that occur with non-zero probability (because once a 0 is output, the rest of the output *symbols will be* 0 *with probability* 1*). The strings and their type and conditional types are:*

y^n	$x^n(m,y^n)$	$P(m,y^n)$	$V(m,y^n)$	$H(V\|P)$
$(0,0,0,\ldots,0,0)$	$(1,0,0,\ldots,0)$	$(\frac{n-1}{n},\frac{1}{n})$	$\mathcal{Z}(1)$	0
$(1,0,0,\ldots,0,0)$	$(1,1,0,\ldots,0)$	$(\frac{n-2}{n},\frac{2}{n})$	$\mathcal{Z}(1/2)$	$\frac{2}{n}h_b(1/2)$
$(1,1,0,\ldots,0,0)$	$(1,1,1,\ldots,0)$	$(\frac{n-3}{n},\frac{3}{n})$	$\mathcal{Z}(1/3)$	$\frac{3}{n}h_b(1/3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(1,1,1,\ldots,1,0)$	$(1,1,1,\ldots,1)$	$(0,1)$	$\mathcal{Z}(1/n)$	$h_b(1/n)$
$(1,1,1,\ldots,1,1)$	$(1,1,1,\ldots,1)$	$(0,1)$	$\mathcal{Z}(0)$	0

(2.29)

where $\mathcal{Z}(p)$ *is shorthand for the Z-channel with crossover probability* p *and* $h_b(p) = -p\log p - (1-p)\log(1-p)$ *is the binary entropy function.*

This is an important example to keep in mind because, for almost every $P \in \mathcal{P}_n$ *(except for* $P = (1,0)$*), there is a* $y^n \in \mathcal{Y}^n$ *such that* $P(m,y^n) = P$*. Furthermore, for every* $P \in \mathcal{P}_n$ *that occurs, there is exactly one* y^n *with* $P(m,y^n) = P$*. So, in this example, if* $P \in \mathcal{P}_n, V \in \mathcal{V}_n(P)$ *and* $D(V||W|P) < \infty$*, either* $|B(m,P,V)| = 0$ *or* $|B(m,P,V)| = 1$*. It appears in this example as if the sphere-packing encoding tree condition will not hold because none of the* $|B(m,P,V)|$ *are exponential in* n*. A closer inspection, however, is in order for the right hand side column above showing* $H(V|P)$ *for all* $B(m,P,V)$ *that are not empty. The maximum* $H(V|P)$ *there is, for a given* n*,*

$$\begin{aligned}
k_n &\triangleq \max_{2\le m\le n} \frac{m}{n} h_b\left(\frac{1}{m}\right) \\
&= \max_{2\le m\le n} \frac{m}{n}\left(\frac{1}{m}\log m + \frac{m-1}{m}\log\frac{m}{m-1}\right) \\
&= \frac{1}{n}\max_{2\le m\le n}\left(\log m + (m-1)\log\frac{m}{m-1}\right) \\
&= \frac{1}{n}\max_{2\le m\le n}\left(m\log m - (m-1)\log(m-1)\right).
\end{aligned}$$

The function $x\log x$ *is convex-*$\cup$*, so the maximum above occurs when* $m = n$ *and hence,*

$$\begin{aligned}
k_n &= \frac{1}{n}\left(n\log n - (n-1)\log(n-1)\right) \\
&= \log n - \frac{n-1}{n}\log(n-1) \\
&= \log\frac{n}{n-1} + \frac{1}{n}\log(n-1) \\
&\le \frac{1}{n-1} + \frac{1}{n}\log(n-1).
\end{aligned}$$

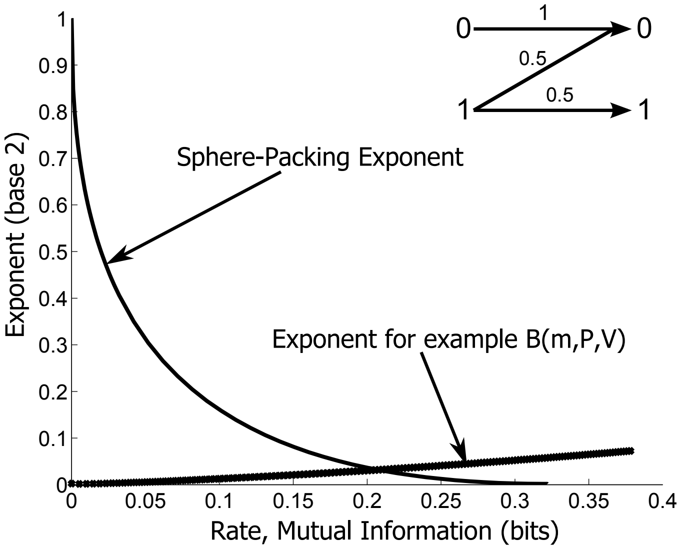

Figure 2.11: This plot shows the sphere-packing exponent when W is a Z-channel with crossover probability 0.5. Also shown is a scatter plot of $D(V||W|P)$ versus $I(P,V)$ for the first 150 $(P,V) = (P(m,y^n), V(m,y^n))$ in (2.29) when $n = 2,000$. The plot highlights the fact that as $n \to \infty$, there are pairs (P,V) in (2.29) such that the point $(I(P,V), D(V||W|P))$ is arbitrarily close to $(0,0)$. Therefore, for this encoding tree, there is a (P,V) that satisfies (2.26), (2.27) and (2.28).

Therefore, $k_n = O((1/n)\log n)$ *and for every* $\delta > 0$*, for large enough* n*,* $H(V|P) < \delta$ *for every* P,V *with* $|B(m,P,V)| = 1$*. So for a fixed* $\delta > 0$ *and large enough* n*, every* P,V *with* $|B(m,P,V)|$ *satisfies (2.26). In order to check that there is a* P,V *such that (2.27) and (2.28) hold as well, we plot* $I(P,V)$ *versus* $D(V||W|P)$ *for* $n = 2,000$ *in Figure 2.11. For this example encoding tree, for large enough* n*, there are* (P,V) *that are arbitrarily close to* $(0,0)$*, so the sphere-packing encoding tree condition holds almost trivially. This corresponds to using the sequences with just a few* 1*'s at the beginning in order to prove error bounds.*

Currently, it seems difficult to prove the sphere-packing encoding tree condition holds except in special cases (like an encoding tree without feedback or one that satisfies the fixed type encoding tree condition). This example shows the core difficulty. If we fixate on one P for which we know that there are sequences with $P(m,y^n) = P$, it is difficult (and untrue) to say that $B(m,P,V)$ is large for arbitrary $V \in \mathcal{V}_n(P)$. All we can say is that there exists at least one $V \in \mathcal{V}_n(P)$ for which $B(m,P,V)$ can be quantified in some way. This V may have $D(V||W|P) > E_{sp}(R)$ or $I(P,V) > R$ however and cannot be used to prove the sphere-packing bound. That said, if the sphere-packing encoding tree condition holds (i.e., it can be proved for all codes with feedback), the sphere-packing bound also holds for all codes with feedback used over general DMCs.

Proposition 3 (SP Encoding Tree condition implies SP holds)**.** *If the sphere-packing encoding tree condition holds, then sphere-packing for fixed-length codes with feedback holds.*

For a proof of the proposition, see Appendix A.11.1. One might be tempted to say that if sphere-packing holds for fixed-length codes with feedback, then the sphere-packing encoding tree condition also must hold. Unfortunately, this seems difficult to prove as well. What can be shown to hold is a much weaker condition.

Definition 6 (Intermediate SP Condition)**.** *We say that the* **intermediate sphere-packing condition** *holds, if for all $\delta > 0, R > 0, W \in \mathcal{W}$, there is a finite $n_{SPI,fb}(\delta, R, W)$ such that the following is true. Given any fixed-length code of rate at least R and length $n \geq n_{SPI,fb}(\delta, R, W)$, there exists a $P \in \mathcal{P}_n$ and $V \in \mathcal{V}_n(P)$ such that*

$$D(V||W|P) \leq E_{sp}(R - \delta) + 2\delta \tag{2.30}$$

$$\frac{1}{n} \log \left[\frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} |B(m, P, V) \cap \mathcal{D}_m^c| \right] \geq H(V|P) - [E_{sp}(R - \delta) + 2\delta - D(V||W|P)] \,. \tag{2.31}$$

Note that if (2.31) holds, then (2.30) also holds by properties of $B(m, P, V)$ (Proposition 14). However, (2.30) reinforces the intuition that (P, V) are such that error probability under channel W is at least the probability prescribed by the sphere-packing bound.

The intermediate sphere-packing condition turns out to be equivalent to sphere-packing holding for codes with feedback. For a proof of this result, see Appendix A.11.2. The intermediate sphere-packing condition is quite un-intuitive, however, so it does not seem promising to attempt to prove that it must hold in general.

Proposition 4 (Equivalence of SP and Intermediate SP condition)**.** *If SP holds with feedback, then the intermediate SP condition holds. Conversely, if the intermediate SP condition holds, SP holds with feedback.*

2.7 Delayed feedback is not useful for very large delays

This section gives an upper bound to the error exponent for block codes used over DMCs with noiseless feedback, where the feedback is delayed by some fixed number of symbols. It is interesting to consider this problem, because in modern, high rate communication systems, the number of symbols that must be encoded before the encoder receives a previous channel output (or more likely, a function of the channel output) can be potentially large. Two possible reasons for this gap between sending a channel input and receiving information about the channel output come to mind: propagation delays and the inherent processing time for demodulation and other processing at the decoder.

Consider communicating 20 symbols per microsecond on a 20 MHz channel over a distance of 1.5 km (round trip 3km). Even without accounting for processing time, the delay for an electromagnetic signal to travel to and fro would be 10 microseconds, meaning at least 200 symbols should have been transmitted before feedback can be received. Additionally, many communications systems have a half-duplex constraint, meaning they cannot listen and transmit at the same time. Thus, feedback information may not return until the transmitter is finished transmitting some appropriate 'block' of symbols.

Suppose that information about the channel outputs is delayed by T symbols. Then, the result of this section is that the error exponent for rate R codes used with noiseless feedback delayed by T symbols is upper bounded by

$$E_{sp}(R - O(\log T/T)) + O(\log T/T),$$

where the constants hidden in the big-O notation depend on the channel transition matrix W. Hence, for large delays in the feedback path, the sphere-packing bound is essentially an upper bound on the error exponent for fixed-length block codes. To avoid confusion, the result applies to a fixed delay in the feedback path (T), as the blocklength (n) goes to infinity. Before giving the result, some definitions germane to this section are in order.

2.7.1 Problem Setup

A rate R, blocklength n coding system is an encoder-decoder pair $(\mathcal{E}, \mathcal{D})$.

Definition 7 (Type 1 Encoder - Delay T feedback)**.** *A rate R, blocklength n encoder $\mathcal{E}$ used with feedback delayed by T symbols is a sequence of maps $\{\phi_i\}_{i=1}^n$, with for $1 \leq i \leq T$,*

$$\phi_i : \{1, 2, \ldots, 2^{nR}\} \to \mathcal{X},$$

and for $i > T$,

$$\phi_i : \{1, 2, \ldots, 2^{nR}\} \times \mathcal{Y}^{i-T} \to \mathcal{X}.$$

Note that $T = 1$ is the usual perfect-feedback setting where the encoder is aware of the channel output immediately before the next channel input must be selected.

Definition 8 (Type 2 Encoder - T block feedback)**.** *This is a more powerful class of encoding systems than the 'Delay T feedback' encoders. Here feedback is provided to the encoder in blocks of T symbols at a time. That is, $(Y_1, \ldots, Y_T)$ is given to the encoder before the encoder chooses X_{T+1} and in general, the received symbol block $(Y_{iT+1}, \ldots, Y_{(i+1)T})$ is provided to the encoder at time $(i+1)T$ before the encoder must choose $X_{(i+1)T+1}$. Hence a blocklength n type 2 encoder with rate R is a sequence of maps $\{\phi_i\}_{i=1}^n$,*

$$\phi_i : \{1, \ldots, 2^{nR}\} \times \mathcal{Y}^{\lfloor (i-1)/T \rfloor T} \to \mathcal{X}.$$

Note that a type 1 encoder is a restricted type 2 encoder that does not use all possible symbols that have been fed back.

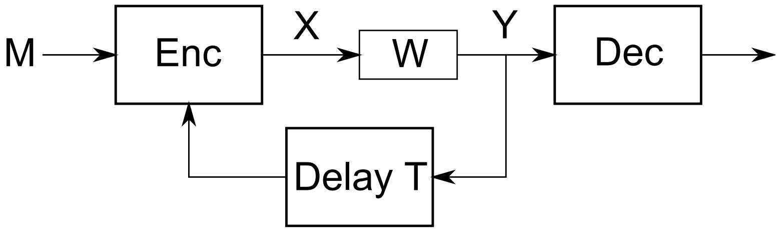

Figure 2.12: Block coding for a DMC W with delayed feedback.

A blocklength n decoder $\mathcal{D}$ is a map

$$\psi : \mathcal{Y}^n \to \{1, 2, \ldots, 2^{nR}\}.$$

The decoding regions for each message are then $\mathcal{D}_m \triangleq \psi^{-1}(m) = \{y^n : \psi(y^n) = m\}$ for $m \in \{1, \ldots, 2^{nR}\}$. The average probability of error for a rate R, blocklength n coding system $(\mathcal{E}, \mathcal{D})$ is thus defined as

$$P_e(n, \mathcal{E}, \mathcal{D}) = \frac{1}{2^{nR}} \sum_{m=1}^{2^{nR}} \mathbb{P}(\psi(Y^n) \neq m | M = m).$$

If we let $\mathcal{C}(n, R, T)$ be the appropriate set of blocklength n coding systems with rate at least equal to R used with delay T feedback type 2 encoders, we define the relevant error exponent at rate R to be

$$E(R, T) = \limsup_{n \to \infty} -\frac{1}{n} \log \min_{(\mathcal{E}, \mathcal{D}) \in \mathcal{C}(n, R, T)} P_e(n, \mathcal{E}, \mathcal{D}).$$

Haroutunian's bound applies to codes without any delay in the feedback path, so it also applies for any $T > 1$, hence for all $T \geq 1$, $E(R, T) \leq E_h(R)$.

2.7.2 Error exponent bound for delayed feedback

We prove a bound on type 2 encoding systems, which immediately becomes a bound on type 1 encoders as well. Without loss of generality, we restrict attention to n that are multiples of T, that is $n = NT$ for some $N \geq 1$ (i.e., there are N total blocks of size T symbols each).

Lemma 5. *Define a channel independent constant*

$$\alpha(T) \triangleq \frac{|\mathcal{X}|(2 + |\mathcal{Y}|) \log(T + 1)}{T}.$$

Fix an $\epsilon > \alpha(T)$. Then, for any blocklength $n = NT$ (with $N \geq 1$) rate R coding system with a type 2 encoder,

$$-\frac{1}{NT}\log P_e(NT,\mathcal{E},\mathcal{D}) \leq E_{sp,T}(R-\epsilon) + \alpha(T) + \gamma(N,T,\epsilon),$$

where

$$E_{sp,T}(R) \triangleq \max_{P\in\mathcal{P}_T}\min_{V\in\mathcal{V}_T(P):I(P,V)\leq R} D(V||W|P), \tag{2.32}$$

and

$$\gamma(N,T,\epsilon) = \frac{1}{NT}\log\frac{1}{1-\exp(-NT(\epsilon-\alpha(T)))}.$$

The proof of this lemma is given in Appendix A.12.1. The proof chooses the test channel V within each block of length T according to the most used type of the input during that block. Notice how in (2.32), we have already interchanged the order of the max and min in the Haroutunian exponent, but the set of V is restricted to have low mutual information over length T. We now give an inequality relating the 'sphere-packing bound for length-T' with the sphere-packing bound (proved in Appendix A.12.2).

Lemma 6. *For any $T \geq 2|\mathcal{X}||\mathcal{Y}|$, for all $P \in \mathcal{P}_T$,*

$$\min_{U\in\mathcal{V}_T(P):I(P,U)\leq R} D(U||W|P) \leq E_{sp}\left(R - \frac{2|\mathcal{X}||\mathcal{Y}|\log T}{T}, P\right) + \frac{\kappa_W|\mathcal{X}||\mathcal{Y}|}{T} + \frac{|\mathcal{X}||\mathcal{Y}|\log(T/|\mathcal{X}|)}{T},$$

where[16]

$$E_{sp}(R,P) \triangleq \min_{V:I(P,V)\leq R} D(V||W|P),$$

$$\kappa_W \triangleq \max_{x,y:W(y|x)>0}\log\frac{1}{W(y|x)}.$$

Putting the two lemmas together, we get

[16] The constant κ_W can be arbitrarily large, but is finite for a given W. It is our opinion that the appearance of κ_W in our bound is an artifact of the proof method, and not intrinsic to the problem. However, to our knowledge, such a term appears in most, if not all, proofs of the sphere-packing bound where an explicit lower bound to probability of error as a function of blocklength is given. For example, see Theorem 5.8.1 of [17] or Theorem 1 and 2 of [36].

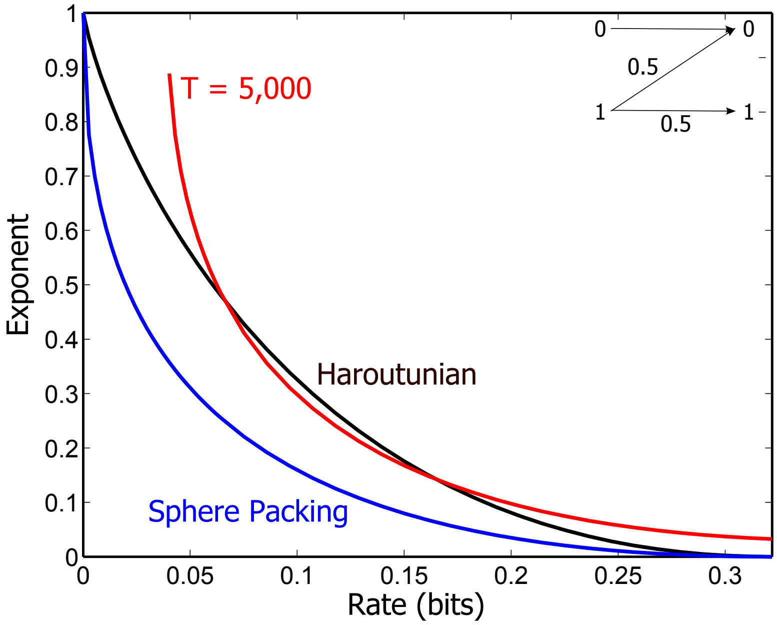

Figure 2.13: A plot of the bound on $E(R,T)$ from Theorem 4 for the Z-channel with $T = 5,000$ (all quantities are base 2). The bound is only approaching the usefulness of the Haroutunian bound (which works for all $T \geq 1$) at this point.

Theorem 4. *For any* $T \geq 2|\mathcal{X}||\mathcal{Y}|$,

$$E(R,T) \leq \quad E_{sp}\left(R - \alpha(T) - \frac{2|\mathcal{X}||\mathcal{Y}|}{T}\log T\right) + \frac{\kappa_W|\mathcal{X}||\mathcal{Y}|}{T} + \frac{|\mathcal{X}||\mathcal{Y}|}{T}\log\frac{T}{|\mathcal{X}|} + \alpha(T).$$

In other words, for type 1 *and type* 2 *coding systems with feedback delay* T *and rate* R,

$$E(R,T) \leq E_{sp}(R - O(\log T/T)) + O(\log T/T).$$

Proof: The result of Lemma 5 is monotonic in the rate, so we need only that $R_N \geq R$ and bound using R. We have just combined the results of the two lemmas together and taken the limit as N tends to ∞. The only thing that needs to be checked is that the term

$$\frac{1}{NT}\log\frac{1}{1-\exp_2(-NT(\epsilon-\alpha(T)))} \tag{2.33}$$

converges to 0 as $N \to \infty$ for any $\epsilon > \alpha(T)$, which is readily seen. Therefore, we can take ϵ arbitrarily close to $\alpha(T)$ and since E_{sp} is continuous for all R expect possibly the R at which E_{sp} becomes infinite, we substitute $\alpha(T)$ for ϵ.

Figures 2.13 and 2.14 show the resulting bound on $E(R,T)$ for the Z-channel when $T = 5,000$ and $50,000$ respectively. As can be seen, even at $T = 5,000$, the constants in the bound for $E(R,T)$ do not bring it close enough to E_{sp} to be much more useful than $E_h(R)$, so this problem is not necessarily 'practically' solved either.

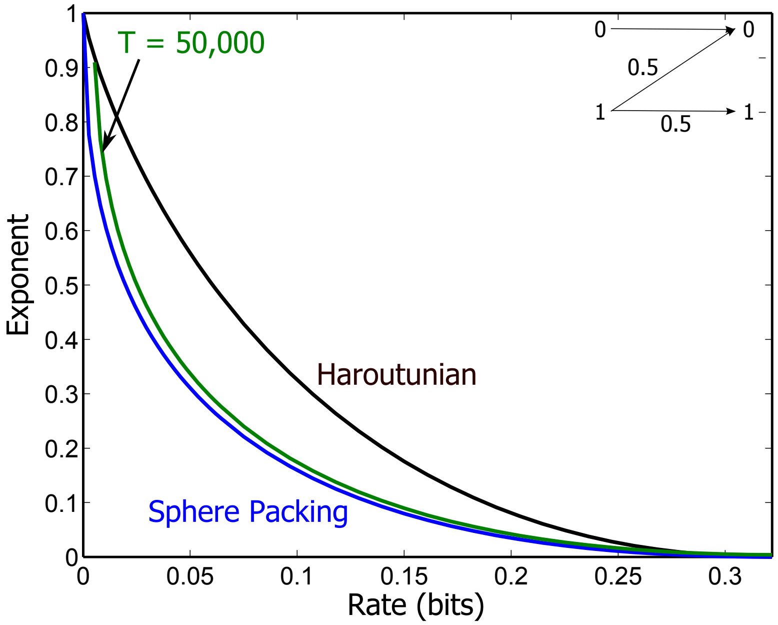

Figure 2.14: A plot of the bound on $E(R, T)$ from Theorem 4 for the Z-channel with $T = 50,000$ (all quantities are base 2). At this point, the upper bound is 'essentially' the same as sphere-packing. If a system has a delay in its feedback path this large for the Z-channel, the feedback does not help increase the error exponent for all practical purposes.

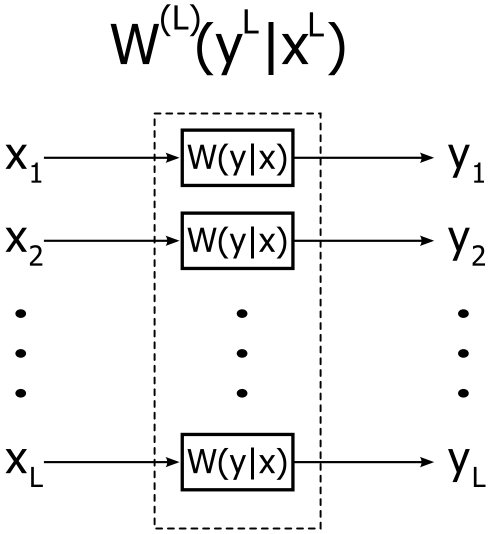

Figure 2.15: The parallel channel $W^{(L)}$ obtained by using W L times. Each use of W is independent. The capacity of $W^{(L)}$ is $LC(W)$ and the scaled sphere-packing exponent $E_{sp}(LR; W^{(L)})$ for $W^{(L)}$ at rate LR is equal to $LE_{sp}(R; W)$.

2.8 The Haroutunian exponent for a parallel channel

In this section, we interpret the result of Section 2.7 and arrive at a surprising conclusion. We will think of the block-T feedback coding systems of Section 2.7 (where the encoder receives output symbols in blocks of size T) as having instantaneous feedback when used over a parallel channel. Shifting notation a bit, let $L \geq 1$ be an integer. A block-L feedback system receives feedback in increments of L symbols at a time, so consider the parallel channel one gets by using L symbols from $\mathcal{X}$ at a time, as shown in Figure 2.15. The block-L feedback system can then be viewed as a code with instantaneous feedback used over the DMC $W^{(L)}$ with input alphabet $\mathcal{X}^L$, output alphabet $\mathcal{Y}^L$ and transition probabilities

$$W^{(L)}(y^L|x^L) = \prod_{i=1}^{L} W(y_i|x_i).$$

It is straightforward to show that $C(W^{(L)}) = LC(W)$, and with a bit more effort, it can be shown that $E_{sp}(LR; W^{(L)}) = LE_{sp}(R; W)$ where $E_{sp}(LR; W^{(L)})$ denotes the sphere-packing exponent for channel $W^{(L)}$ evaluated at rate LR and $E_{sp}(R; W)$ denotes the sphere-packing exponent for channel W at rate R. Theorem 4 then shows that

$$E_{fb}(LR; W^{(L)}) \leq LE_{sp}(R; W) + O\left(\frac{\log(L+1)}{L}\right).$$

Another upper bound on $E_{fb}(LR; W^{(L)})$ can be had by applying the Haroutunian bound to the channel $W^{(L)}$, yielding $E_{fb}(LR; W^{(L)}) \leq E_h(LR; W^{(L)})$, where $E_h(LR; W^{(L)})$ is the

Haroutunian exponent for $W^{(L)}$ evaluated at rate LR . Considering that $C(W^{(L)}) = LC(W)$ and $E_{sp}(LR; W^{(L)}) = LE_{sp}(R; W)$, it wouldn't be much of a stretch to guess that $E_h(LR; W^{(L)}) = LE_h(R; W)$. If W is an asymmetric channel such as the Z-channel, then $LE_h(R; W) > LE_{sp}(R; W) + O\left(\frac{\log(L+1)}{L}\right)$ for large enough L. This means that we have the following interesting dichotomy: either (a) or (b) below must be true.

(a) We have shown that $E_{fb}(LR; W^{(L)}) < E_h(LR; W^{(L)})$ for some asymmetric channel $W^{(L)}$ ($W^{(L)}$ is still asymmetric if W is for all $L \geq 1$). That is, we have shown a class of channels for which the Haroutunian exponent is strictly loose, which would partially confirm our hypothesis that sphere-packing is a valid upper bound on the error exponent for block codes with feedback for asymmetric channels as well.

(b) The Haroutunian exponent for the parallel channel $E_h(LR; W^{(L)})$ does not decompose like the sphere-packing exponent and capacity, i.e.,

$$E_h(LR; W^{(L)}) < LE_h(R; W).$$

Surprisingly, it turns out that (b) is true[17]. In the optimizations that characterize $C(W^{(L)})$ (maximization over distributions on $\mathcal{X}^L$) and $E_{sp}(LR; W^{(L)})$ (maximization over distributions on $\mathcal{X}^L$ and minimizations over channels from $\mathcal{X}^L$ to $\mathcal{Y}^L$), it suffices to consider product distributions due to the concavity and convexity properties of mutual information and divergence. This causes the optimizations to decompose into optimizations over L individual symbols, resulting in the multiplicative scaling. In the optimization for the Haroutunian exponent, it turns out that it is not sufficient to consider channels V from $\mathcal{X}^L$ to $\mathcal{Y}^L$ that are independent over time, i.e., $V(y^L|x^L) = \prod_{i=1}^L V_i(y_i|x_i)$.

Lemma 7. *For the parallel channel $W^{(L)}$ with input alphabet $\mathcal{X}^L$, output alphabet $\mathcal{Y}^L$ and transition probabilities*

$$W^{(L)}(y^L|x^L) = \prod_{i=1}^{L} W(y_i|x_i),$$

we have

$$E_h(LR; W^{(L)}) \leq LE_{sp}\left(R - \frac{|\mathcal{X}|}{L}\log(L+1); W\right).$$

Thus, in the limit as $L \to \infty$ (because E_{sp} is left-continuous),

$$\lim_{L\to\infty} \frac{1}{L} E_h(LR; W^{(L)}) \leq E_{sp}(R; W).$$

[17] While (a) may be true in the sense that the Haroutunian exponent is strictly loose for some class of asymmetric channels, we have not shown it to be true by the result in Section 2.7.

Note that for all L,

$$\frac{1}{L}E_h(LR; W^{(L)}) \geq \frac{1}{L}E_{sp}(LR; W^{(L)}) = E_{sp}(R; W).$$

The proof of this result is given in Appendix A.13, but we will describe here the test channel V that is used to prove the result. Fix $\epsilon_L = \frac{|\mathcal{X}|}{L}\log(L+1)$ and for each $P \in \mathcal{P}_L$, let

$$U_P \in \arg \min_{U : I(P,U) \leq R - \epsilon_L} D(U||W|P),$$

and define $V(\cdot|x^L)$ according to the type of x^L. If $x^L \in T_P$, let for all y^L

$$V(y^L|x^L) = \prod_{i=1}^{L} U_P(y_i|x_i).$$

Note that V is a product channel only when restricting the type of x^L. The fact that V depends on the type of x^L makes it a non-product channel. In Appendix A.13, it is shown that for this V, $C(V) \leq LR$, so it is included in the minimization for the Haroutunian exponent. It is also shown that the maximal divergence between V and $W^{(L)}$ is at most $LE_{sp}(R - \epsilon_L; W)$.

The intuitive reason for this lemma is that the channel $W^{(L)}$ is becoming 'more symmetric' as L gets large. The channel $\{W(\cdot|x^L)\}_{x^L \in T_P}$ for any given type $P \in \mathcal{P}_L$ is output symmetric, and the number of types is at most $(L+1)^{|\mathcal{X}|}$, which is inconsequential with respect to the size of the channel input alphabet (which is growing exponentially with L) in the limit. This lemma can also be used to prove the result in the delayed feedback section by applying the Haroutunian bound to the parallel channel. In Section 3.5, we show how the lemma can be used to prove that fixed delay codes without feedback cannot have an error exponent larger than the sphere-packing exponent.

2.9 Limited memory at the encoder means the sphere-packing bound holds

There are two results in this section that chip away at the notion that feedback might be able to beat the sphere-packing bound. They are both due to Baris Nakiboglu and Giacomo Como of MIT, coming out of their interaction with us (HP and Anant Sahai) on the problem of proving sphere-packing for fixed-length codes with feedback. The first result was suggested as an exercise by Anant Sahai to Baris and Giacomo to investigate what sort of degradation to the assumption of delayless, noiseless feedback allows one to prove the sphere-packing bound.

2.9.1 An encoder that dumps its memory is bound by the sphere-packing exponent

Consider the class of feedback encoders that empty their memory (of received symbols) every k symbols. So for all i, y^{i-1}, X_i is a function of the message m and the previous $(i-1)-\lfloor \frac{i-1}{k} \rfloor k$ received symbols,

$$x_i(m, y^{i-1}) = x_i\left(m, y^{i-1}_{\lfloor \frac{i-1}{k} \rfloor k+1}\right).$$

That is, the encoder stores at successive times, 0, then 1, then 2 received symbols, and so on, until it reaches $k-1$, at which point it erases its memory. We will see that this class of codes is bound by the sphere-packing exponent. This will be done by viewing blocks of k symbols for the original channel as supersymbols and inputs to a new channel. Let

$$\begin{aligned} \mathcal{X}' &= \mathcal{X} \times \mathcal{X}^{|\mathcal{Y}|} \times \mathcal{X}^{|\mathcal{Y}|^2} \times \cdots \times \mathcal{X}^{|\mathcal{Y}|^{k-1}} \\ \mathcal{Y}' &= \mathcal{Y}^k \end{aligned}$$

be the input and output alphabets for the new super channel. Let $\mathcal{P}'$ be the set of distributions on $\mathcal{X}'$, $\mathcal{Q}'$ the set of distributions on $\mathcal{Y}'$ and $\mathcal{W}'$ the set of probability transition matrices from $\mathcal{X}'$ to $\mathcal{Y}'$. The input letter for a message on one super symbol is actually a vector of k functions from the received symbols available to the encoder to the $\mathcal{X}$ space. That is,

$$\begin{aligned} x' \in \mathcal{X}' \Rightarrow x' &= (x_1, f_2, \ldots, f_k) \\ x_1 &\in \mathcal{X} \\ f_i &: \mathcal{Y}^{i-1} \to \mathcal{X},\ i = 2, \ldots, k. \end{aligned}$$

So, $x_i(m, y^{i-1}) = f_i(y^{i-1})$ for $i = 2, \ldots, k$ and x_1 for $i = 1$ if x' is the supersymbol for the first k symbols of message m. Now, the channel for the supersymbols induced by W is

$$\begin{aligned} W'(y'|x') &= W'(y_1, \ldots, y_k | x_1, f_2, \ldots, f_k) \\ &= W(y_1|x_1) W(y_2|f_2(y_1)) \cdots W(y_k|f_k(y^{k-1})). \end{aligned}$$

For this superchannel W', the sphere-packing bound holds because the code with supersymbols does not use feedback (all the feedback used has been hidden in the choice of feedback functions f in the supersymbol). Therefore, the sphere-packing bound for codes without feedback (Theorem 1) holds for this superchannel. Let $E'_{sp}(R)$ denote the sphere-packing exponent for this superchannel at rate R,

$$E'_{sp}(R) = \max_{P' \in \mathcal{P}'} \min_{V' \in \mathcal{W}' : I(P', V') \leq R} D(V' || W' | P').$$

Proposition 5. *The following relationship between $E_{sp}(R)$ and $E'_{sp}(R)$ holds for all $k \geq 2$ (i.e., all interesting k):*

$$E'_{sp}(R) = kE_{sp}\left(\frac{R}{k}\right).$$

This result is due to Baris Nakiboglu and Giacomo Como [33].

The proof of this proposition does not appear in the literature, so it is included in Appendix A.10. The proof shows that within the supersymbols, the sphere-packing optimization decomposes into k individual sphere-packing optimizations for rate R/k.

Now, we see that the sphere-packing exponent for a k-block supersymbol is at most k times the sphere-packing exponent at rate R/k for the original channel. Note that the new length of the code (in superblocks) is n/k and the new rate is thus $\frac{k}{n}\log|\mathcal{M}| = \frac{k}{n}nR = kR$. Hence, the asymptotic bound is still

$$\limsup_{n\to\infty} \frac{1}{n}\log P_e(W) \leq \frac{1}{k}E'_{sp}(kR) \leq E_{sp}(R).$$

Of course, the asymptotics hide the terrible dependence of the bound on the memory length k. The size of the input alphabet grows doubly exponentially in k, as

$$\begin{aligned}\log|\mathcal{X}'| &= \sum_{i=0}^{k-1} |\mathcal{Y}|^i \log|\mathcal{X}|\\ \log|\mathcal{Y}'| &= k\log|\mathcal{Y}|.\end{aligned}$$

Hence, this bound does not scale well, but it does show that having access to only the recent past does not allow us to beat the sphere-packing bound with feedback (in an exponential sense).

2.9.2 Finite state memory encoders cannot beat the sphere-packing bound

The following result of Giacomo Como and Baris Nakiboglu is a philosophical extension of the idea that forcing the encoder to forget feedback symbols renders it unable to beat the sphere-packing bound asymptotically. Instead of encoders that forget all memory periodically, it applies to encoders that can only hold on to all received information in a summarized way through a finite state. Hence, it smooths out the periodic dumping of all memory from the previous problem setup to a dumping of individual symbols at each time.

Define the class of finite memory encoders with feedback as follows. For each $i \geq 1$, the input at time i is a function of the message m and a state $s_i \in \mathcal{S}$ where $\mathcal{S}$ is a finite state alphabet. That is,

$$x_i(m, y^{i-1}) = \Phi_i(m, s_i)$$

and

$$s_{i+1} = \Gamma_i(s_i, y_i)$$

for some state update functions $\Gamma_i : \mathcal{S} \times \mathcal{Y} \to \mathcal{S}$. If $\mathcal{S}$ were infinite, we could simply store all the past received symbols, but by keeping it finite, the encoder has limited memory of all the received symbols, summarized by the state. Assume that there exists a finite k such that for all $i \geq 1$, for all $s, s' \in \mathcal{S}$,

$$\exists\ y_{i+1}^{i+k} \in \mathcal{Y}^k : \Gamma_{i+k}(\Gamma_{i+k-1}(\cdots(\Gamma_{i+1}(s, y_{i+1}), y_{i+2}), \cdots), y_{i+k}) = s'. \tag{2.34}$$

This is a reachability constraint on the encoder's state update mechanism. Essentially, it also enforces that eventually (i.e., after k symbols) every received symbol is forgotten at the encoder. Furthermore, assume that

$$\tau_W \triangleq \min_{x,y} W(y|x) > 0.$$

Theorem 5. *[36] Fix a channel W with $\tau_W > 0$ and suppose the encoder with feedback is as described above with k as in (2.34). Then,*

$$\begin{aligned} -\frac{1}{n}\log P_e(W) &\leq \min_{l \in \{1,\ldots,n\}} E_{sp}(R - \alpha(l)) + \alpha(l) \\ \alpha(l) &\triangleq \frac{2k \log \frac{e}{\tau_W}}{l\tau_W^k} + \frac{l}{n}(\log 4 + |\mathcal{S}| \log |\mathcal{X}|) + \\ &\qquad \frac{\log(1 + n/l)}{n} \exp\left(l|\mathcal{S}|(|\mathcal{Y}| \log |\mathcal{S}| + \log |\mathcal{X}|)\right) \end{aligned}$$

Note that as $n \to \infty$, l can scale slightly sub-logarithmically with n and we will have $\alpha(l) \to 0$ as $n \to \infty$. So, again, for this restricted class of encoders with feedback used over channels with no zero elements, we see that the sphere-packing bound holds asymptotically. The proof in [36] uses the reachability condition on the encoder to show that certain probabilistic terms relating to mutual information and divergence converge to their average. This is done by a mixing argument for Markov chains.

2.10 Concluding remarks

In this chapter, we have presented a collection of mini-results that we hope has advanced understanding of the error exponent problem for block codes with feedback. The partial results showing that having feedback information from only the very far past or only the near past does not allow codes to beat the sphere-packing bound seem to suggest that the question is not far from begin 'practically' solved (meaning solved for any case one might see in practice, due to limited memory at the encoder or delay in the feedback path). The partial result for fixed type encoding trees, however, shows that in some cases, we can precisely show the same sphere-packing bound holds for codes with feedback and without. In the general case, the sphere-packing bound that holds for codes with feedback may turn out to be slightly looser than the bound for codes without feedback, but only polynomially looser in the blocklength n. In our estimation, the most promising paths to showing that sphere-packing holds with feedback are coming up with a refined analysis for the 'optimized' test-channel with memory (Section 2.4.3) and proving the sphere-packing encoding tree condition holds (Section 2.6). If these proofs can be made to go through, they may resolve the uncertainty about asymmetric channels from the perspective of Polyanskiy et al., as discussed in Section 2.1.

Chapter 3

Tightening to sphere-packing for fixed delay coding without feedback

3.1 Introduction

For communicating information over a noisy channel, capacity is the first-order description of how much information we can communicate reliably. A refined understanding can be gained through studying the error exponent for a channel. The error exponent governs the exponential decay with which the optimal error probability decays as more and more channel 'resources' (e.g. block length, expected block length, delay) are provided while maintaining a given rate of information transmission. If d is the number of channel 'uses' and $P_e(d, R)$ is the optimal error probability for a rate R that 'uses' the channel d times, then the error exponent is generally defined to be

$$E(R) = \lim_{d\to\infty} -\frac{1}{d}\log P_e(d, R).$$

In Chapter 2, we have discussed at length the problem of upper bounding the error exponent for fixed-length block codes with and without feedback. For the reader who is reading this chapter independently, we briefly summarize the relevant discussion. Recall that for block codes without feedback, the sphere-packing bound,

$$E_{sp}(R) = \max_{P} \min_{V:I(P,V)\leq R} D(V||W|P),$$

is an upper bound to the error exponent (with respect to blocklength). For codes with feedback, the Haroutunian exponent,

$$\begin{aligned} E_h(R) &= \min_{V:C(V)\leq R} \max_{P} D(V||W|P) \qquad (3.1) \\ &= \min_{V:C(V)\leq R} \max_{x\in\mathcal{X}} D(V(\cdot|x)||W(\cdot|x)) \end{aligned}$$

is the best known upper bound to the error exponent (with respect to blocklength). Recall also that, except for symmetric channels such as the BSC and BEC, $E_{sp}(R) < E_h(R)$ in general. The reason for this gap is believed to be the proof technique and not a fundamental advantage gained with feedback when communicating over asymmetric channels. In order to lower bound the error probability for codes with feedback, Haroutunian's bound calculates the probability that the 'true' channel W behaves like a 'test' channel V with a capacity less than the rate (which automatically induces a non-negligible error probability by the strong converse). The exponent of this probability is the conditional divergence $D(V||W|P)$ where P is the average input distribution under channel V during the error event. It is quite difficult to prove anything useful about this input distribution, so a worst-case bound is taken, hence the inner max in (3.1).

In other words, the bound of (3.1) essentially assumes that because the code has feedback, it can somehow noncausally realize that the channel is behaving like V and use the exponent maximizing letter repeatedly. Intuition strongly suggests that, because the channel is assumed to be memoryless, a code could not tell in advance that future channel behavior will not be able to support the desired rate and pursue this strategy successfully. In fact, one might conjecture that the following exponent should be an upper bound to the error exponent for block codes with feedback,

$$\widetilde{E}(R) = \min_{V:C(V)\leq R} \max_{P:I(P,W)\geq R} D(V||W|P).$$

The idea here is that for an optimal code to be reliable over W, the input distribution under other channels must still support a rate that is high enough if the channel were W because the code cannot 'predict' that the memoryless channel will behave like V and not W.

This new exponent $\widetilde{E}(R)$ is easily seen to be sandwiched between $E_h(R)$ and $E_{sp}(R)$, and therefore is equal to both for symmetric channels. However, for general asymmetric channels, $E_{sp}(R) < \widetilde{E}(R) < E_h(R)$ (as shown in Figure 3.1 for the Z-channel). The improvement in the new exponent over Haroutunian is pronounced at rates near capacity (as shown in Figure 3.2). This is because any distribution that has a mutual information $I(P,W)$ near capacity must be 'close' to a capacity achieving distribution, and hence far away from the degenerate distribution that achieves the inner maximization in $E_h(R)$.

Perhaps even more interesting though, is that the ratio of $\widetilde{E}(R)$ to $E_{sp}(R)$ approaches 1 as R tends to capacity, at least for the Z-channel (as shown in Figure 3.3), while the ratio of $E_h(R)$ to $E_{sp}(R)$ approaches a constant greater than 2 as R tends to capacity for the Z-channel. While this fact may not seem interesting in itself, it has an interesting connection to another view of the rate-reliability tradeoff for block codes. As discussed in the introduction of Chapter 2, there is a connection between the second-order derivative of the error exponent at capacity and the other view of the rate-reliability tradeoff studied recently by Polyanskiy ([11], [29]) through the channel dispersion. The fact that $\widetilde{E}(R)$ and $E_{sp}(R)$ have the same second-order derivatives at capacity (at least for the Z-channel) suggests that if $\widetilde{E}(R)$ can be shown to be an upper bound to the error exponent for block

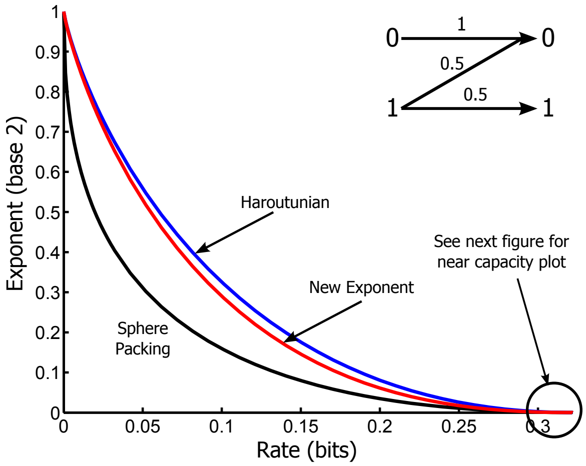

Figure 3.1: A plot of $E_{sp}(R)$ (sphere-packing), $E_h(R)$ (Haroutunian) and $\widetilde{E}(R)$ (the new exponent) for a Z-channel with crossover probability 1/2. The new exponent is sandwiched in between the sphere-packing and Haroutunian exponents.

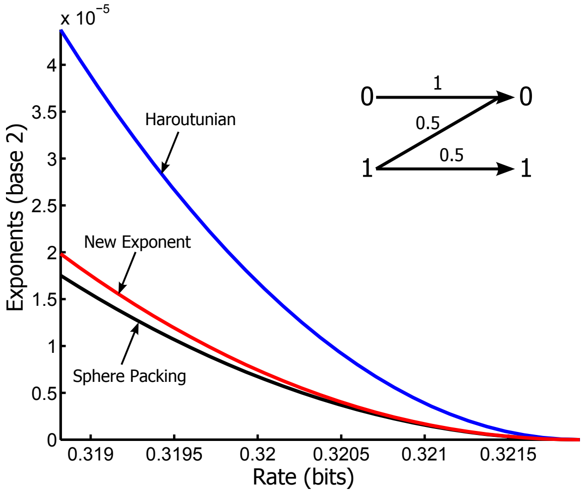

Figure 3.2: A plot of $E_{sp}(R)$, $E_h(R)$ and $\widetilde{E}(R)$ for the Z-channel with crossover probability 1/2 for rates near capacity. The new exponent seems to approximate the sphere-packing exponent much better that the Haroutunian exponent near capacity.

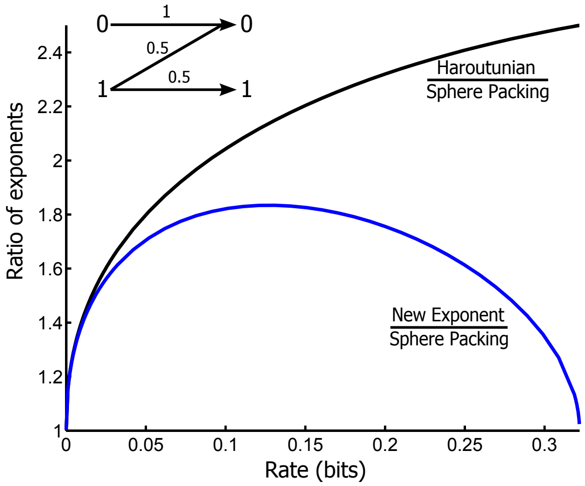

Figure 3.3: A plot of $E_h(R)/E_{sp}(R)$ compared to $\widetilde{E}(R)/E_{sp}(R)$. As the rate approaches capacity, the ratio of the new exponent to the sphere-packing exponent approaches 1 for the Z-channel, while the ratio of Haroutunian to sphere-packing approaches a value greater than 2 for the Z-channel. This is equivalent to the new exponent's second-order derivative at capacity being equal to the sphere-packing exponent's second-order derivative at capacity.

codes with feedback, then feedback does not significantly improve the maximum achievable rate for a given error probability and block length for all channels W. Such a conclusion can only be drawn for symmetric channels such as the BSC and BEC today [11].

Regrettably, we have not been able to prove that $\widetilde{E}(R)$ is an upper bound to $E(R)$ for block codes with feedback precisely because it is difficult to get a handle on the input distribution when the channel W behaves like an arbitrary test channel V. The Haroutunian exponent, however, also appears as an upper bound to the error exponent in the context of fixed delay codes [37] and neighborhood decoding of message-passing decoders [38]. The common tie between all these problems is that the input distribution is not known for the test channel V when the error event is happening. Therefore, a change-of-measure argument using V as the test channel ends up making a worst-case assumption on the input distribution.

3.2 Fixed delay coding and anytime codes

In order to make further progress in understanding asymmetric channels, we turn our attention to fixed delay codes. In many communication scenarios, data arrives in a steady stream at a transmitter to be sent to a receiver and the communication process must begin before the transmitter knows all the data to be sent to the receiver. One reason for this might be that the message is actually composed of many individual submessages, each of which must

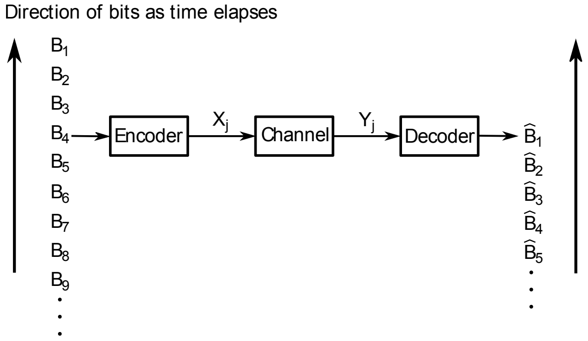

Figure 3.4: A fixed delay code differs from a block code in that bits stream into the encoder causally as time elapses. The decoder must produce an estimate of each bit within a fixed delay of when the bit arrives at the encoder.

be received within a fixed delay, e.g. interactive voice communication. A fixed delay code (as shown in Figure 3.4) can be modelled as taking an infinite stream of message bits that are causally revealed (at a uniform rate) to the encoder and communicating them across a channel, with each bit being decoded correctly with high probability within some fixed delay of arriving at the encoder.

In [37], the Haroutunian exponent is shown to be an upper bound for the error exponent of fixed delay codes (with respect to delay d). This is done by showing that for a test channel V with capacity lower than the rate, some bit must have non-vanishing error. An argument in [37] shows that the channel need only behave like V for the d uses between when the bit arrives at the encoder and when it is decoded for the error to occur with the same probability. The exponent with which this error happens is $D(V||W|P)$, where P is the input distribution during the error event for the d symbols before the bit's deadline. Again, we run into the issue of not being able to say anything about the input distribution P during this time span because only the bit with non-vanishing error probability is both revealed and decoded within this window. For the purposes of the proof, the code can pretend it has already decoded all previous bits before this bit arrives and it can pretend that it will reliably decode all bits arriving after this one after the deadline for this bit. Therefore using only the Haroutunian optimizing input symbol over the d symbols only adversely affects the error probability of this one bit (at least we cannot prove this to be untrue).

Near the completion of this thesis, the result of Section 2.8 was derived, showing that the Haroutunian exponent for an L-wise parallel channel constructed from using W L times independently at rate LR is at most $LE_{sp}(R - O(\log L/L); W)$, where $E_{sp}(R; W)$ is the sphere-packing exponent of W at rate R. Thus, by grouping uses of the channel together into large blocks, the error exponent one gets (after properly normalizing) approaches the sphere-packing exponent. This fact can be used to obtain the main result of this chapter in

Section 3.5: if $E_{delay}(R)$ denotes the error exponent for fixed delay coding without feedback,

$$E_{delay}(R) \leq E_{sp}(R).$$

This result shows that there is nothing special about asymmetric channels, especially without feedback, that allows codes used over them to predict when they will commit errors and 'give up' in that event. However, the proof does not give a sense of the true nature of the communication process in fixed delay codes and how individual bits are forced to share channel resources in any d-length decoding window. For this reason, we also will present a result for a more restricted class of fixed delay codes that will be more intuitive.

While we cannot prove that non-trivial communication is happening over the length d decoding window, there is an extended notion of fixed-length codes that this chapter will also focus on, called *anytime codes*, for which this property[1] will be shown to hold (an example of an anytime code is shown in Figure 3.5). Anytime codes differ from fixed delay codes in that at 'any time', the decoder produces estimates of *all* message bits that have arrived at the encoder at that point. The natural fixed delay codes: tree codes and convolutional codes, all have the anytime property that is needed here. As time elapses, the delay between when a bit arrives at the encoder and the current time also becomes large, and we evaluate the probability of error with respect to each delay for each bit[2]. The error exponent for an anytime code is then the error exponent with respect to delay. Note that for any delay d, an anytime code can be made into a delay-d fixed delay code by only providing the estimate for a bit when exactly d channel uses have elapsed since the bit arrived at the encoder. Therefore, the Haroutunian exponent is also an upper bound to the error exponent for anytime codes without feedback.

As just mentioned, good anytime codes enforce the requirement that nontrivial communication is happening within any (large) given window of channel uses. If it were not, some bit would arrive within the window and a smaller (than the window) delay estimate of that bit would be too unreliable. This anytime property has the consequence of forcing the input distribution during the large window to be nondegenerate. In fact, because reliable communication over W must happen for moderate delays, this implies that the input distribution over large delays must have mutual information over channel W of at least the rate. This intuition can be formalized to give the second result of this chapter, which is that if $E_{any}(R)$ denotes the optimal error exponent for anytime codes without feedback,

$$E_{any}(R) \leq \widetilde{E}(R).$$

This result is strictly weaker than the result that shows that $E_{delay}(R) \leq E_{sp}(R)$, but it was obtained first and is more intuitively understood.

[1] In [39], the authors assume a similar 'steady progress in communication' condition for a different error exponent problem.

[2] In a fixed delay code, once a bit's deadline has passed, there is no reason to ensure that the bit's probability of error continues to get lower.

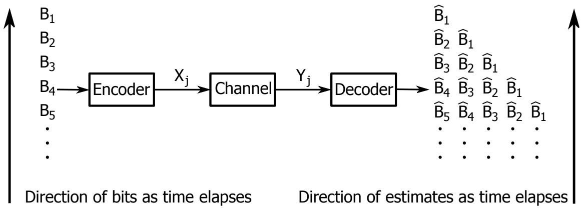

Figure 3.5: An anytime code with a universal delay decoder. Message bits arrive in a steady stream at the encoder. At each time, the decoder outputs estimates of all bits that have arrived at the encoder. The longer the delay between when a bit arrives at the encoder and the current time, the better the estimate of that bit is expected to be. In particular, we are interested in codes that achieve an exponentially decaying error probability with delay.

3.3 Control and anytime coding

Another important reason to study anytime codes is that they turn out to be of fundamental importance in the study of control and estimation over noisy channels for particular metrics[3] on the estimation or control error [40]. Rather than going into an in-depth exposition of control and its relation to anytime communication, which would vastly overstate the contribution of the result in this chapter to control theory, we will describe the specific control problem this result has application to. For those interested in the broader context of control and where anytime coding fits in it, see [40].

Consider the problem of controlling a scalar, linear unstable plant over a noisy DMC W, as shown in Figure 3.6. The plant state starts at $Z_0 = 0$ and updates at integer times t as

$$Z_{t+1} = \lambda Z_t + U_t + S_t \tag{3.2}$$

$$S_t \in \left[-\frac{\Omega}{2}, \frac{\Omega}{2}\right]. \tag{3.3}$$

The state evolution depends on the unstable eigenvalue of the plant, $|\lambda| > 1$, real-valued control inputs $\{U_t\}$ and real-valued noise process $\{S_t\}$, which is assumed to be bounded by $|S_t| \leq \Omega/2$ almost surely. The state observer and controller are separated, but the observer may communicate to the controller through a DMC W, with one channel use per unit time. At each time t, the observer inputs a channel symbol $X_t \in \mathcal{X}$ based on the current as well as past plant states $\{Z_k\}_{k\leq t}$. The controller receives $Y_t \in \mathcal{Y}$ with probability $W(Y_t|X_t)$ and chooses the control input U_t based on the current as well as past channel outputs $\{Y_k\}_{k\leq t}$. The plant state then evolves according to (3.2).

[3] For control of unstable systems under a bounded moment condition on the error, anytime capacity is the relevant quantity, while for other requirements such as asymptotically almost sure controllability, first-order notions such as capacity turn out to be sufficient. Even for these other asymptotic controllability requirements, however, anytime decoding turns out to be required, but the probability of error can decrease to 0 subexponentially [40].

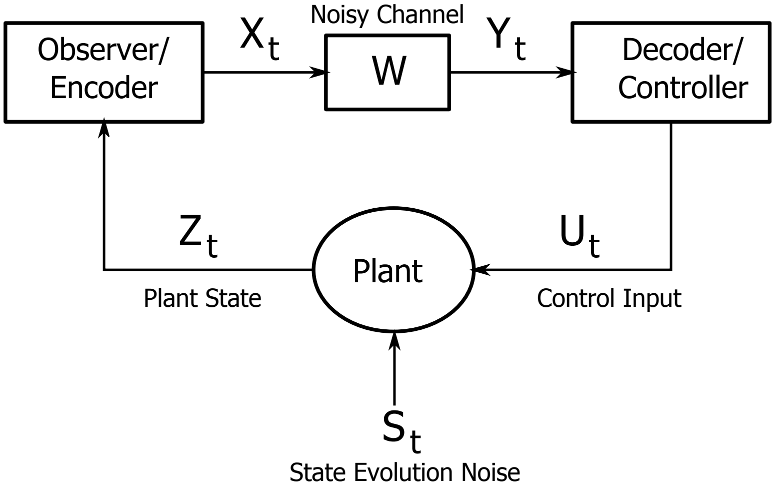

Figure 3.6: Control of an unstable plant over a noisy channel. The observer and controller of the plant are separated, but have the use of a noisy communication medium. The observer must also then perform some channel encoding to communicate information about the plant state and the controller must decode that information to decide its control input.

For any $\eta > 0$, we say that the channel W is sufficient to η-stabilize the plant if there exists an observer and controller pair and $K < \infty$ such that

$$\sup_{t \geq 1} \mathbb{E}\left[|Z_t|^{\eta}\right] \leq K$$

for every noise process $\{S_t\}_{t \geq 0}$ that obeys the bound in (3.3). According to [40], if there is an $R > \log \lambda$ for which the anytime error exponent without feedback is at least $\eta \log \lambda$, i.e., $E_{any}(R) > \eta \log \lambda$, then W is sufficient to η-stabilize the plant. This sufficiency characterization, however, is not the fundamental relationship between control of (3.2) and anytime communication. Rather, the true relationship is between control of (3.2) and anytime communication with noiseless feedback, as shown in Figure 3.7. The main results of [40] show that a channel W with perfect output feedback is sufficient to η-stabilize the plant if and only if there is a rate $R \geq \log \lambda$ such that the anytime error exponent with feedback is at least $\eta \log \lambda$, i.e., $E_{any,fb}(R) > \eta \log \lambda$.

The result of this chapter says only that $E_{any}(R) \leq \widetilde{E}(R)$, so it does not necessarily rule out that a channel is sufficient to η-stabilize the plant without feedback. It merely says when a channel W without feedback is not automatically sufficient to η-stabilize the plant without feedback.

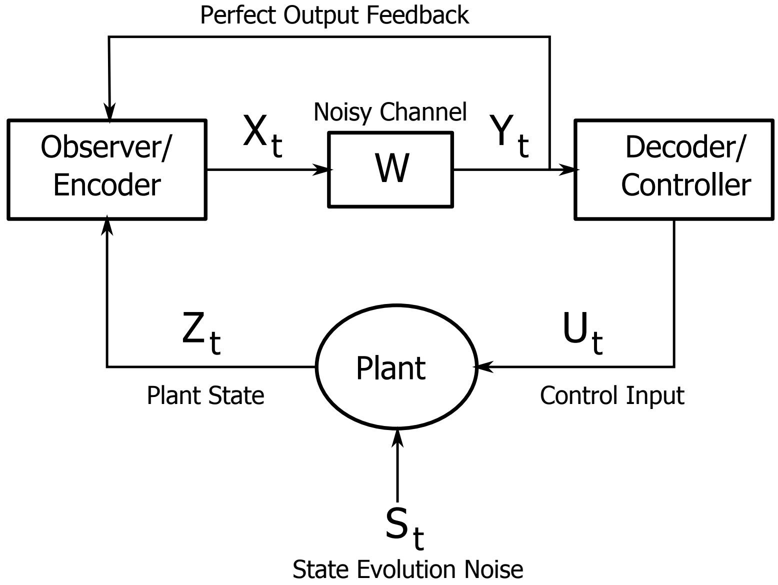

Figure 3.7: Control of an unstable plant over a noisy channel W. In addition to the noisy channel from the observer to the controller, there is a feedback path from the controller to the observer that carries the channel output symbols noiselessly and without delay.

3.4 Problem setup

Some of the notation used in Chapter 2 will continue to be used. Please refer to Tables 2.1 and 2.2 if some notation seems unfamiliar and it does not appear in this section.

We have an infinite sequence of IID equiprobable bits $\{B_i\}_{i=1}^{\infty}$ that will be communicated over a noisy channel. The bits are revealed to the encoder in a steady stream, at rate R bits per channel use (one channel use per unit of time). At time j, the encoder has access to[4]

$$B^{\lfloor jR \rfloor} = (B_1, B_2, \ldots, B_{\lfloor jR \rfloor}).$$

The communication medium is a discrete memoryless channel (DMC) induced by a probability transition matrix W from a finite input alphabet $\mathcal{X}$ to a finite output alphabet $\mathcal{Y}$. If the input to the channel at time j is $X_j = x \in \mathcal{X}$, the probability that the output at time j is $Y_j = y \in \mathcal{Y}$ is

$$\mathbb{P}_W(Y_j = y | X_j = x) = W(y|x).$$

A rate-R streaming encoder without feedback is a sequence of encoders $\mathcal{E} = \{\mathcal{E}_t\}_{t=1}^{\infty}$ with

$$\mathcal{E}_t : \{0,1\}^{\lfloor tR \rfloor} \to \mathcal{X}.$$

[4] We will use capital letters to denote random variables and lower case letters to denote realizations of those random variables. Superscripts and subscripts will be used to denote vectors, for example $x_i^j = (x_i, x_{i+1}, \ldots, x_j)$. For the special case of $i = 1$, we drop the subscript, i.e., $x^j = (x_1, \ldots, x_j)$.

At time j, the input to the channel with encoder $\mathcal{E}$ is $X_j = \mathcal{E}_j(B^{\lfloor jR \rfloor})$.

3.4.1 Fixed delay codes

A rate-R, delay-d fixed delay code is a pair $\mathcal{C} = (\mathcal{E}, \mathcal{D})$ where $\mathcal{E}$ is a rate-R streaming encoder and $\mathcal{D}$ is a rate-R delay-d decoder. The decoder is a sequence of maps $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^{\infty}$ with

$$\mathcal{D}_i : \mathcal{Y}^{\lceil i/R \rceil + d - 1} \to \{0, 1\}.$$

The fixed delay code's estimates of the bits are $\widehat{B}_i(d) = \mathcal{D}_i(Y^{\lceil i/R \rceil + d - 1})$. Let $\mathcal{C}_{R,d}$ denote the set of rate-R, delay-d fixed delay codes $\mathcal{C} = (\mathcal{E}, \mathcal{D})$. Define the error probability of $\mathcal{C} \in \mathcal{C}_{R,d}$ to be

$$P_e(d, \mathcal{C}) = \sup_i \mathbb{P}_W(\widehat{B}_i(d) \neq B_i).$$

The error exponent for fixed delay codes without feedback is defined to be

$$E_{delay}(R) = \limsup_{d \to \infty} -\frac{1}{d} \log \left[\inf_{\mathcal{C} \in \mathcal{C}_{R,d}} P_e(d, C) \right].$$

It was shown by Sahai [37] that

$$E_{delay}(R) \leq E_h(R) = \min_{V : C(V) \leq R} \max_P D(V || W | P).$$

By constructing random tree-codes with maximum-likelihood decoding, it can be shown that the random-coding exponent is achievable, so

$$E_{delay}(R) \geq E_r(R).$$

3.4.2 Anytime codes

A rate-R anytime decoder $\mathcal{D}$ is a set of maps $\mathcal{D} = \{\mathcal{D}_{i,d}\}_{i \geq 1, d \geq 0}$, with

$$\mathcal{D}_{i,d} : \mathcal{Y}^{\lceil i/R \rceil + d - 1} \to \{0, 1\}.$$

$\mathcal{D}_{i,d}$ is called the delay-d decoder for bit i since d channel uses after bit i arrives at the encoder are allowed before decoding bit i. We let $\widehat{B}_i(d) = \mathcal{D}_{i,d}(Y^{\lceil i/R \rceil + d - 1})$. A rate-$R$ anytime (universal delay) code without feedback is a pair $\mathcal{C} = (\mathcal{E}, \mathcal{D})$, where $\mathcal{E}$ is a rate-R streaming encoder and $\mathcal{D}$ is a rate-R anytime decoder. Let $\mathcal{C}_R$ denote the set of all possible rate-R

anytime codes without feedback. For a fixed anytime code $\mathcal{C}$, define the error performance metrics[5]

$$\begin{aligned}P_{e,i}(d,\mathcal{C}) &= \mathbb{P}_W\left(\widehat{B}_i(d) \neq B_i\right)\\ P_e(d,\mathcal{C}) &= \sup_i P_{e,i}(d,\mathcal{C})\\ E(\mathcal{C}) &\triangleq \liminf_{d\to\infty} -\frac{1}{d}\log P_e(d,\mathcal{C}).\end{aligned}$$

The anytime error exponent without feedback, at rate R, is defined to be

$$E_{any}(R) \triangleq \sup_{\mathcal{C}\in\mathcal{C}_R} E(\mathcal{C}).$$

A fixed delay code can be created by an anytime code for any delay d by just using the delay d decoder from the anytime decoder, so it immediately follows that

$$E_{any}(R) \leq E_{delay}(R) \leq E_h(R).$$

3.5 Tightening to sphere-packing for fixed delay codes

In order to show that the sphere-packing exponent upper bounds the error exponent for fixed delay codes, we will use the result proved by Sahai along with the parallel channel lemma about the Haroutunian exponent (Lemma 7 of Section 2.8).

Theorem 6 (Sahai [37], Theorem 3.1)**.** *For all $\delta > 0$ and rates $R > 0$, there is a finite $d_h(\delta, R, W)$ such that any fixed delay code $\mathcal{C}$ of rate R and delay $d \geq d_h(\delta, R, W)$ has*

$$-\frac{1}{d}\log P_e(d,\mathcal{C}) \leq E_h(R-\delta; W) + \delta, \tag{3.4}$$

where $E_h(R;W)$ is the Haroutunian exponent for channel W evaluated at rate R.

Theorem 7. *For all $\delta > 0$, $R > 0$, there is a finite $d_{sp}(\delta, R, W)$ such that for any fixed delay code $\mathcal{C}$ of rate R and delay $d \geq d_{sp}(\delta, R, W)$,*

$$-\frac{1}{d}\log P_e(d,\mathcal{C}) \leq E_{sp}(R-\delta; W) + \delta,$$

where $E_{sp}(R;W)$ is the sphere-packing exponent for channel W evaluated at rate R.

[5] The sup in the definition of $P_e(d,\mathcal{C})$ can be replaced by lim sup without affecting the result in this chapter.

Proof: Consider a large L and $W^{(L)}$, the parallel channel of W used L times independently within one symbol (as in Lemma 7), with input alphabet $\mathcal{X}^L$ and output alphabet $\mathcal{Y}^L$. Let $\mathcal{C}'_{LR,d'}$ be the set of rate LR, delay d' fixed delay codes for $W^{(L)}$. By Theorem 6 applied to $W^{(L)}$, there is a finite $d_h(\delta, LR, W^{(L)})$ such that if $d' \geq d_h(\delta, LR, W^{(L)})$,

$$-\frac{1}{d'} \log P_e(d', \mathcal{C}') \leq E_h(LR - \delta; W^{(L)}) + \delta$$

for any $\mathcal{C}' \in \mathcal{C}'_{LR,d'}$. Let us construct a code $\mathcal{C}'_{LR,d'}$ for some d' by using a code in $\mathcal{C}_{R,d}$. For some d large to be specified later, let

$$d' = \left\lceil \frac{d}{L} \right\rceil + 1.$$

We will use t to denote the symbol number for $\mathcal{C}$ (in terms of symbols from $\mathcal{X}$) and s to denote the symbol number for $\mathcal{C}'$ (in terms of supersymbols from $\mathcal{X}^L$). Let $X'_s = (X_{(s-1)L+1}, \ldots, X_{sL})$ and $Y'_s = (Y_{(s-1)L+1}, \ldots, Y_{sL})$ for $s \geq 1$. Consider the encoder for $\mathcal{C}'$ one gets by using the encoder from $\mathcal{C}$. That is, $\mathcal{E}' = \{\mathcal{E}'_s\}_{s=1}^{\infty}$ with

$$\mathcal{E}'_s : \{0,1\}^{\lfloor sLR \rfloor} \to \mathcal{X}^L$$

where

$$\mathcal{E}'_s(B^{\lfloor sLR \rfloor}) = \left(\mathcal{E}_{(s-1)L+1}\left(B^{\lfloor((s-1)L+1)R\rfloor}\right), \ldots, \mathcal{E}_{sL}\left(B^{\lfloor sLR \rfloor}\right)\right).$$

That is, $\mathcal{E}'$ uses the encoder $\mathcal{E}$ by giving the encoder knowledge of the bits in blocks of time of length L. Now, we want to use the delay d decoder from $\mathcal{C}$ to construct a delay d' decoder for $\mathcal{C}'$. In order to use this decoder, we must ensure that $\mathcal{C}'$ has the necessary channel outputs to run $\mathcal{D}$. Bit i arrives at the encoder $\mathcal{E}'$ at time (in supersymbols) $\left\lceil \frac{i}{LR} \right\rceil$. Bit i arrives at encoder $\mathcal{E}$ at time (in regular symbols) $\lceil i/R \rceil$. It is decoded by $\mathcal{D}$ at time (in regular symbols) $\lceil i/R \rceil + d - 1$. We need to ensure that if $d' = \lceil d/L \rceil + 1$, the supersymbol $Y'_{\lceil i/LR \rceil + d' - 1}$ contains $Y_{\lceil i/R \rceil + d - 1}$. This is guaranteed if

$$L\left(\left\lceil \frac{i}{LR} \right\rceil + \left\lceil \frac{d}{L} \right\rceil\right) \geq \left\lceil \frac{i}{R} \right\rceil + d - 1$$

because the last regular symbol in Y'_s is Y_{sL}. Now, using the fact that $L\lceil x/L \rceil \geq x$ and $\lceil x \rceil \leq x + 1$ yields

$$\begin{aligned}
L\left(\left\lceil \frac{i}{LR} \right\rceil + \left\lceil \frac{d}{L} \right\rceil\right) &\geq L\left\lceil \frac{i}{LR} \right\rceil + d \\
&= L\left\lceil \frac{(i/R)}{L} \right\rceil + 1 + (d-1) \\
&\geq (i/R) + 1 + (d-1) \\
&\geq \left\lceil \frac{i}{R} \right\rceil + d - 1.
\end{aligned}$$

Therefore, simply by running the original rate-R delay-d code $\mathcal{C}$ over channel W, we get a code $\mathcal{C}'$ of rate LR and delay d' for channel $W^{(L)}$. The estimates for these codes are the same, so

$$P_e(d, \mathcal{C}) = P_e(d', \mathcal{C}').$$

Now, if $d' \geq d_h(\delta, LR, W^{(L)})$, it follows by Theorem 6 that

$$\begin{aligned} -\frac{1}{d'} \log P_e(d, \mathcal{C}) &= -\frac{1}{d'} \log P_e(d', \mathcal{C}') \\ &\leq E_h\left(LR - \delta; W^{(L)}\right) + \delta. \end{aligned}$$

Therefore, using Lemma 7 to bound the Haroutunian exponent of the parallel channel,

$$\begin{aligned} -\frac{1}{\left\lceil \frac{d}{L} \right\rceil + 1} \log P_e(d, \mathcal{C}) &\leq E_h\left(LR - \delta; W^{(L)}\right) + \delta \\ &\leq LE_{sp}\left(R - \frac{\delta}{L} - \frac{|\mathcal{X}|}{L} \log(L+1); W\right) + \delta. \end{aligned}$$

Rescaling to d yields

$$\begin{aligned} -\frac{1}{d} \log P_e(d, \mathcal{C}) &\leq \frac{L(\lceil d/L \rceil + 1)}{d} E_{sp}\left(R - \frac{\delta}{L} - \frac{|\mathcal{X}|}{L} \log(L+1); W\right) + \delta\left(\frac{\lceil d/L \rceil + 1}{d}\right) \\ &\leq E_{sp}\left(R - \frac{\delta}{L} - \frac{|\mathcal{X}|}{L} \log(L+1); W\right) + \\ &\qquad \frac{2L}{d} E_{sp}\left(R - \frac{\delta}{L} - \frac{|\mathcal{X}|}{L} \log(L+1); W\right) + \delta\left(\frac{1}{L} + \frac{2}{d}\right). \end{aligned}$$

By first making L large enough, then δ small enough, followed by d large enough, and at least as large as $d_h(\delta, LR, W^{(L)})$, it follows that for all $\epsilon > 0$, there is a finite $d_{sp}(\epsilon, R, W)$ such that

$$-\frac{1}{d} \log P_e(d, \mathcal{C}) \leq E_{sp}(R - \epsilon; W) + \epsilon$$

for every $\mathcal{C} \in \mathcal{C}_{R,d}$ with $d \geq d_{sp}(\epsilon, R, W)$.

3.6 A weaker result for anytime codes

This result, guided largely by intuition about anytime codes, was obtained earlier than the stronger result for fixed delay codes. It is included in the thesis for completeness and because it may be useful in think about the problem of fixed blocklength coding with feedback.

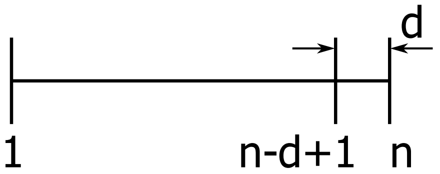

Figure 3.8: Constructing a block code of length n by using $\mathcal{C}$ for the first n channel uses. Only bits arriving at the encoder by time $n-d+1$ are expected to be decoded.

Theorem 8. *Fix a channel $W \in \mathcal{W}$. Then,*

$$\begin{aligned} E_{any}(R) &\leq \widetilde{E}(R) \\ &= \min_{V \in \mathcal{W}: C(V) \leq R} \max_{P \in \mathcal{P}: I(P,W) \geq R} D(V||W|P). \end{aligned} \tag{3.5}$$

The theorem also holds if common randomness is provided to the encoder and decoder. This is because the error probability bounds will hold for any fixed realization of the common randomness. To avoid cluttering notation, however, we will assume that common randomness is not present. In order to prove the theorem, we will show that for any $\mathcal{C} \in \mathcal{C}_R$, for any $\epsilon > 0$, $E(\mathcal{C}) \leq \widetilde{E}(R-\epsilon)$. Since $\epsilon > 0$ can be made arbitrarily small, we will show that the theorem holds by left continuity of the exponent $\widetilde{E}(R)$ in R. We will give here an outline of the proof, and make it rigorous in Appendix B.1.

First, fix an $\epsilon \in (0, R)$ and a $\mathcal{C} \in \mathcal{C}_R$. We will show that $E(\mathcal{C}) \leq \widetilde{E}(R-\epsilon)$. Let V denote an $\widetilde{E}(R-\epsilon)$ optimizing channel, that is

$$V \in \arg \min_{U \in \mathcal{W}: C(U) \leq R-\epsilon} \left\{ \max_{P: I(P,W) \geq R-\epsilon} D(U||W|P) \right\}.$$

We will show that

$$E(\mathcal{C}) \leq \max_{P: I(P,W) \geq R-\epsilon} D(V||W|P)$$

in several steps.

1. We will take two integer sequences, $d_l, \widetilde{d_l}$, going to infinity, with $d_l/\widetilde{d_l}$ approximately fixed (up to integer effects), $1 \ll \widetilde{d_l} \ll d_l$ in the limit. Now, fix a d and $\widetilde{d}$ (that is, drop the subscript l). First, we will construct a random block code of length $n \gg d$ by using the encoder $\mathcal{E}$ for the first n channel uses. The message bits will be those bits that have arrived at the encoder by time $n-d+1$. The block decoder will use the delay-d decoder for each of the message bits.

   If n is large enough, then the rate of this code is at least $R - \epsilon/2$ since the number of wasted bits at the encoder can be made arbitrarily small relative to the length of the block. Using arguments already made in [37], we can say that there are 'genie-aided' feedforward decoders, $\mathcal{D}^f$, that use only the past d received channel outputs as well as the true message bits up to the one being decoded that perform at least as well as $\mathcal{D}$. Let this feedforward decoder be denoted $\mathcal{D}^f_{i,d}$ for each i, d of interest and, more concretely, they are maps

$$\mathcal{D}^f_{i,d} : \{0,1\}^{i-1} \times \mathcal{Y}^d \to \{0,1\}$$
$$\widehat{B}^f_i(d) = \mathcal{D}^f_{i,d}\left(B^{i-1}, Y_{\lceil i/R \rceil}^{\lceil i/R \rceil + d - 1}\right).$$

These maps are as good as $\mathcal{D}$ for the channel W, namely for all i, d,

$$\mathbb{P}_W\left(\widehat{B}^f_i(d) \neq B_i\right) \leq \mathbb{P}_W\left(\widehat{B}_i(d) \neq B_i\right).$$

Now, the channel V has capacity at most $R - \epsilon$ and the rate of the code is at least $R - \epsilon/2$, so we should expect errors in some of the bits when the channel is V. Then, [37] shows that under channel V, if we let $\widetilde{B}^f_i(d) = B_i \oplus \widehat{B}^f_i(d)$ be the errors made by these genie-aided feedforward decoders,

$$\sum_{i=1}^{\lfloor (n-d+1)R \rfloor} H(\widetilde{B}^f_i(d)) \geq n\frac{\epsilon}{2}.$$

Therefore, since entropy is non-negative, there is at least one i such that $H(\widetilde{B}^f_i(d)) \geq \epsilon/2R$. For this i,

$$\mathbb{P}_V\left(\widehat{B}^f_i(d) \neq B_i\right) \geq h_b^{-1}\left(\frac{\epsilon}{2R}\right), \tag{3.6}$$

where h_b^{-1} is the inverse of the binary entropy function. We now note that in order for the channel W to behave like V and induce an error when the decoder is $\mathcal{D}^f_{i,d}$, it only needs to behave like V for the d channel uses after bit i is available to the encoder, as opposed to the full n channel uses that were needed to construct a block code of sufficient rate.

2. To show that the error probability of our actual decoder under W is not too small, we will study another block code's performance. Figure 3.9 shows the relevant times that we use $\mathcal{C}$ to construct this second block code. We wish to make a statement about the type of the input during the d channel uses that are of importance in (3.6). So, we use $\mathcal{C}$ to communicate over the channel W starting at time $\lceil i/R \rceil$ (the first channel use that bit i is available to the encoder) for a total of d channel uses. We have a fixed $\widetilde{d} < d$, so the message bits are

$$M = \left(B_i, B_{i+1}, \ldots, B_{\lfloor (\lceil i/R \rceil + d - \widetilde{d})R \rfloor}\right).$$

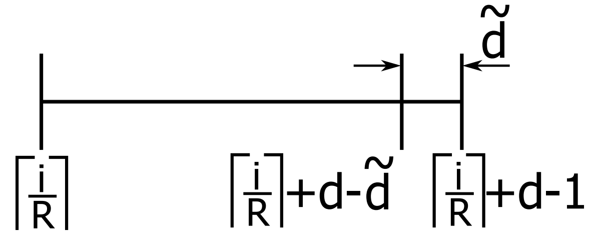
$\widetilde{d}$
$\lceil \frac{i}{R} \rceil$
$\lceil \frac{i}{R} \rceil + d - \widetilde{d}$
$\lceil \frac{i}{R} \rceil + d - 1$

Figure 3.9: A random block code constructed to show that the type of the input during the period when bit i's error event happens must have a large enough mutual information. The block code uses $\mathcal{C}$ starting from the time bit i arrives at the encoder for a total of d channel uses. The delay-$\widetilde{d}$ decoder for $\mathcal{C}$ is used to yield a code with low error probability for large enough $\widetilde{d}$. Randomness in the form of the bits before i and those that arrive after time $\lceil i/R \rceil + d - \widetilde{d}$ is given to the encoder.

The rate of this code, can be made arbitrarily close to R provided that $\widetilde{d}/d$ is made small enough. To decode the message, we use for each bit in the message, the delay-$\widetilde{d}$ decoder from $\mathcal{D}^f$. That is,

$$\widehat{M} = \left(\widehat{B}_i^f(\widetilde{d}), \ldots, \widehat{B}^f_{\lfloor (\lceil i/R \rceil + d - \widetilde{d})R \rfloor}(\widetilde{d}) \right).$$

Now, assume that $E(\mathcal{C}) \geq 2\widetilde{E}(R)/3$, as otherwise there is nothing to prove. Then, there is a $\widetilde{d}$ large enough so that the error probability of the block code, using the union bound over the bits, is upper bounded as

$$\begin{aligned}
\mathbb{P}_W\left(\widehat{M} \neq M\right) &\leq \left\lceil (d - \widetilde{d})R \right\rceil \exp\left(-\widetilde{d}\frac{\widetilde{E}(R)}{2}\right) \\
&\leq dR \exp\left(-\widetilde{d}\frac{\widetilde{E}(R)}{2}\right) \\
&= \exp\left(-d\left[\frac{\widetilde{d}}{2d}\widetilde{E}(R) - \frac{1}{d}\log dR\right]\right) \\
&\leq \exp\left(-d\left[\frac{\widetilde{d}}{4d}\widetilde{E}(R)\right]\right),
\end{aligned}$$

where the last statement holds for d large enough. Now, we have used the bits arriving before bit i and those arriving after time $\lceil i/R \rceil + d - \widetilde{d}$ as common randomness. Since the error probability is averaged over all values of this common randomness, Markov's inequality tells us that with high probability, any particular realization of the block code will have error probability upper bounded by

$$\exp\left(-d\left[\frac{\widetilde{d}}{8d}\widetilde{E}(R)\right]\right) \tag{3.7}$$

for large enough d. Now, provided that $\widetilde{d}$ and d are large enough, while the ratio $\widetilde{d}/d$ is small enough, we have that for most realizations of the block code of rate nearly R, the error probability is very small (crucially, exponential in d), as in Eqn. (3.7).

At this point, we prove a lemma that shows that if the error probability of a block code is exponentially decaying in the block length, the type of the input must have mutual information across the channel W high enough to support nearly the rate. Using this lemma, we can show that for our encoding system, for $\delta > 0$ small,

$$\mathbb{P}\Big(\text{ Type of } X_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1} = P, I(P,W) \geq R-\delta\Big) \simeq 1, \tag{3.8}$$

where the above means that the probability[6] tends to 1 with parameters $d, \widetilde{d}$ chosen properly. Therefore, the type of the input during the error event in (3.6) supports rate $R-\delta$ with high probability.

3. At this point, we perform a change-of-measure argument to get a lower bound on $P_{e,i}(d)$. We already know from the arguments in [37] that

$$\begin{aligned}\mathbb{P}_W\left(\widehat{B}_i(d) \neq B_i\right) &\geq \mathbb{P}_W\left(\widehat{B}_i^f(d) \neq B_i\right)\\ \mathbb{P}_V\left(\widehat{B}_i^f(d) \neq B_i\right) &\geq h_b^{-1}\left(\frac{\epsilon}{2R}\right).\end{aligned}$$

Using (3.8) and a straightforward change of measure argument, we can show that

$$\mathbb{P}_W\left(\widehat{B}_i^f(d) \neq B_i\right) \geq \frac{1}{4}h_b^{-1}\left(\frac{\epsilon}{2R}\right)^2 \exp\left(-d\left[\max_{P:I(P,W)\geq R-\delta} D(V||W|P) + O\left(\sqrt{\frac{\log d}{d}}\right)\right]\right).$$

Taking, limits as $d, \widetilde{d}$ get large and $\widetilde{d}/d$ gets small (less than $\epsilon/2$) yields that

$$E(\mathcal{C}) \leq \max_{P:I(P,W)\geq R-\epsilon} D(V|W|P),$$

which completes the proof.

[6] Note that since the code does not have feedback, the channel inputs depend only on the bits arriving at the encoder. Therefore, probability statements about the type of the input do not depend on the channel. For codes with feedback, probability statements about the input to the channel necessarily also depend on the channel, and hence this proof cannot be used to prove the same result for codes with feedback.

3.6.1 The relationship between $E_{sp}, \widetilde{E}$ and E_h for the Z-channel

We have the following properties, which were alluded to earlier, of $\widetilde{E}(R)$ when W is a Z-channel.

Proposition 6. *If W is a Z-channel with crossover probability $\delta \in (0,1)$,*

$$\lim_{R \to C_Z(\delta)} \frac{E_h(R)}{E_{sp}(R)} \geq \frac{1}{p^*(\delta)} \geq 2, \tag{3.9}$$

where $p^*(\delta)$ is the capacity achieving distribution's probability on 1 and $C_Z(\delta)$ is the capacity for W. Meanwhile,

$$\lim_{R \to C_Z(\delta)} \frac{\widetilde{E}(R)}{E_{sp}(R)} = 1.$$

Equivalently, since both $E_{sp}(R)$ and $\widetilde{E}(R)$ have zero first derivative at $C_Z(\delta)$,

$$\left. \frac{d^2 E_{sp}(R)}{dR^2} \right|_{R=C_Z(\delta)} = \left. \frac{d^2 \widetilde{E}(R)}{dR^2} \right|_{R=C_Z(\delta)}. \tag{3.10}$$

The proofs of these two results can be found in the appendix on Z-channels (Appendix B.2). We conjecture that (3.10) holds for any asymmetric DMC that has a unique capacity achieving distribution. The lower bound of 2 for the family of Z-channels in (3.9) does not extend to all asymmetric channels. One can see this by considering binary asymmetric channels as the crossover probabilities go to the same value.

3.7 Concluding remarks

To conclude, we would like to reflect on why $\widetilde{E}(R)$ seems to be a good approximation to $E_{sp}(R)$ for rates very near capacity (at least for the Z-channel), but is much closer to $E_h(R)$ for most rates. The improvement in $\widetilde{E}(R)$ from $E_h(R)$ is coming entirely from restricting the set of P in the inner maximization of (3.1) to those that satisfy $I(P, W) \geq R$. When R is very near capacity, the only P allowed are very close to the capacity achieving distribution (which usually are far from degenerate). However, because mutual information is concave-$\cap$ in P, the derivative (or gradient) around the capacity achieving distribution is 0, and hence any increase in the separation of R from capacity (at least initially) yields a large increase in the set of allowable P. This issue is seen clearly in the form of a plot for the Z-channel, as shown in Figure 3.10.

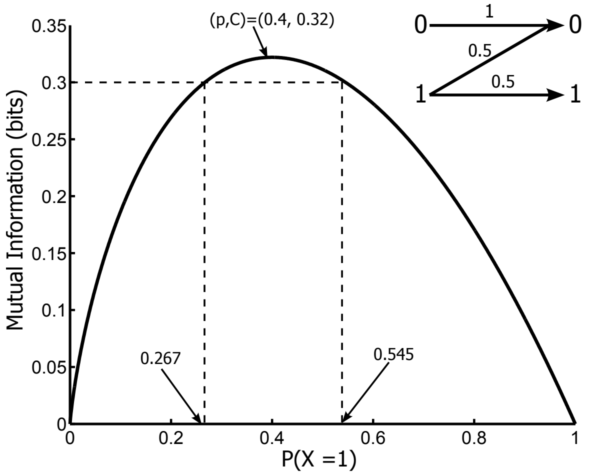

Figure 3.10: A plot of $I(P,W)$ when W is a Z-channel with crossover probability 1/2. The capacity achieving P has $P(X=1)=0.4$ and $C(W)=0.322$ bits per channel use. Even if $R=0.3$, the set of P such that $I(P,W)\geq R$ is already any P with $P(X=1)\in[0.267,0.545]$. This is the reason why $\widetilde{E}(R)$ is a much better bound than $E_h(R)$ only for rates very near capacity.

Chapter 4

Lossy compression of arbitrarily varying sources

4.1 Introduction

4.1.1 Motivation

The arbitrarily varying source (AVS) was introduced by Berger [4] as a source that samples other 'subsources' under the control of an agent called a switcher. The AVS was used in the model of an information-theoretic 'source coding game' between two players, the aforementioned switcher and a coder. The goal of the coder was to encode the output of the AVS to within a specified distortion, and the goal of the switcher was to make the coder use as large a rate to attain the specified distortion as possible. Berger studied the adversarial rate-distortion function (the rate the coder needs to achieve a target distortion regardless of switcher strategy) under certain rules for the switcher. Primarily, [4] gives the rate-distortion function when the switcher is not allowed to observe present or future subsource realizations.

The purpose of this chapter is to deepen understanding of the source coding game by looking into variations where the capabilities of the switcher are enhanced. In [4], Berger himself asks what happens to the rate-distortion function when the switcher is allowed to 'cheat' and observe present or future realizations of the subsources. In addition to tackling this question, we further study scenarios where the switcher receives noisy observations of the subsources or the switcher is not adversarial, but helpful.

As a motivation for studying the source coding game, Berger mentions that the results might have application to situations where multiple data streams are multiplexed into a single data stream. Another potential application is in the field of active sensing or active vision [41], a subfield of computer vision in which sensors actively explore their environment using information they have previously sensed.

4.1.2 Active sources and causality

Active vision/sensing/perception [41] is an approach to computer vision, the main principle of which is that sensors choose to explore their environment actively *based on what they currently sense or have previously sensed.* As Bajcsy says it in [41], “We do not just see, we look.” The contrast to passive sensors can be seen by comparing a fixed security camera (non-active) to a person holding a camera (active). Even if the person is otherwise stationary, they may zoom the camera into any part of their visual field to obtain a better view (e.g. if they see a trespasser). Perhaps more concretely, cameras autonomously operated by algorithms and principles developed in the field of computer vision fit into the category of ‘active sources’[1]. Another example of active sensing is a sequence of sensor-measurements that are dynamically sampled by a distributed sensor network. Yet another is the case of measurements taken by an autonomously moving sensor that chooses where to go in part based on what it is observing.

Suppose one were to design a system which required compressing the output of an active source. Standard rate-distortion theory [42] tells us that if we can model the source as a stochastic process with suitable probabilistic structure (e.g. stationary and ergodic), we can find its rate-distortion function and are guaranteed the existence of codes with rates that approach the information-theoretic limit. However, for many active sources, such probabilistic structure is difficult to establish because of the difficulty in precisely modelling the inner workings of the source. This difficulty may arise, for example, due to the active source making decisions using a complex algorithm that introduces memory in ways that are hard to quantify (and hence make the source nonergodic). Alternatively, this difficulty might occur because the entity controlling the active source has some degree of free will[2].

At this point, we might relax the assumption of a known distribution for the source’s output and turn to universal coding over some class of sources with limited structure (e.g. the class of m-th order Markov processes). Going further, we could give up hope of specifying a possible distribution for the source process and instead take an individual-sequence approach to compression. Here, we are guaranteed the existence of universal coding algorithms such as those based on the Lempel-Ziv algorithm ([43], [44]) that asymptotically approach the compression rate required for any particular individual sequence, as measured by the best one can do with (arbitrarily large) finite-state coding systems.

The individual sequences/universal coding framework gives an answer to the *algorithmic* question of how we might optimally encode and decode the source. In this paradigm, however, we cannot know in advance the actual number of bits per symbol needed to represent the source output until the time of encoding. This lack of *a priori* knowledge would be unacceptable to, e.g., the designer of a distributed control system in which rate must be pro-

[1] We use the terms active sensors, vision/video and sources interchangeably, but strictly speaking active sensors and active vision are types of active sources.

[2] As an extreme example, consider accurately probabilistically modelling the output stream of a camera operated by a human, even assuming a simple probabilistic model of the environment being filmed.

visioned. For example, if we were to provision enough rate for reconstructing all individual sequences with zero distortion, the required rate would be the log of the alphabet size. It seems that requiring every individual sequence be accounted for gives high model generality at the cost of excessive provisioning of rate.

In order to answer the question of how much rate to *provision* for an active source, while allowing for flexibility in the model of the source, we propose to model it as an arbitrarily varying source (AVS) and study its rate-distortion function. Bounds on the fixed-length block coding rate-distortion function of an AVS give usable estimates for the designer of a system wishing to know how much rate to provision for the transmission or storage of a compressed active source. The key strength of an AVS model is that it allows for a tradeoff between model generality and the corresponding tightness of the rate bounds for compression provided by the model. Two basic ways to model the goals of the source are worst case (adversarial) and helpful (joint optimization of active source and coding system), which respectively correspond to upper and lower bounds to the rate needed to compress the output of an active source.

We start with the notion that knowledge allows the switcher to use the switching mechanism to gain freedom from ambient distributions. In developing a suitable model for an active source as an AVS, we would like to have simple ways of restricting the knowledge of the switcher. This restriction should align with how the active source makes its decisions on sampling from the environment. Causality of the decision making is an important restriction, as there is the possibility that the source has noncausal information about the environment. For example, a cameraman at a sporting event generally has only causal knowledge of the environment. A cameraman on a movie set, however, has noncausal information about the environment through the script. The noncausal information can be advantageous to the cameraman in (actively) capturing the important features of a scene. Additionally, it may be that the entity controlling the source does not have perfect knowledge of the subsources, i.e., it observes the sources noisily. In general, the more power/knowledge given to the switcher, the larger the gap between the upper and lower bounds on the rate-distortion function. These various ways of modelling the switcher allow us to quantify this intuition through a tradeoff between model generality and potential conservatism in rate provisioning.

There are many issues to consider in the study of lossy compression for active sources, but we concentrate here entirely on the simplest aspect of the problem: what is the impact on the rate-distortion function of having the source being actively sampled by an entity that knows something about the realizations of the environment? Thus, we assume a simplified traditional rate-distortion setting with known finite alphabets and bounded distortion measures. The goal is the traditional block-coding one: meet an average distortion constraint using as little rate as possible. Admittedly, the most interesting practical problems involve subsources with memory, but following tradition[3] and for simplicity, we first focus in this

[3] See ([42], Section VII.) for an account of the slow pace that rate-distortion theory impacted practical compression applications during its first 25 years.

Switcher model	$R(D)$
Strictly causal (adversarial)	Berger [4]
Noncausal (adversarial)	Section 4.3
Noncausal noisy observations (adversarial)	Section 4.4
Noncausal (helpful)	Section 4.5

chapter on memoryless subsources to understand the basic differences between active and non-active sources for lossy compression.

4.1.3 Contributions

Intuitively, a strictly causal adversary switching amongst memoryless sources is no more threatening than a switcher that randomly switches. This intuition was proved correct in [4] by Berger as he determined the rate-distortion function for memoryless subsources and a strictly causal adversarial model. Section 4.3 gives the rate-distortion function for an AVS when the adversary has noncausal access to realizations of a finite collection of memoryless subsources and can sample among them. As shown in Theorem 9, the rate-distortion function for this problem is the maximization of the rate-distortion function over the IID sources the adversary can simulate. The adversary requires only causal information to impose this rate-distortion function. This establishes that when the subsources are memoryless, the rate-distortion function can strictly increase when the adversary has knowledge of the present subsource realizations, but no further increase occurs when the adversary is allowed knowledge of the future realizations.

In order to give more ways to restrict the knowledge of the switcher, we then extend the AVS model to include noisy or partial observations of the subsource realizations and determine the rate-distortion function for this setting in Section 4.4. As shown in Theorem 10, the form of the solution is the same as for the adversary with clean observations, with the set of attainable distributions essentially being related to the original distributions through Bayes' rule.

Next, Section 4.5 changes the perspective from the traditional adversarial setting to a cooperative setting. It explores the problem when the goal of the switcher is to help the coder achieve a low distortion. Theorem 11 gives a characterization of the rate-distortion functions if the helper has access to future realizations in terms of the rate-distortion function for an associated lossy compression problem. As a corollary, we also give bounds for the cases of causal observations and noisy observations. However, for most helpful switcher settings, a tight characterization of the rate-distortion function is lacking.

Simple examples illustrating these results are given in Section 4.6. In Section 4.7, we discuss how to compute the rate-distortion function for arbitrarily varying sources to within a given accuracy using the uniform continuity of the IID rate-distortion function. This task needs some discussion because of the fact that the IID rate-distortion function is generally nonconcave as a function of the distribution [45]. The main tool there is an explicit bound on the uniform continuity of the IID rate-distortion function that is of potentially independent interest, as we use it to quickly analyze the behavior (in probability) of a simple rate-distortion estimator for IID sources. Finally, we conclude in Section 4.8. The results of this chapter appear in [46], which is to be published in IEEE Transactions on Information Theory.

4.2 Problem Setup

4.2.1 Notation

Let $\mathcal{X}$ and $\widehat{\mathcal{X}}$ be the finite source and reconstruction alphabets respectively. Let $\mathbf{x}^n = (x_1, \ldots, x_n)$ denote a vector from $\mathcal{X}^n$ and $\widehat{\mathbf{x}}^n = (\widehat{x}_1, \ldots, \widehat{x}_n)$ a vector from $\widehat{\mathcal{X}}^n$. When needed, $\mathbf{x}^k = (x_1, \ldots, x_k)$ will be used to denote the first k symbols in the vector $\mathbf{x}^n$.

Let $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$ be a distortion measure on the product set $\mathcal{X} \times \widehat{\mathcal{X}}$ with maximum distortion $d^* < \infty$. Let

$$\widetilde{d} = \min_{(x,\widehat{x}):\ d(x,\widehat{x})>0} d(x, \widehat{x}) \tag{4.1}$$

be the minimum nonzero distortion. Define $d_n : \mathcal{X}^n \times \widehat{\mathcal{X}}^n \to [0, d^*]$ for $n \geq 1$ to be

$$d_n(\mathbf{x}^n, \widehat{\mathbf{x}}^n) = \frac{1}{n} \sum_{k=1}^{n} d(x_k, \widehat{x}_k).$$

Let $\mathcal{P}(\mathcal{X})$ be the set of probability distributions on $\mathcal{X}$, let $\mathcal{P}_n(\mathcal{X})$ be the set of types (see [21], [47]) of length-n strings from $\mathcal{X}$, and let $\mathcal{W}$ be the set of probability transition matrices from $\mathcal{X}$ to $\widehat{\mathcal{X}}$. Let $p_{\mathbf{x}^n} \in \mathcal{P}_n(\mathcal{X})$ be the empirical type of a vector $\mathbf{x}^n$. For a $p \in \mathcal{P}(\mathcal{X})$, let

$$D_{\min}(p) = \sum_{x \in \mathcal{X}} p(x) \min_{\widehat{x} \in \widehat{\mathcal{X}}} d(x, \widehat{x})$$

be the minimum average distortion achievable for the source distribution p. The (functional) IID rate-distortion function of $p \in \mathcal{P}(\mathcal{X})$ at distortion $D > D_{\min}(p)$ with respect to distortion measure d is defined to be

$$R(p, D) = \min_{W \in \mathcal{W}(p,D)} I(p, W),$$

where $\mathcal{W}(p, D)$ is a set of admissable probability transition matrices,

$$\mathcal{W}(p, D) = \left\{ W : \sum_{x \in \mathcal{X}} \sum_{\widehat{x} \in \widehat{\mathcal{X}}} p(x) W(\widehat{x}|x) d(x, \widehat{x}) \leq D \right\}$$

and $I(p, W)$ is the mutual information[4]

$$I(p, W) = \sum_{x \in \mathcal{X}} \sum_{\widehat{x} \in \widehat{\mathcal{X}}} p(x) W(\widehat{x}|x) \ln \left[\frac{W(\widehat{x}|x)}{(pW)(\widehat{x})} \right],$$

with $(pW)(\widehat{x}) = \sum_{x' \in \mathcal{X}} p(x') W(\widehat{x}|x')$. Let $\mathcal{B} = \{\widehat{\mathbf{x}}^n(1), \ldots, \widehat{\mathbf{x}}^n(K)\}$ be a codebook with K length-n vectors from $\widehat{\mathcal{X}}^n$. Define

$$d_n(\mathbf{x}^n; \mathcal{B}) = \min_{\widehat{\mathbf{x}}^n \in \mathcal{B}} d_n(\mathbf{x}^n, \widehat{\mathbf{x}}^n).$$

If $\mathcal{B}$ is used to represent an IID source with distribution p, then the average distortion of $\mathcal{B}$ is defined to be

$$d(p; \mathcal{B}) = \sum_{\mathbf{x}^n \in \mathcal{X}^n} d_n(\mathbf{x}^n; \mathcal{B}) \prod_{k=1}^{n} p(x_k) = \mathbb{E}_p[d_n(\mathbf{x}^n; \mathcal{B})].$$

For $n \geq 1$, $D > D_{\min}(p)$, let $K(n, D)$ be the minimum number of codewords needed in a codebook $\mathcal{B} \subset \widehat{\mathcal{X}}^n$ so that $d(p; \mathcal{B}) \leq D$. By convention, if no such codebook exists, $K(n, D) = \infty$. Let the (operational) rate-distortion function[5] of an IID source be $R(D) = \limsup_n \frac{1}{n} \ln K(n, D)$. Shannon's rate-distortion theorem ([48], [49]) states that for all n, $\frac{1}{n} \ln K(n, D) \geq R(p, D)$ and

$$\limsup_{n \to \infty} \frac{1}{n} \ln K(n, D) = R(D) = R(p, D).$$

4.2.2 Arbitrarily varying sources

As mentioned earlier, the AVS is a model of a source in the 'source coding game' introduced by Berger in [4]. The two players are called the 'switcher' and 'coder'. In a coding context, the coder corresponds to the designer of a lossy source code and the switcher corresponds to a potentially malicious adversary selecting symbols to be encoded.

Fig. 4.1 shows a model of an AVS. There are m IID 'subsources' with common alphabet $\mathcal{X}$. In [4], the subsources are assumed to be independent, but that restriction turns out not to be required[6]. There can also be multiple subsources governed by the same distribution. In that sense, the switcher has access to a *list* of m subsources, rather than a set of m different distributions. The marginal distributions of the m subsources are known to be $\{p_l\}_{l=1}^m$ and

[4] We use natural log, denoted ln, and nats in most of the chapter. In examples only, we use bits.

[5] We define $R(D_{\min}(p)) = \lim_{D \downarrow D_{\min}(p)} R(D)$. This is equivalent to saying that a sequence of codes represent a source to within distortion D if their average distortion is tending to D in the limit. The only distortion where this distinction is meaningful is $D_{\min}(p)$.

[6] In [4], the motivation was multiplexing data streams and independence is a reasonable assumption, but the proof does not require it.

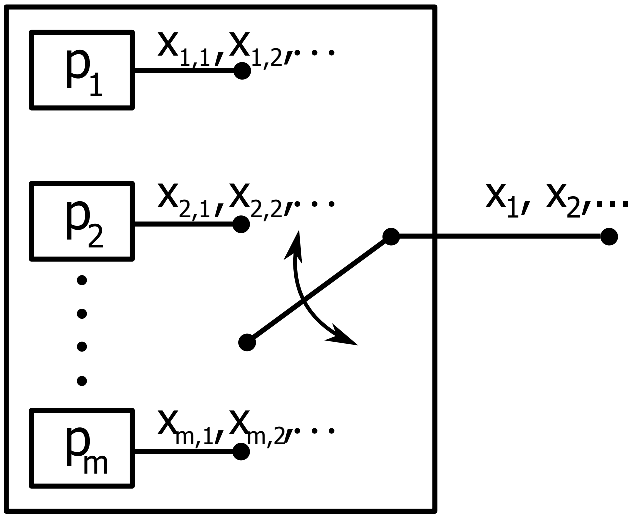

Figure 4.1: A class of models for an AVS. The switcher can set the switch position according to the rules of the model.

we let $\mathcal{G} = \{p_1, \ldots, p_m\}$. Let $P(x_{1,1}, \ldots, x_{m,1})$ be the joint probability distribution for the IID source $\{(x_{1,k}, \ldots, x_{m,k})\}_{k \geq 1}$. Fix an $n \geq 1$ and consider a block of length n. We let $x_{l,k}$ denote the output of the l^{th} subsource at time k. We will use $\mathbf{x}_l^n$ to denote the vector $(x_{l,1}, \ldots, x_{l,n})$. At each time k, the AVS outputs a letter x_k which is determined by the position of the switch inside the AVS. The switch positions are denoted $\mathbf{s}^n = (s_1, \ldots, s_n)$ with $s_k \in \{1, 2, \ldots, m\}$ for each $1 \leq k \leq n$. With this notation, $x_k = x_{s_k,k}$ for $1 \leq k \leq n$.

The switcher can set the switch position according to the rules for the AVS. In the next few sections, we will discuss different rules for the switcher, particularly different levels of causality in knowledge of the subsource realizations. The switcher may or may not have knowledge of the codebook, but this knowledge turns out to be inconsequential for the worst-case rate-distortion function.

The coder's goal is to design a codebook $\mathcal{B}$ of minimal size to represent $\mathbf{x}^n$ to within distortion D on average. The codebook must be able to do this for *every* allowable strategy for the switcher according to the model. Define

$$M(n, D) = \min \left\{ |\mathcal{B}| : \begin{array}{c} \mathbb{E}[d_n(\mathbf{x}^n; \mathcal{B})] \leq D \\ \text{for all allowable} \\ \text{switcher strategies} \end{array} \right\}.$$

Here, $\mathbb{E}[d_n(\mathbf{x}^n; \mathcal{B})]$ is defined to be $\sum_{\mathbf{x}^n} \left(\sum_{\mathbf{s}^n} P(\mathbf{s}^n, \mathbf{x}^n) \right) d_n(\mathbf{x}^n; \mathcal{B})$, where $P(\mathbf{s}^n, \mathbf{x}^n)$ is an appropriate probability mass function on $\{1, \ldots, m\}^n \times \mathcal{X}^n$ that agrees with the model of the AVS. We are interested in the exponential rate of growth of $M(n, D)$ with n, and so we define the rate-distortion function of an adversarial AVS to be

$$R(D) \triangleq \limsup_{n\to\infty} \frac{1}{n} \ln M(n, D).$$

In every case considered, it will also be clear that $R(D) = \liminf_{n\to\infty} \frac{1}{n} \ln M(n, D)$. For notational convenience, we only refer to the rate-distortion function as $R(D)$, removing its dependence on the subsource distributions as well as all the different cases of switcher power.

4.2.3 Literature Review

Before presenting the results of this chapter, let us consider some of the relevant results about rate-distortion for arbitrarily varying sources, starting with IID sources.

One IID source Suppose $m = 1$. Then there is only one IID subsource $p_1 = p$ and the switch position is determined to be $s_k = 1$ for all time. This is exactly the classical rate-distortion problem considered by Shannon [48], and he showed

$$R(D) = R(p, D).$$

Computing $R(p, D)$ can be done with the Blahut-Arimoto algorithm [47], and also falls under the umbrella of convex programming.

Compound source Now suppose that $m > 1$, but the switcher is constrained to choose $s_k = s \in \{1, \dots, m\}$ for all k. That is, the switch position is set once and remains constant afterwards. Sakrison [50] studied the rate-distortion function for this class of *compound* sources and showed that planning for the worst subsource is both necessary and sufficient. Hence, for compound sources,

$$R(D) = \max_{p\in\mathcal{G}} R(p, D).$$

Recall that $\mathcal{G} = \{p_1, \dots, p_m\}$ is the set of marginal distributions of the m subsources. This result holds whether the switch position is chosen with or without knowledge of the realizations of the m subsources. Here, $R(D)$ can be computed easily since m is finite and each individual $R(p, D)$ can be computed.

Strictly causal adversarial source In Berger's setup [4], the switcher is allowed to choose $s_k \in \{1, \dots, m\}$ arbitrarily at any time k , but must do so in a strictly causal manner without access to the current time step's subsource realizations. More specifically, the switch position s_k is chosen as a (possibly random) function of $(s_1, \dots, s_{k-1})$ and $(x_1, \dots, x_{k-1})$. The conclusion of [4] is that under these rules,

$$R(D) = \max_{p\in\overline{\mathcal{G}}} R(p, D), \tag{4.2}$$

where $\overline{\mathcal{G}}$ is the convex hull of $\mathcal{G}$. It should be noted that this same rate-distortion function applies in the following cases [4]:

- The switcher chooses s_k at each time k without *any* observations at all.
- The switcher chooses s_k as a function of the first $k-1$ outputs of *all* m subsources.

Note that in (4.2), evaluating $R(D)$ involves a maximization over an infinite set, so the computation of $R(D)$ is not trivial since $R(p, D)$ is not necessarily a concave-$\cap$ function. A simple, provable, approximate (to any given accuracy) solution is discussed in Section 4.7.

4.3 $R(D)$ for the cheating switcher

In the conclusion of [4], Berger poses the question of what happens to the rate-distortion function when the rules are tilted in favor of the switcher. Paraphrasing Berger:

> As another example, suppose the switcher is permitted to observe the candidates for x_k generated by each [subsource] before (randomly) selecting one of them. Then it can be shown that [$R(D)$ (except in certain special cases) strictly increases]. The determination of $R(D)$ under these rules appears to be a challenging task.

Suppose that the switcher were given access to the m subsource realizations before having to choose the switch positions; we call such a switcher a 'cheating switcher'. In this chapter, we deal with two levels of noncausality and show they are essentially the same when the subsources are IID over time:

- The switcher chooses s_k based on the realizations of the m subsources at time k. We refer to this case as **1-step lookahead** for the switcher.
- The switcher chooses $(s_1, \ldots, s_n)$ based on the entire length-n realizations of the m subsources. We refer to this case as **full lookahead** for the switcher.

Theorem 9. *Define the set of distributions*

$$\mathcal{C} = \left\{ p \;\; : \begin{array}{c} \sum_{x \in \mathcal{V}} p(x) \geq P\big(\forall\ l, x_l \in \mathcal{V}\big) \\ \forall\ \mathcal{V} \textit{ such that} \\ \mathcal{V} \subseteq \mathcal{X} \end{array} \right\}, \tag{4.3}$$

where the event $\{\forall\ l, x_l \in \mathcal{V}\}$ is shorthand for $\{(x_1, \ldots, x_m) : x_l \in \mathcal{V}, l = 1, \ldots, m\}$. Also, define

$$\widetilde{R}(D) \triangleq \max_{p \in \mathcal{C}} R(p, D).$$

For a general set of distributions $\mathcal{Q} \subset \mathcal{P}(\mathcal{X})$, let $D_{\min}(\mathcal{Q}) \triangleq \sup_{p \in \mathcal{Q}} D_{\min}(p)$. Suppose the switcher has either 1-step lookahead or full lookahead. In both cases, for $D > D_{\min}(\mathcal{C})$,

$$R(D) = \widetilde{R}(D)$$

For $D < D_{\min}(\mathcal{C})$, $R(D) = \infty$ by convention because the switcher can simulate a distribution for which the distortion D is infeasible for the coder.

Remarks:

- In non-degenerate cases, $\overline{\mathcal{G}}$ is a strict subset of $\mathcal{C}$, and thus $R(D)$ can strictly increase when the switcher is allowed to look at the present subsource realizations before choosing the switch position.
- As a consequence of the theorem, we see that when the subsources within an AVS are IID, knowledge of past subsource realizations is useless to the switcher, knowledge of the current step's subsource realizations is useful, and knowledge of future subsource realizations beyond the current step is useless if 1-step lookahead is already given.
- Note that computing $\widetilde{R}(D)$ requires the further discussion given in Section 4.7, just as it does for the strictly causal case of Berger.

Proof: We give a short outline of the proof here. See Appendix C.1 for the complete proof. To show $R(D) \leq \widetilde{R}(D)$, we use the type-covering lemma from [4]. It says for a fixed type p in $\mathcal{P}_n(\mathcal{X})$ and $\epsilon > 0$, all sequences with type p can be covered within distortion D with at most $\exp(n(R(p, D) + \epsilon))$ codewords for large enough n. Since there are at most $(n+1)^{|\mathcal{X}|}$ distinct types, we can cover all n-length strings with types in $\mathcal{C}$ with at most $\exp(n(\widetilde{R}(D) + \frac{|\mathcal{X}|}{n}\ln(n+1) + \epsilon))$ codewords. Furthermore, we can show that types not in $\mathcal{C}$ occur exponentially rarely even if the switcher has full lookahead, meaning that their contribution to the average distortion can be bounded by d^* times an exponentially decaying term in n. Hence, the rate needed regardless of the switcher strategy is at most $\widetilde{R}(D) + \epsilon$ with $\epsilon > 0$ arbitrarily small.

Now, to show $R(D) \geq \widetilde{R}(D)$, we describe one potential strategy for the adversary. This strategy requires only 1-step lookahead and it forces the coder to use rate at least $\widetilde{R}(D)$. For each subset $\mathcal{V} \subseteq \mathcal{X}$ with $\mathcal{V} \neq \emptyset$ and $|\mathcal{V}| \leq m$, the adversary has a random rule $f(\cdot|\mathcal{V})$, which is a probability mass function (PMF) on $\mathcal{V}$. At each time k, if the switcher observes a candidate set $\{x_{1,k}, \ldots, x_{m,k}\}$, the switcher chooses to output $x \in \{x_{1,k}, \ldots, x_{m,k}\}$ with probability $f(x|\{x_{1,k}, \ldots, x_{m,k}\})$. If $\beta(\mathcal{V}) = P(\{x_{1,k}, \ldots, x_{m,k}\} = \mathcal{V})$, let

$$\mathcal{D} \triangleq \left\{ p \in \mathcal{P} \quad : \quad \begin{array}{c} p(\cdot) = \sum_{\mathcal{V}} \beta(\mathcal{V}) f(\cdot|\mathcal{V}), \\ f(\cdot|\mathcal{V}) \text{ is a PMF on } \mathcal{V}, \\ \forall\, \mathcal{V} \text{ s.t. } \mathcal{V} \subseteq \mathcal{X},\ |\mathcal{V}| \leq m \end{array} \right\}. \tag{4.4}$$

$\mathcal{D}$ is the set of IID distributions the AVS can 'simulate' using these memoryless rules requiring 1-step lookahead. It is clear by construction that $\mathcal{D} \subseteq \mathcal{C}$. Also, it is clear that both $\mathcal{C}$ and $\mathcal{D}$ are convex sets of distributions. Lemma 28 in Appendix C.1 uses a separating hyperplane argument to show $\mathcal{D} = \mathcal{C}$. The adversary can therefore simulate any IID source with distribution in $\mathcal{C}$ and hence $R(D) \geq \widetilde{R}(D)$.

Qualitatively, allowing the switcher to 'cheat' gives access to distributions $p \in \mathcal{C}$ which may not be in $\overline{\mathcal{G}}$. Quantitatively, the conditions placed on the distributions in $\mathcal{C}$ are precisely those that restrict the switcher from producing symbols that do not occur often enough on average. For example, let $\mathcal{V} = \{1\}$ where $1 \in \mathcal{X}$, and suppose that the subsources are independent of each other. Then for every $p \in \mathcal{C}$,

$$p(1) \geq \prod_{l=1}^{m} p_l(1).$$

$\prod_{l=1}^{m} p_l(1)$ is the probability that all m subsources produce the letter 1 at a given time. In this case, the switcher has no option but to output the letter 1, hence any distribution the switcher mimics must have $p(1) \geq \prod_{l=1}^{m} p_l(1)$. The same logic can be applied to all subsets $\mathcal{V}$ of $\mathcal{X}$.

4.4 Noisy observations of subsource realizations

A natural extension of the AVS model of Fig. 4.1 is to consider the case when the adversary has noisy access to subsource realizations through a discrete memoryless channel. Suppose we let the switcher observe y_k at time k, which is probabilistically related to the subsource realizations through a discrete memoryless multiple access channel W by

$$W(y_k | x_{1,k}, x_{2,k}, \ldots, x_{m,k}).$$

Since the subsource probability distributions are already known, through an application of Bayes' rule, this model is equivalent to one in which the switcher observes a state, $t_k = y_k$, noiselessly. Namely,

$$\begin{gathered} Pr(x_{1,k} = x_1, \ldots, x_{m,k} = x_m | t_k = t) = \\ \frac{P(x_1, \ldots, x_m) W(t | x_1, \ldots, x_m)}{\sum_{x'_1, \ldots, x'_m} W(t | x'_1, \ldots, x'_m) P(x'_1, \ldots, x'_m)}. \end{gathered}$$

Conditioned on the state, the m subsources emit symbols independent of the past according to a conditional distribution. This model is depicted in Fig. 4.2.

The overall AVS is comprised now of a 'state generator' and a 'symbol generator' that outputs m symbols at a time. The state generator produces the state t_k at time k from a finite set $\mathcal{T}$. We assume the states are generated IID across time with distribution $\alpha(t)$. At time k,

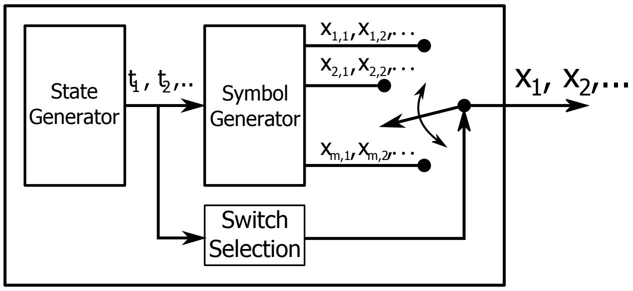

Figure 4.2: A model of an AVS encompassing both cheating and non-cheating switchers. Additionally, this model allows for noisy observations of subsource realizations by the switcher.

the symbol generator outputs $(x_{1,k}, \ldots, x_{m,k})$ according to $P(x_{1,k}, \ldots, x_{m,k}|t_k)$. This model allows for correlation among the subsources at a fixed time. Let $p_l(\cdot|t), l = 1, \ldots, m$, be the marginals of this joint distribution so that conditioned on t_k, $x_{l,k}$ has marginal distribution $p_l(\cdot|t_k)$. For a fixed $t \in \mathcal{T}$, let $\overline{\mathcal{G}}(t) = \mathbf{conv}(p_1(\cdot|t), \ldots, p_m(\cdot|t))$.

The switcher can observe states either with full lookahead or 1-step lookahead, but these two cases will once again have the same rate-distortion function when the switcher is an adversary. So assume that at time k, the switcher chooses the switch position s_k with knowledge of $\mathbf{t}^n, \mathbf{x}_1^{k-1}, \ldots, \mathbf{x}_m^{k-1}$. The strictly causal and 1-step lookahead switchers with noiseless subsource observations can be recovered as special cases of this model. If the conditional distributions $p_l(x|t)$ do not depend on t, the strictly causal switcher is recovered. The full lookahead switcher with noiseless subsource observations is recovered by setting $\mathcal{T} = \mathcal{X}^m$ and letting $p_l(x|t) = 1(x = t(l))$ where the state t is an m dimensional vector consisting of the outputs of each subsource.

With this setup, we have the following extension of Theorem 9.

Theorem 10. *For the AVS problem of Fig. 4.2, where the adversary has access to the states either with* 1*-step lookahead or full lookahead,*

$$R(D) = \max_{p \in \mathcal{D}_{states}} R(p, D), \tag{4.5}$$

where

$$\mathcal{D}_{states} = \left\{ p : \begin{array}{c} p(\cdot) = \sum_{t \in \mathcal{T}} \alpha(t) f(\cdot|t) \\ f(\cdot|t) \in \overline{\mathcal{G}}(t), \forall\ t \in \mathcal{T} \end{array} \right\}. \tag{4.6}$$

Proof: See Appendix C.2.

One can see that in the case of the cheating switcher of the previous section, the set $\mathcal{D}$ of (4.4) equates directly with $\mathcal{D}_{states}$ of (4.6). In that sense, from the switcher's point of view, $\mathcal{D}$ is a more natural description of the set of distributions that can be simulated than $\mathcal{C}$. Again, actually computing $R(D)$ in (4.5) falls into the discussion of Section 4.7.

4.5 The helpful switcher

Arbitrarily varying sources and channels have generally been associated with adversarial source and channel coding, but in this section, we consider the *helpful* cheating switcher to more thoroughly explore the information-theoretic game established in [4]. The goal of the helpful switcher is to help the coding system achieve low distortion. The model is as follows:

- The coder chooses a codebook that is made known to the switcher.
- The switcher chooses a strategy to help the coder achieve distortion D on average with the minimum number of codewords. We consider the cases where the switcher has full lookahead or 1-step lookahead.

As opposed to the adversarial setting, a rate R is now achievable at distortion D if *there exist* switcher strategies and codebooks for each n with expected distortion at most D and the rates of the codebooks tend to R. The following theorem establishes $R(D)$ if the cheating switcher has full lookahead.

Theorem 11. *Let $\mathcal{X}^* = \{\mathcal{V} \subseteq \mathcal{X} : \mathcal{V} \neq \emptyset, |\mathcal{V}| \leq m\}$. Let $\rho : \mathcal{X}^* \times \widehat{\mathcal{X}} \to [0, d^*]$ be defined by*

$$\rho(\mathcal{V}, \widehat{x}) = \min_{x \in \mathcal{V}} d(x, \widehat{x}).$$

Let $\mathcal{V}_k = \{x_{1,k}, \dots, x_{m,k}\}$ for all k. Note that $\mathcal{V}_i, i = 1, 2, \dots$ is a sequence of IID random variables with distribution $\beta(\mathcal{V}) = P(\{x_{1,1}, \dots, x_{m,1}\} = \mathcal{V})$. Let $R^*(\beta, D)$ be the rate-distortion function for this new IID source with distribution β at distortion D with respect to the distortion measure $\rho(\cdot, \cdot)$. For the helpful cheating switcher with full lookahead,

$$R(D) = R^*(\beta, D). \tag{4.7}$$

Proof: Rate-distortion problems are essentially covering problems, so we equate the rate-distortion problem for the helpful switcher with the classical covering problem for the observed sets $\mathcal{V}_i$. If the switcher is helpful, has full lookahead, and knowledge of the codebook, the problem of designing the codebook is equivalent to designing the switcher strategy and codebook jointly. At each time k, the switcher observes a candidate set $\mathcal{V}_k$ and must select an element from $\mathcal{V}_k$. For any particular reconstruction codeword $\widehat{\mathbf{x}}^n$, and a string of candidate sets $(\mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_n)$, the switcher can at best output a sequence $\mathbf{x}^n$ such that

$$d_n(\mathbf{x}^n, \widehat{\mathbf{x}}^n) = \frac{1}{n} \sum_{k=1}^{n} \rho(\mathcal{V}_k, \widehat{x}_k)$$

Hence, for a codebook $\mathcal{B}$, the helpful switcher with full lookahead can select switch positions to output $\mathbf{x}^n$ such that, at best

$$\begin{aligned} d_n(\mathbf{x}^n;\mathcal{B}) &= \min_{\widehat{\mathbf{x}}^n\in\mathcal{B}} \frac{1}{n}\sum_{k=1}^{n} \min_{x\in\mathcal{V}_k} d(x,\widehat{x}_k) \\ &= \min_{\widehat{\mathbf{x}}^n\in\mathcal{B}} \frac{1}{n}\sum_{k=1}^{n} \rho(\mathcal{V}_k,\widehat{x}_k). \end{aligned}$$

Therefore, for the helpful switcher with full lookahead, the problem of covering the $\mathcal{X}$ space with respect to the distortion measure $d(\cdot,\cdot)$ now becomes one of covering the $\mathcal{X}^*$ space with respect to the distortion measure $\rho(\cdot,\cdot)$.

Remarks:

- Computing $R(D)$ in (4.7) can be done by the Blahut-Arimoto algorithm [21].
- In the above proof, full lookahead was required in order for the switcher to align the entire output word of the source with the minimum distortion reconstruction codeword as a whole. This process cannot be done with 1-step lookahead and so the $R(D)$ function for a helpful switcher with 1-step lookahead remains an open question, but we have the following corollary of Theorems 9 and 11.

Corollary 1. *For the helpful switcher with* 1*-step lookahead,*

$$R^*(\beta,D) \le R(D) \le \min_{p\in\mathcal{C}} R(p,D)$$

Proof: If the switcher has at least 1-step lookahead, it immediately follows from the proof of Theorem 9 that $R(D) \le \min_{p\in\mathcal{C}} R(p,D)$. The question is whether or not any lower rate is achievable. We can make the helpful switcher with 1-step lookahead more powerful by giving it n-step lookahead, which yields the lower bound $R^*(\beta,D)$.

An example in Section 4.6.2 shows that in general, we have the strict inequality $R^*(\beta,D) < \min_{p\in\mathcal{C}} R(p,D)$.

One can also investigate the helpful switcher problem when the switcher has access to noisy or partial observations as in Section 4.4. This problem has the added flavor of remote source coding because the switcher can be thought of as an extension of the coder and observes data correlated with the source to be encoded. However, the switcher has the additional capability of choosing the subsource that must be encoded. For now, this problem is open and we can only say that $R(D) \le \min_{p\in\mathcal{D}_{states}} R(p,D)$.

4.6 Examples

We illustrate the results with several simple examples using binary alphabets and Hamming distortion, i.e., $\mathcal{X} = \widehat{\mathcal{X}} = \{0,1\}$ and $d(x,\widehat{x}) = 1(x \ne \widehat{x})$. Recall that the rate-distortion function of an IID binary source with distribution $(1-p, p)$, $p \in [0, \frac{1}{2}]$ is

$$R((1-p,p),D) = \begin{cases} h_b(p) - h_b(D) & D \in [0,p] \\ 0 & D > p \end{cases},$$

where $h_b(p)$ is the binary entropy function (in bits for this section).

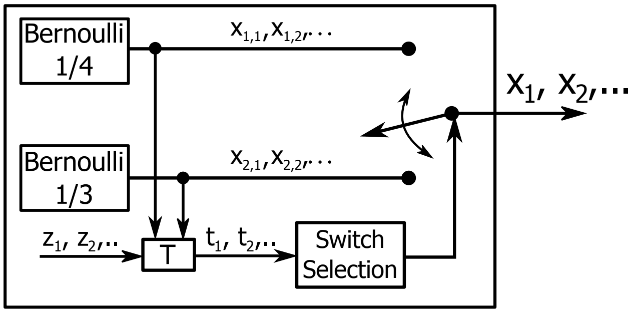

Figure 4.3: Two independent Bernoulli subsources, which produce 1's with probabilities 1/4 and 1/3.

4.6.1 Bernoulli $1/4$ and $1/3$ sources

Consider the example shown in Fig. 4.3 where the switcher has access to two independent IID Bernoulli subsources. Subsource 1 outputs 1 with probability 1/4 and subsource 2 outputs 1 with probability 1/3, so $p_1 = (3/4, 1/4)$ and $p_2 = (2/3, 1/3)$. At time i, the switcher is given access to an observation $t_k = T(x_{1,k}, x_{2,k}, z_k)$ where T is a function and z_k is independent noise (that is, the switcher observes a potentially noisy version of the subsource realizations).

First, we consider the switcher as an adversary in the traditional strictly causal setting of [4] and the 1-step lookahead setting, where switcher has the subsource realizations $t_k = (x_{1,k}, x_{2,k})$ before choosing the switch position. For any time k,

$$\begin{aligned} P(x_{1,k} = x_{2,k} = 0) &= \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2} \\ P(x_{1,k} = x_{2,k} = 1) &= \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{12} \\ P(\{x_{1,k}, x_{2,k}\} = \{0,1\}) &= 1 - \frac{1}{2} - \frac{1}{12} = \frac{5}{12}. \end{aligned}$$

If the switcher is allowed 1-step lookahead and has the option of choosing either 0 or 1, suppose the switcher chooses 1 with probability f_1. The coder then sees an IID binary source with a probability of a 1 occurring being equal to:

$$p(1) = \frac{1}{12} + \frac{5}{12} f_1.$$

By using f_1 as a parameter, the switcher can produce 1's with any probability between 1/12 and 1/2. The attainable distributions are shown in Fig. 4.4. The switcher with lookahead can simulate a significantly larger set of distributions than the strictly causal switcher, which is restricted to outputting 1's with a probability in $[1/4, 1/3]$. Thus, for the strictly causal switcher, $R(D) = h_b(1/3) - h_b(D)$ for $D \in [0, 1/3]$ and for the switcher with 1-step or full lookahead, $R(D) = 1 - h_b(D)$ for $D \in [0, 1/2]$.

We now look at several variations of this example to illustrate the utility of noisy or partial observations of the subsources for the switcher. In the first variation, the switcher observes the mod-2 sum of the two subsources $t_k = x_{1,k} \oplus x_{2,k}$. Theorem 10 then implies that $R(D) = h_b(1/3) - h_b(D)$ for $D \in [0, 1/3]$. Hence, the mod-2 sum of these two subsources is useless to the switcher in deciding the switch position. This is intuitively clear from the symmetry of the mod-2 sum. If $t_k = 0$, either both subsources are 0 or both subsources are 1, so the switch position doesn't matter in this state. If $t_k = 1$, one of the subsources has output 1 and the other has output 0, but because of the symmetry of the mod-2 function, the switcher's prior as to which subsource output the 1 does not change and it remains that subsource 2 was more likely to have output the 1.

In the second variation, the switcher observes the second subsource directly but not the first, so $t_k = x_{2,i}$ for all k. Using Theorem 10 again, it can be deduced that in this case $R(D) = 1 - h_b(D)$ for $D \in [0, 1/2]$. This is also true if $t_k = x_{1,k}$ for all i, so observing just one of the subsources noncausally is as beneficial to the switcher as observing both subsources noncausally. This is clear in this example because the switcher is attempting to output as many 1's as possible. If $t = 1$, the switcher will set the switch position to 2 and if $t = 0$, the switcher will set the switch position to 1 as there is still a chance that the first subsource outputs a 1.

For this example, the helpful cheater with 1-step lookahead has a rate-distortion function that is upper bounded by $h_b(1/12) - h_b(D)$ for $D \in [0, 1/12]$. The rate-distortion function for the helpful cheater with full lookahead can be computed from Theorem 11. In Fig. 4.5, the rate-distortion function is plotted for the situations discussed so far.

Finally, consider an example where an adversarial switcher observes only the second subsource through a binary symmetric channel with crossover probability $\delta \in [0, 1/2]$, i.e., $t_k = x_{2,k} \oplus z_k$ where z_k is a Bernoulli sequence that produces 1's with probability δ. Applying Theorem 10 again, it can be shown that if $\delta \in [0, 2/5]$,

$$R(D) = h_b\left(\frac{1}{2} - \frac{5}{12}\delta\right) - h_b(D),\ D \in \left[0, \frac{1}{2} - \frac{5}{12}\delta\right]$$

and if $\delta \in [2/5, 1/2]$,

$$R(D) = h_b\left(\frac{1}{3}\right) - h_b(D),\ D \in \left[0, \frac{1}{3}\right].$$

Here, increasing δ decreases the switcher's knowledge of the subsource realizations. Somewhat surprisingly, the utility of the observation is exhausted at $\delta = 2/5$, even before the state

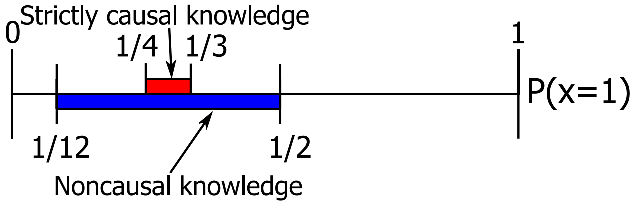

Figure 4.4: The binary distributions the switcher can mimic. $\overline{\mathcal{G}}$ is the set of distributions the switcher can mimic with strictly causal access to subsource realizations, and $\mathcal{C}$ is the set attainable with noncausal access.

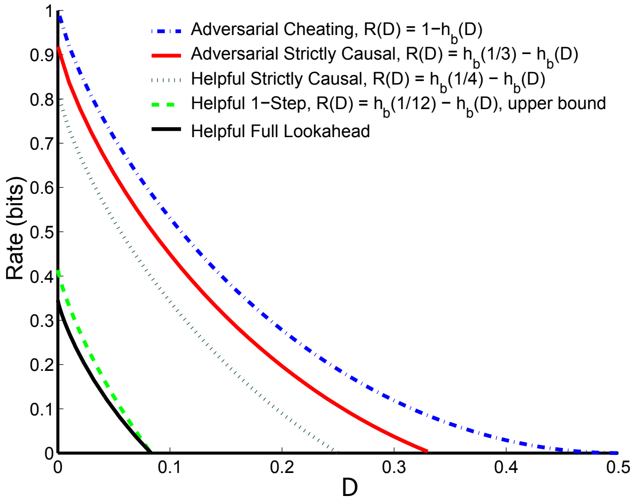

Figure 4.5: $R(D)$ for the cheating switcher and the non-cheating switcher with Bernoulli 1/4 and 1/3 subsources. Also, the rate-distortion function for the examples of Fig. 4.3 where $t_k = x_{1,k} \oplus x_{2,k}$ and $t_k = x_{2,k}$.

and observation are completely independent at $\delta = 1/2$. This can be explained through the switcher's *a posteriori* belief that the second subsource output was a 1 given the state. If the switcher observes $t_k = 1$ and $\delta \leq 1/2$, $p(x_{2,k} = 1|t_k = 1) \geq 1/3 > 1/4$ so the switch position will be set to 2. When the switcher observes $t_k = 0$, if $\delta \leq 2/5$, $p(x_{2,k} = 1|t_k = 0) \leq 1/4$, so the switch will be set to position 1. However, if $\delta > 2/5$, $p(x_{2,k} = 1|t_k = 0) > 1/4$, so the switch position will be set to 2 even if $t = 0$ because the switcher's *a posteriori* belief is that the second subsource is *still* more likely to have output a 1 than the first subsource. Fig. 4.6 shows $R(D)$ for this example as a function of δ for two values of D.

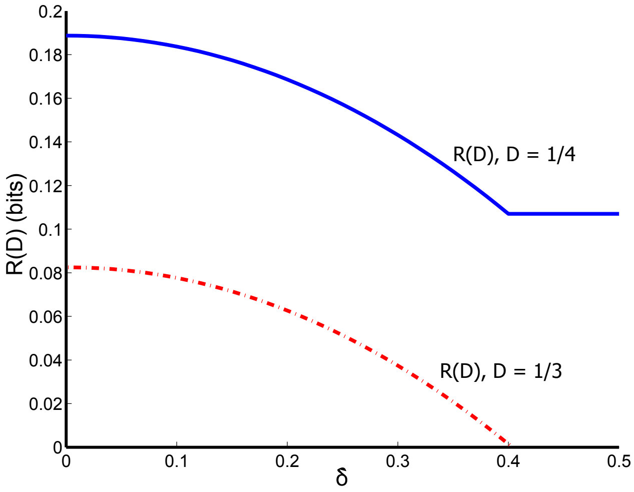

Figure 4.6: $R(D)$ as a function of the noisy observation crossover probability δ for $D = 1/3$ and $D = 1/4$ for the example of Fig. 4.3 with $t_k = x_{2,k} \oplus z_k$ and $z_k \sim \mathcal{B}(\delta)$.

4.6.2 Two Bernoulli $1/2$ subsources

Suppose $m = 2$, and the subsources are independent Bernoulli $1/2$ IID processes. For this example, the rate-distortion function is $R(D) = 1 - h_b(D)$ for $D \in [0, 1/2]$ whether the adversarial switcher is strictly causal, causal or noncausal. When the helpful switcher has 1-step lookahead, $R(D) \leq R_U(D) = h_b(1/4) - h_b(D)$ for $D \in [0, 1/4]$. One can also think of this upper bound as being the rate-distortion function for the helpful switcher with 1-step lookahead that is restricted to using memoryless, time-invariant rules. Using Theorem 9.4.1 of [17] and Theorem 11, one can show that when the switcher has full lookahead with $t_k = (x_{1,k}, x_{2,k})$,

$$R(D) = R^*(\beta, D) = \frac{1}{2}\left[1 - h_b(2D)\right], \ D \in [0, 1/4].$$

The plot of these functions in Fig. 4.7 shows that the rate-distortion function can be significantly reduced if the helpful switcher is allowed to observe the entire block of subsource realizations. It is also interesting to note *how* the switcher with full lookahead helps the coder achieve a rate of $R^*(\beta, D)$. In this example $\mathcal{X}^* = \{\{0\}, \{1\}, \{0,1\}\}$, $\rho(\{0\}, \widehat{x}) = 1(0 \neq \widehat{x})$, $\rho(\{1\}, \widehat{x}) = 1(1 \neq \widehat{x})$, $\rho(\{0,1\}, \widehat{x}) = 0$ and $\beta = (1/4, 1/4, 1/2)$. The $R^*(\beta, D)$ achieving distribution on $\widehat{\mathcal{X}}$ is $(1/2, 1/2)$, but $R^*(\beta, D) < 1 - h_b(D)$. The coder is attempting to cover strings with types near $(1/2, 1/2)$ but with far fewer codewords than are needed to actually cover all such strings. This problem is circumvented through the aid provided by the switcher in pushing the output of the source inside the Hamming D-ball of a codeword. This is in contrast to the strategy that achieves $R_U(D)$, where the switcher makes the output an IID sequence with as few 1's as possible and the coder is expected to cover *all* strings with types near $(3/4, 1/4)$.

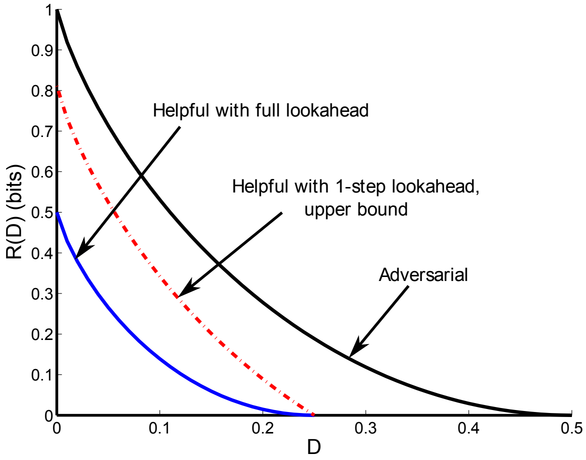

Figure 4.7: The $R(D)$ function for an AVS with two Bernoulli 1/2 sources when the switcher is helpful with full lookahead. For 1-step lookahead, the upper bound is shown.

4.7 Computing $R(D)$ for an AVS

The $R(D)$ function for an adversarial AVS with either causal or noncausal access to the subsource realizations is of the form

$$R(D) = \max_{p \in \mathcal{Q}} R(p, D), \tag{4.8}$$

where $\mathcal{Q}$ is a set of distributions in $\mathcal{P}(\mathcal{X})$. In (4.2), (4.3), and (4.6) $\mathcal{Q}$ is defined by a finite number of linear inequalities and hence is a polytope. The number of constraints in the definition of $\mathcal{Q}$ is exponential in $|\mathcal{X}|$ or $|\mathcal{T}|$ when the adversary has something other than strictly causal knowledge. Unfortunately, the problem of finding $R(D)$ is not a convex program because $R(p, D)$ is not a concave-$\cap$ function of p in general. In fact, $R(p, D)$ may not even be quasi-concave and may have multiple local maxima with values different from the global maximum, as shown by Ahlswede [45].

Since standard convex optimization tools are unavailable for this problem, we consider the question of how to approximate $R(D)$ to within some (provable) precision. That is, for any $\epsilon > 0$, we will consider how to provide an approximation $R_a(D)$ such that $|R_a(D) - R(D)| \leq \epsilon$. Note that for fixed p, $R(p, D)$ can be computed efficiently by the Blahut-Arimoto algorithm to any given precision, say much less than ϵ. Therefore, we assume that $R(p, D)$ can be computed for a fixed p and D. We also assume $D \geq D_{\min}(\mathcal{Q})$ since otherwise $R(D) = \infty$. Checking this condition is a linear program since $\mathcal{Q}$ is a polytope and $D_{\min}(p)$ is linear in p.

We will take a 'brute-force' approach to computing $R(D)$. That is, we wish to compute $R(p, D)$ for (finitely) many p and then maximize over the computed values to yield $R_a(D)$. Since $R(p, D)$ is uniformly continuous in p, it is possible to do this and have $|R_a(D) - R(D)| \leq \epsilon$ provided enough distributions p are 'sampled'. Undoubtedly, there are other algorithms to compute $R(D)$ that likely have better problem-size dependence. In this section, we are only interested in showing that $R(D)$ can provably be computed to within any required precision with a finite number of computations.

4.7.1 Uniform continuity of $R(p, D)$

The main tool used to show that the rate-distortion function can be approximated is an explicit bound on the uniform continuity of $R(p, D)$ in terms of $\|p - q\|_1 = \sum_{x \in \mathcal{X}} |p(x) - q(x)|$ for distortion measures that allow for 0-distortion to be achieved regardless of the source. In [21], a bound on the continuity of the entropy of a distribution is developed in terms of $\|p - q\|_1$.

Lemma 8 (Uniform continuity of entropy, [21])**.** *Let p and q be two probability distributions on $\mathcal{X}$ such that $\|p - q\|_1 \leq 1/2$, then*

$$|H(p) - H(q)| \leq \|p - q\|_1 \ln \frac{|\mathcal{X}|}{\|p - q\|_1}.$$

In the following lemma, a similar uniform continuity is stated for $R(p, D)$. The proof makes use of Lemma 8.

Lemma 9 (Uniform continuity of $R(p, D)$)**.** *Let $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$ be a distortion function. $\widetilde{d}$ is the minimum nonzero distortion from (4.1). Also, assume that for each $x \in \mathcal{X}$, there is an $\hat{x}_0(x) \in \widehat{\mathcal{X}}$ such that $d(x, \hat{x}_0(x)) = 0$. Then, for $p, q \in \mathcal{P}(\mathcal{X})$ with $\|p - q\|_1 \leq \frac{\widetilde{d}}{4d^*}$, for any $D \geq 0$,*

$$|R(p, D) - R(q, D)| \leq \frac{7d^*}{\widetilde{d}} \|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1}. \tag{4.9}$$

Proof: See Appendix C.3.

The restriction that $d(x, \cdot)$ has at least one zero for every x can be relaxed if we are careful about recognizing when $R(p, D)$ is infinite. For an arbitrary distortion measure $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$, define another distortion measure $d_0 : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$ by

$$d_0(x, \widehat{x}) = d(x, \widehat{x}) - \min_{\widetilde{x} \in \widehat{\mathcal{X}}} d(x, \widetilde{x}).$$

Now let $d_0^* = \max_{x,\widehat{x}} d_0(x, \widehat{x})$ and $\widetilde{d_0} = \min_{(x,\widehat{x}):d_0(x,\widehat{x})>0} d_0(x, \widehat{x})$. We have defined $d_0(x, \widehat{x})$ so that Lemma 9 applies, so we can prove the following lemma.

Lemma 10. *Let $p, q \in \mathcal{P}(\mathcal{X})$ and let $D \geq \max(D_{\min}(p), D_{\min}(q))$. If $\|p - q\|_1 \leq \widetilde{d_0}/4d^*$,*

$$|R(p, D) - R(q, D)| \leq \frac{11d^*}{\widetilde{d_0}} \|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1}.$$

Proof: See Appendix C.4.

As $\|p - q\|_1$ goes to 0, $-\ln \|p - q\|_1$ goes to infinity slowly and it can be shown that for any $\delta \in (0, 1)$ and $\gamma \in [0, 1/2]$,

$$\gamma \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\gamma} \leq \frac{(|\mathcal{X}||\widehat{\mathcal{X}}|)^\delta}{e\delta} \gamma^{1-\delta}. \tag{4.10}$$

In the sequel, we let $f(\gamma) = \gamma \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\gamma}$ for $\gamma \in [0, 1/2]$ with $f(0) = 0$ by continuity. It can be checked that f is strictly monotonically increasing and continuous on $[0, 1/2]$ and hence has an inverse function $g : f([0, 1/2]) \to [0, 1/2]$, i.e., $g(f(\gamma)) = \gamma$ for all $\gamma \in [0, 1/2]$. Note that g is not expressible in a simple 'closed-form', but can be computed numerically. Also, by inverting (4.10), we have a lower bound on $g(r)$ for any $r \in [0, f(1/2)]$ and $\delta \in (0, 1)$,

$$g(r) \geq \left(\frac{e\delta}{\left(|\mathcal{X}||\widehat{\mathcal{X}}|\right)^\delta} r \right)^{1/(1-\delta)}. \tag{4.11}$$

4.7.2 A bound on the number of distributions to sample

Returning to the problem of computing $R(D)$ in (4.8), consider the following simple algorithm. Without loss of generality, assume $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$. Let $\gamma \in (0, 1)$ and let $\gamma\mathbb{Z}^{|\mathcal{X}|-1}$ be the $|\mathcal{X}| - 1$ dimensional integer lattice scaled by γ. Let $\widetilde{\mathcal{O}} = [0, 1]^{|\mathcal{X}|-1} \bigcap \gamma\mathbb{Z}^{|\mathcal{X}|-1}$. Now, define

$$\mathcal{O} = \left\{ q \in \mathcal{P}(\mathcal{X}) : \begin{array}{c} \exists\ \widetilde{q} \in \widetilde{\mathcal{O}} \text{ s.t.} \\ q(i) = \widetilde{q}(i), i = 1, \ldots, |\mathcal{X}| - 1, \\ q(|\mathcal{X}|) = 1 - \sum_{i=1}^{|\mathcal{X}|-1} \widetilde{q}(i) \geq 0 \end{array} \right\}.$$

In words, sample the $|\mathcal{X}|-1$ dimensional unit cube, $[0,1]^{|\mathcal{X}|-1}$, uniformly with points from a scaled integer lattice. Embed these points in $\mathbb{R}^{|\mathcal{X}|}$ by assigning the last coordinate of the new vector to be 1 minus the sum of the values in the original point. If this last value is non-negative, the new point is a distribution in $\mathcal{P}(\mathcal{X})$. The algorithm to compute $R_a(D)$ is then one where we compute $R(p,D)$ for distributions $q \in \mathcal{O}$ that are in or close enough to $\mathcal{Q}$.

1. Fix a $q \in \mathcal{O}$. If $\min_{p\in\mathcal{Q}} \|p-q\|_1 \leq 2|\mathcal{X}|\gamma$, compute $R(q,D)$, otherwise do not compute $R(q,D)$. Repeat for all $q \in \mathcal{O}$.

2. Let $R_a(D)$ be the maximum of the computed values of $R(q,D)$, i.e.,

$$\begin{aligned} R_a(D) &= \max\Big\{ R(q,D) : q \in \mathcal{O}, \\ &\qquad\qquad \min_{p\in\mathcal{Q}} \|p-q\|_1 \leq 2|\mathcal{X}|\gamma \Big\}. \end{aligned}$$

Checking the condition $\min_{p\in\mathcal{Q}} \|p-q\|_1 \leq \gamma 2|\mathcal{X}|$ is essentially a linear program, so it can be efficiently solved. By setting γ according to the accuracy $\epsilon > 0$ we want, we get the following result.

Theorem 12. *The preceding algorithm computes an approximation $R_a(D)$ such that $|R_a(D)-R(D)| \leq \epsilon$ if*

$$\gamma \leq \frac{1}{2|\mathcal{X}|} g\left(\frac{\epsilon \widetilde{d}_0}{11 d^*}\right).$$

The number of distributions for which $R(q,D)$ is computed to determine $R(D)$ to within accuracy ϵ is at most[7]

$$N(\epsilon) \leq \left(\frac{2|\mathcal{X}|}{g\left(\frac{\epsilon \widetilde{d}_0}{11d^*}\right)} + 2\right)^{|\mathcal{X}|-1}.$$

Proof: The bound on $N(\epsilon)$ is clear because the number of points in $\widetilde{\mathcal{O}}$ is at most $(\lceil 1/\gamma \rceil + 1)^{|\mathcal{X}|-1}$ and every distribution in $\mathcal{O}$ is associated with one in $\widetilde{\mathcal{O}}$, so $|\mathcal{O}| \leq |\widetilde{\mathcal{O}}|$.

Now, we prove $|R_a(D)-R(D)| \leq \epsilon$. For this discussion, we let $\gamma = \frac{1}{2|\mathcal{X}|} g\left(\frac{\epsilon \widetilde{d}_0}{11d^*}\right)$. First, for all $p \in \mathcal{Q}$, there is a $q \in \mathcal{O}$ with $\|p-q\|_1 \leq g\left(\frac{\epsilon \widetilde{d}_0}{11d^*}\right) = 2|\mathcal{X}|\gamma$. To see this, let $\widetilde{q}(i) = \lfloor \frac{p(i)}{\gamma} \rfloor \gamma$

[7] This is clearly not the best bound as many of the points in the unit cube do not yield distributions on $\mathcal{P}(\mathcal{X})$. The factor by which we are overbounding is roughly $|\mathcal{X}|!$, but this factor does not affect the dependence on ϵ.

for $i = 1, \ldots, |\mathcal{X}| - 1$. Then $\widetilde{q} \in \widetilde{\mathcal{O}}$, and we let $q(i) = \widetilde{q}(i)$ for $i = 1, \ldots, |\mathcal{X}| - 1$. Note that

$$\begin{aligned} q(|\mathcal{X}|) &= 1 - \sum_{i=1}^{|\mathcal{X}|-1} q(i) = 1 - \sum_{i=1}^{|\mathcal{X}|-1} \left\lfloor \frac{p(i)}{\gamma} \right\rfloor \gamma \\ &\geq 1 - \sum_{i=1}^{|\mathcal{X}|-1} p(i) = p(|\mathcal{X}|) \geq 0. \end{aligned}$$

Therefore $q \in \mathcal{O}$ and furthermore,

$$\begin{aligned} \|p - q\|_1 &\leq \left(1 - \sum_{i=1}^{|\mathcal{X}|-1} (p(i) - \gamma) - p(|\mathcal{X}|)\right) + \\ &\quad \sum_{i=1}^{|\mathcal{X}|-1} \left(p(i) - \left\lfloor \frac{p(i)}{\gamma} \right\rfloor \gamma\right) \\ &\leq 2(|\mathcal{X}| - 1)\gamma \leq 2|\mathcal{X}|\gamma \leq g\left(\frac{\epsilon \widetilde{d}_0}{11 d^*}\right). \end{aligned}$$

By Lemma 10, $R(q, D) \geq R(p, D) - \epsilon$. This distribution q (or possibly one closer to p) will always be included in the maximization yielding $R_a(D)$, so we have $R_a(D) \geq \max_{p \in \mathcal{Q}} R(p, D) - \epsilon = R(D) - \epsilon$.

Conversely, for a $q \in \mathcal{O}$, if $\min_{p \in \mathcal{Q}} \|p - q\|_1 \leq 2|\mathcal{X}|\gamma$, Lemma 10 again gives

$$R(q, D) \leq \max_{p \in \mathcal{Q}} R(p, D) + \epsilon = R(D) + \epsilon$$

Therefore, $|R_a(D) - R(D)| \leq \epsilon$. To get a sense of how $N(\epsilon)$ scales as ϵ goes to 0, we can use the bound of (4.11) with an arbitrary value of $\delta \in (0, 1)$. For example, with $\delta = 1/2$, the scaling becomes

$$\begin{aligned} N(\epsilon) &\leq \left(\frac{2|\mathcal{X}|}{\left(\frac{e \widetilde{d}_0}{22 d^* \sqrt{|\mathcal{X}||\widehat{\mathcal{X}}|}}\right)^2} \cdot \frac{1}{\epsilon^2} + 2\right)^{|\mathcal{X}|-1} \\ &= O\left((1/\epsilon)^{2(|\mathcal{X}|-1)}\right). \end{aligned}$$

4.7.3 Estimation of the rate-distortion function of an unknown IID source

An explicit bound on the continuity of the rate-distortion function has other applications. Recently, Harrison and Kontoyiannis [51] have studied the problem of estimating the rate-distortion function of the marginal distribution of an unknown source. Let $p_{\mathbf{x}^n}$ be the (marginal) empirical distribution of a vector $\mathbf{x}^n \in \mathcal{X}^n$. They show that the 'plug-in' estimator $R(p_{\mathbf{x}^n}, D)$, the rate-distortion function of the empirical marginal distribution of a sequence, is a consistent estimator for a large class of sources beyond just IID sources with known alphabets. However, if the source is known to be IID with alphabet size $|\mathcal{X}|$, estimates of the convergence rate (in probability) of the estimator can be provided using the uniform continuity of the rate-distortion function.

Suppose the true source is IID with distribution $p \in \mathcal{P}(\mathcal{X})$ and fix a probability $\tau \in (0,1)$ and an $\epsilon \in (0, \ln|\mathcal{X}|)$. We wish to answer the question: How many samples n need to be taken so that $|R(p_{\mathbf{x}^n}, D) - R(p, D)| \leq \epsilon$ with probability at least $1-\tau$? The following lemma gives a sufficient number of samples n.

Lemma 11. *Let $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$ be a distortion measure for which Lemma 9 holds. For any $p \in \mathcal{P}(\mathcal{X})$, $\tau \in (0,1)$, and $\epsilon \in (0, \ln|\mathcal{X}|)$,*

$$P(|R(p_{\mathbf{x}^n}, D) - R(p, D)| \geq \epsilon) \leq \tau$$

if

$$n > \frac{2}{g\left(\frac{\epsilon\widetilde{d}}{7d^*}\right)^2}\left(\ln\frac{1}{\tau} + |\mathcal{X}|\ln 2\right). \tag{4.12}$$

Proof: From Lemma 9, we have

$$\begin{aligned} P(|R(p_{\mathbf{x}^n}, D) - R(p, D)| \geq \epsilon) &\leq P\left(\|p_{\mathbf{x}^n} - p\|_1 \geq g\left(\frac{\epsilon\widetilde{d}}{7d^*}\right)\right) \\ &\leq 2^{|\mathcal{X}|}\exp\left(-\frac{n}{2}g\left(\frac{\epsilon\widetilde{d}}{7d^*}\right)^2\right) \end{aligned}$$

The last line follows from Theorem 2.1 of [52]. This bound is similar to, but a slight improvement over, the method-of-types bound of Sanov's Theorem. Rather than an $(n+1)^{|\mathcal{X}|}$ term, we just have a $2^{|\mathcal{X}|}$ term multiplying the exponential. Taking ln of both sides gives the desired result.

We emphasize that this number n is a sufficient number of samples regardless of what the true distribution $p \in \mathcal{P}(\mathcal{X})$ is. The bound of (4.12) depends only on the distortion measure d, alphabet sizes $|\mathcal{X}|$ and $|\widehat{\mathcal{X}}|$, desired accuracy ϵ and 'estimation error' probability τ.

4.8 Concluding remarks

In this chapter, we have seen how the rate-distortion function for an AVS is affected by various constraints on the switcher's knowledge involving causality and noise in observations (see Table 4.1). Several other natural constraints come to mind. First, there might

Switcher model	$R(D)$
Time-invariant (adversarial) [50]	$\max_{p\in\mathcal{G}} R(p,D)$
Strictly causal (adversarial) [42]	$\max_{p\in\overline{\mathcal{G}}} R(p,D)$
Causal or noncausal (adversarial)	$\max_{p\in\mathcal{C}} R(p,D)$
Casual or noncausal noisy observations (adversarial)	$\max_{p\in\mathcal{D}_{states}} R(p,D)$
Noncausal (helpful)	$R^*(\beta,D)$

Table 4.1: Summary of results.

be a constraint on how much information the switcher has when making its decisions on subsampling. This could be handled by performing an optimization in Theorem 10 over all channels from the subsources to the state observations that satisfy a mutual information constraint. Secondly, one might be interested in studying the rate-distortion function if the switching speed is fixed or constrained in some way. Another interesting area to study might be 'mismatched objectives' where the switcher is trying to be helpful for some particular distortion metric but the source is actually being encoded with a different metric in mind. Here, some understanding of how the rate-distortion function behaves with continuity of the metric might prove useful.

Finally, if the active sensor and coding system are part of a tightly delay-constrained control loop, we would want to study these issues from the causal source code perspective of [53]. It seems likely that the adversarial results of Theorems 9 and 10 would follow straightforwardly with the same sets of distributions $\mathcal{C}$ and $\mathcal{D}$, with the IID rate-distortion function for noncausal source codes replaced by the IID rate-distortion functions for causal source codes.

Chapter 5

Conclusion

Let us now briefly look back on the contributions of this thesis and look forward to the many interesting questions that remain. Chapters 2 and 3 focused on an upper bound to the error exponent for two channel coding problems: fixed blocklength coding with feedback and fixed delay coding without feedback. The upper bound, called the Haroutunian bound, has a noncausal interpretation. It assumes that a code may be able to predict future channel behavior and adjust its input distribution accordingly, even though the channel is memoryless. The Haroutunian exponent luckily evaluates to the sphere-packing exponent for symmetric channels, but is strictly larger for asymmetric ones. It has been presumed for both problems that the sphere-packing exponent is a valid upper bound to the error exponent for all channels, but proving so ran into difficulty because little can be said about the local input distribution used by an arbitrary code during the error event.

For block codes with feedback, the core difficulty is that the input distribution depends on the channel behavior faced over the block through the feedback. However, the feedback is causal, in the sense that an output is only produced after an input is put into the channel. In Chapter 2, we documented several failed attempts made at circumventing this issue to prove that sphere-packing must hold for block codes with feedback. Unable to make progress for general codes, we then considered codes with feedback for which the input type is fixed regardless of the output sequence. We showed that for these ‘fixed-type encoding tree’ codes, the code appears to have the same input-output profile as a code without feedback when one considers the conditional relationship between input and output sequences. This result solidified the intuition that the only way a code with feedback could beat the sphere-packing bound is by changing its input distribution according to the channel behavior it sees. Next, we showed that if the feedback information is delayed, the sphere-packing exponent holds in the limit of large delays in the feedback path. This result was reinterpreted by looking at the Haroutunian exponent for a parallel channel constructed by using the original channel T times independently. We showed that, surprisingly, the Haroutunian exponent for the parallel channel converges to the sphere-packing exponent for the original channel as T gets large after normalization. In essence, this means that when grouping asymmetric channel uses together, the resulting channel looks more and more symmetric.

Ultimately, we were not able to show that the sphere-packing exponent holds for block codes with feedback without placing restrictions on the encoder. Why is this? Perhaps it is because beating the sphere-packing bound is possible with feedback over an asymmetric channel, although we think it unlikely. In our estimation, more needs to be understood about encoding trees with feedback. If they are not constrained to use fixed type inputs, is it possible that they can recognize when an error is unavoidable and 'give up' to yield something like the Haroutunian exponent? This seems unlikely because the Haroutunian exponent minimizing channel will be a channel V for which the capacity, $C(V)$ is (almost) equal to the rate R due to the convexity of capacity and conditional divergence. So, even if part of the way through the block, the encoder realizes the channel W is behaving like V, it cannot give up on decoding properly because for the rest of the block, the channel is likely to behave like W. Because W has a capacity larger than R, a good code will make up for poor channel behavior earlier in the block by using good channel behavior later in the block. After all, the whole point of channel coding is to smooth out channel noise over long periods of time. So we should not expect that the a good code will behave in such a way to yield anything like the Haroutunian exponent.

Additional progress in the practical problem could be made by studying noisy feedback. Much like delay, noise is an unavoidable hindrance in the feedback path of most systems. It would be interesting to know if small amounts of noise in the feedback path can be used to prove an upper bound to the error exponent much closer to the sphere-packing exponent for asymmetric channels.

In Chapter 3, we studied fixed delay coding without feedback, another problem where the Haroutunian exponent was the best known upper bound to the error exponent. In fixed delay coding without feedback, there is no conceivable method by which a code could tailor its input distribution for the channel behavior it faces because there is no feedback information. Rather, the Haroutunian exponent comes up because little is known about the input distribution over short periods of communication between when an individual bit arrives at the encoder and when it is decoded. Using the result about the Haroutunian exponent for parallel channels, we were able to show that the sphere-packing exponent holds for fixed delay codes. While this is exactly what we wanted to prove, we should hope for more. The proof of Section 3.5 yields only an asymptotic result. Ideally, we would like a lower bound to the error probability for fixed delay codes with good constant factors, i.e., something of the form

$$P_e(d,R) \geq \exp\left(-d\left[E_{sp}\left(R - O\left(\frac{\log d}{d}\right)\right) + O\left(\sqrt{\frac{\log d}{d}}\right)\right]\right).$$

Such a lower bound could then be turned around to give an upper bound on the maximum rate of information that can be communicated for a given error probability and delay, akin to the perspective of Polyanskiy et al. [11] for block coding.

Finally, the Haroutunian exponent shows up as the best known upper bound to the error exponent for message-passing decoding on VLSI circuitry [38]. Once again, like fixed delay coding, one gets the Haroutunian bound because the local input type for the 'neighborhood' of channel outputs that a bit's decoder sees is not known. The parallel channel trick used for fixed delay codes does not work in this case, so it would seem that a new approach to proving the sphere-packing bound for asymmetric channels must be devised.

The second part of the thesis continued the study of the source coding game and arbitrarily varying sources. In Berger's paper [4], the question was asked of what happens to the rate-distortion function for the adversarial arbitrarily varying source if the switcher has noncausal knowledge of the memoryless subsources' realizations. In Chapter 4, we characterized the $R(D)$ function for this case, and extended it to allow for the switcher to have both noisy and noncausal knowledge of subsource realizations. The characterization showed that noncausal knowledge allowed the switcher to enlarge the set of distributions the output of the AVS could 'simulate'. We saw that knowledge of the future does not increase the $R(D)$ function any more once knowledge of present subsources realizations are given to the switcher before deciding the switch position. We also studied the helpful switcher, who is trying to help the coder use the lowest possible rate to achieve a specified distortion. There, we could only characterize the $R(D)$ function if the switcher was given fully noncausal knowledge of subsource realizations. Finally, as an aside used for the computation of the $R(D)$ function, we proved a result about the uniform continuity of the rate-distortion function analogous to one for entropy.

The big question for the future pertaining to arbitrarily varying sources is what happens when the subsources have memory. Subsources with memory are of practical interest because the sources that motivate the consideration of AVSs the most are video streams in active vision. Dobrushin [54] has analyzed the case of the non-anticipatory AVS composed of independent subsources with memory with different distributions when the switcher is passive and blindly chooses the switch position. In the case of subsources with memory, additional knowledge will no doubt increase the adversary's power to increase the rate-distortion function. If we let $R^{(k)}(D)$ be the rate-distortion function for an AVS composed of subsources with memory and an adversary with k step lookahead, one could imagine that in general,

$$R^{(0)}(D) < R^{(1)}(D) < R^{(2)}(D) < \cdots < R^{(\infty)}(D).$$

Results that can be evaluated with finite dimensional optimizations are especially desirable, but it is not clear if they are attainable.

Appendix A

Block coding appendix

A.1 Definitions and notation for Chapter 2

This section introduces the notation used throughout Chapter 2. However, for those familiar with information theory, a glance through Tables 2.1 and 2.2 might suffice to read through the chapter.

A.1.1 Alphabets, channels and distributions

The finite channel input and output alphabets are $\mathcal{X}$ and $\mathcal{Y}$ respectively. We let $x^n = (x_1, \ldots, x_n)$ denote a length n vector[1] from $\mathcal{X}^n$ and similarly for y^n. Uppercase versions of symbols are generally meant to denote random variables. For example X^n is a random vector, while x^n is a fixed deterministic vector. We let $\mathcal{P}$ denote the set of distributions (probability mass functions) on $\mathcal{X}$ and $\mathcal{Q}$ the set of distributions on $\mathcal{Y}$. A distribution P on $\mathcal{X}$ is a vector of non-negative real numbers such that $\sum_x P(x) = 1$.

A channel, or more formally, a channel transition matrix W, from $\mathcal{X}$ to $\mathcal{Y}$ is a non-negative real matrix with $\mathcal{X}$ rows and $\mathcal{Y}$ columns with values denoted by $W(y|x)$ for each $x \in \mathcal{X}, y \in \mathcal{Y}$. The rows of W are distributions on $\mathcal{Y}$, that is, $\sum_y W(y|x) = 1$ for all $x \in \mathcal{X}$. We let $\mathcal{W}$ denote the set of channel transition matrices from $\mathcal{X}$ to $\mathcal{Y}$. Occasionally, we will write $V(\cdot|x)$ to refer to the distribution on $\mathcal{Y}$ given by fixing x.

A.1.2 Types and conditional types

The concept and properties of types and conditional types is explained quite well in the textbook of Csiszár and Körner [47], as well as the survey article on the method of types by

[1] Throughout the thesis, when we speak of vectors from $\mathcal{X}$ or $\mathcal{Y}$, we mean only tuples from the alphabet, no formal algebraic vector space definition is implied. The only vector spaces in this thesis are real vector spaces corresponding to the native spaces of distributions on $\mathcal{X}$ and $\mathcal{Y}$.

Csiszar [32]. Here, we only define the notation as used in this thesis, assuming the reader is familiar with the terms. We let $\mathcal{P}_n \subset \mathcal{P}$ be the set of types of $\mathcal{X}^n$, i.e. the set of types of length n vectors from $\mathcal{X}$. For a type $P \in \mathcal{P}_n$, T_P denotes the type class of P, i.e. the set of vectors in $\mathcal{X}^n$ with type P. Similarly, $\mathcal{Q}_n \subset \mathcal{Q}$ denotes the set of types of length n for $\mathcal{Y}$.

For a type $P \in \mathcal{P}_n$, we let $\mathcal{V}_n(P)$ denote the set of conditional types of length n 'compatible' with P. That is, if for each x that $P(x) > 0$, $V(\cdot|x) \in \mathcal{Q}_{nP(x)}$, then $V \in \mathcal{V}_n(P)$. Together, the pair (P, V) defines a joint type on $\mathcal{X}^n \times \mathcal{Y}^n$. For a vector x^n of type $P \in \mathcal{P}_n$ and conditional type $V \in \mathcal{V}_n(P)$, define the V-shell around x^n, $T_V(x^n)$, to be the set of y^n such that

$$\forall\ x, y,\ \frac{|\{i : (x_i = x, y_i = y)\}|}{n} = nP(x)V(y|x).$$

The V-shell for conditional types is analogous to the type class for types.

A.1.3 Functions on distributions and channels

Throughout this thesis, we will use log to denote the logarithm to base e, but the results can apply to any base so long as the exponential function and logarithm are taken to the same base. For a distribution P on an alphabet[2] $\mathcal{X}$, we define the entropy of P to be

$$H(P) = \sum_x P(x) \log \frac{1}{P(x)}.$$

The conditional entropy of a channel V and input distribution P is defined to be

$$H(V|P) = \sum_x P(x) \sum_y V(y|x) \log \frac{1}{V(y|x)} = \sum_x P(x) H(V(\cdot|x)).$$

The mutual information of input distribution P over channel V is defined to be $I(P, V) = H(PV) - H(V|P)$ where PV denotes the distribution on $\mathcal{Y}$ given by

$$\forall\ y \in \mathcal{Y},\ (PV)(y) = \sum_x P(x)V(y|x).$$

If X and Y are random variables on $\mathcal{X}$ and $\mathcal{Y}$ respectively, the mutual information $I(X; Y)$ is equal to the mutual information $I(P, V)$ where P is the distribution of X and $V(y|x) = \mathbb{P}(Y = y|X = x)$ is the conditional probability. Throughout the thesis, the symbol $\mathbb{P}$ denotes a relevant probability measure, which should be clear from context. The divergence[3] between two distributions P and $\widetilde{P}$ on $\mathcal{X}$ is

$$D(P||\widetilde{P}) = \sum_x P(x) \log \frac{P(x)}{\widetilde{P}(x)}.$$

[2] The alphabet in these definitions is arbitrary, so long as it is finite.

[3] By continuity, $0 \log \frac{0}{0} = 0, 0 \log \frac{1}{0} = \infty, 0 \log \frac{0}{1} = 0$.

The conditional divergence between two channels V and W when the input distribution is P is

$$D(V||W|P) = \sum_x P(x) D(V(\cdot|x)||W(\cdot|x)) = \sum_{x,y} P(x)V(y|x) \log \frac{V(y|x)}{W(y|x)}.$$

The $\mathcal{L}_1$ distance between two distributions $P, P' \in \mathcal{P}$ is

$$||P - P'||_1 = \sum_x |P(x) - P'(x)|.$$

A.1.4 Fixed blocklength channel codes

In our model, the communication medium is assumed to be a stationary discrete memoryless channel (DMC) W. That means that at any time $j \in \mathbb{N}$,

$$\forall\ x \in \mathcal{X}, y \in \mathcal{Y},\ \mathbb{P}_W(Y_j = y|X_j = x) = W(y|x),$$

where the subscript in $\mathbb{P}_W$ is used to denote that the channel is W. Further the memoryless portion of the definition can be taken to mean that for any $j \in \mathbb{N}$,

$$\forall\ x \in \mathcal{X}, y \in \mathcal{Y}, y^{j-1} \in \mathcal{Y}^{j-1},\ \mathbb{P}_W(Y_j = y|X_j = x, Y^{j-1} = y^{j-1}) = W(y|x).$$

A fixed blocklength channel code can either have feedback or not have feedback. In either case, a fixed blocklength code has a blocklength n, message set $\mathcal{M}$ and decoding regions $\{\mathcal{D}_m\}_{m \in \mathcal{M}}$. Without loss of generality (WLOG), $\mathcal{M} = \{1, \ldots, |\mathcal{M}|\}$, where $|\mathcal{M}|$ is referred to as the message set size or code size. The rate of the code is

$$R = \frac{1}{n} \log |\mathcal{M}|.$$

The decoding regions, $\mathcal{D}_m, m \in \mathcal{M}$, are disjoint subsets of $\mathcal{Y}^n$ that cover $\mathcal{Y}^n$. If $Y^n = y^n$, the decoder is assumed to produce a 'decoded message' of $\widehat{M} = m$, where $y^n \in \mathcal{D}_m$.

A fixed blocklength n code without feedback consists of $|\mathcal{M}|$ codewords in $\mathcal{X}^n$. The codeword for message m is denoted by $\phi(m) \in \mathcal{X}^n$ or equivalently, $\phi(m) = x^n(m) = (x_1(m), x_2(m), \ldots, x_n(m))$, where $x_i(m)$ denotes the input symbol at time i when the message is m. For a given fixed blocklength n code without feedback, message m, and channel output vector y^n,

$$\mathbb{P}_W\left(Y^n = y^n | M = m\right) = \prod_{j=1}^{n} W(y_j|x_j(m)),$$

where $M = m$ denotes conditioning on the event that the random message M is equal to m.

A fixed blocklength n code with (noiseless, delayless output) feedback[4] consists of $|\mathcal{M}|$ encoding functions or encoding trees, which tell the encoder what input symbol to use depending on the message *and received channel outputs*. For each m, at time i, if the past received channel symbols are y^{i-1}, then input is denoted $x_i(m, y^{i-1})$. For the whole block, we use $x^n(m, y^n) = (x_1(m), x_2(m, y^1), \ldots, x_n(m, y^{n-1}))$ to denote the input vector when the message is m and the output vector is y^n (note that the input vector does not actually depend on the last received symbol y_n with this notation). We let $P(m, y^n)$ denote the type of the input $x^n(m, y^n)$ when the message is m and received output vector is y^n. We let $V(m, y^n)$ denote the conditional type whose V-shell y^n lies in when the input is $x^n(m, y^n)$. For a $P \in \mathcal{P}_n$, $U \in \mathcal{V}_n(P)$, let for each $m \in \mathcal{M}$,

$$B(m, P, U) = \{y^n : P(m, y^n) = P, V(m, y^n) = U\}.$$

For a fixed blocklength n code with feedback, channel output vector y^n and message m,

$$\mathbb{P}_W(Y^n = y^n | M = m) = \prod_{i=1}^{n} W(y_i | x_i(m, y^{i-1})).$$

Note that the definitions of $P(m, y^n)$, $V(m, y^n)$ and $B(m, P, U)$ are well defined for fixed-length block codes without feedback as well. If the code does not have feedback, $P(m, y^n)$ and $V(m, y^n)$ do not depend on y^n.

In the remainder of the thesis, we will refer to fixed blocklength channel codes as block codes and fixed-length codes interchangeably. We will generally qualify if a result applies to block codes with or without feedback. Of course, any block code without feedback is trivially also a block code with feedback. For a given block code (either with or without feedback), the (average) error probability[5] under channel V is

$$P_e(V) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_V(Y^n \notin \mathcal{D}_m | M = m),$$

and the (average) probability of correct reception under channel V is

$$P_c(V) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_V(Y^n \in \mathcal{D}_m | M = m).$$

[4] Later in the chapter, we will discuss the notation for fixed blocklength codes with noiseless, delayed feedback.

[5] All the results for block codes in this thesis pertain to average error probability. Maximum error probability is another criterion that leads to the same error exponent results for the block coding problems studied in this thesis.

A.2 Sphere-packing without feedback via method of types

Theorem 1 (first part) of Section 2.3: Fix a $\delta > 0$. There is a finite $n_{SP}(W, \delta)$ such that for $n \geq n_{SP}(W, \delta)$, any block code of length n and rate R without feedback has

$$-\frac{1}{n} \log P_e(W) \leq E_{sp}(R - \delta) + \delta.$$

Proof: By basic properties of types, we know that $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$. We have

$$|\mathcal{M}| = \left| \bigcup_{P \in \mathcal{P}_n} \mathcal{M}(P) \right| = \sum_{P \in \mathcal{P}_n} |\mathcal{M}(P)|,$$

where $\mathcal{M}(P) = \{m \in \mathcal{M} : \phi(m) \in T_P\}$. By the pigeon-hole principle, there must then exist a $P \in \mathcal{P}_n$ (depending on the code) such that

$$|\mathcal{M}(P)| \geq \frac{|\mathcal{M}|}{(n+1)^{|\mathcal{X}|}}.$$

Now, we have

$$\begin{aligned}
P_e(W) &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}(Y^n \notin \mathcal{D}_m | X^n = \phi(m)) \\
&\overset{(a)}{\geq} \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}(P)} \mathbb{P}(Y^n \notin \mathcal{D}_m | X^n = \phi(m)) \qquad \text{(A.1)} \\
&\overset{(b)}{=} \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}(P)} \sum_{V \in \mathcal{V}_n(P)} \sum_{y^n \in T_V(\phi(m)) \cap \mathcal{D}_m^c} \mathbb{P}(Y^n = y^n | X^n = \phi(m)) \\
&\overset{(c)}{=} \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}(P)} \sum_{V \in \mathcal{V}_n(P)} \sum_{y^n \in T_V(\phi(m)) \cap \mathcal{D}_m^c} \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \\
&\overset{(d)}{=} \frac{1}{|\mathcal{M}|} \sum_{V \in \mathcal{V}_n(P)} \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \sum_{m \in \mathcal{M}(P)} \sum_{y^n \in T_V(\phi(m)) \cap \mathcal{D}_m^c} 1 \\
&\overset{(e)}{=} \frac{1}{|\mathcal{M}|} \sum_{V \in \mathcal{V}_n(P)} \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \sum_{m \in \mathcal{M}(P)} \left|T_V(\phi(m)) \cap \mathcal{D}_m^c\right|. \qquad \text{(A.2)}
\end{aligned}$$

In the above, (a) is due to the sum being over fewer non-negative terms, (b) expands the probability by summing over individual output vectors, (c) follows from the properties of V-shells, (d) interchanges sums and (e) notes that the innermost sum in (d) is simply the size of a set. Now, we need to understand when the term $\sum_{m \in \mathcal{M}(P)} |T_V(\phi(m)) \cap \mathcal{D}_m^c|$ is large.

Now, $\forall\, m \in \mathcal{M}(P)$, we know $|T_V(\phi(m))| = |T_V(\phi(m)) \cap \mathcal{D}_m| + |T_V(\phi(m)) \cap \mathcal{D}_m^c|$, so it follows that

$$\begin{aligned}\sum_{m\in\mathcal{M}(P)} |T_V(\phi(m)) \cap \mathcal{D}_m^c| &= \sum_{m\in\mathcal{M}(P)} |T_V(\phi(m))| - |T_V(\phi(m)) \cap \mathcal{D}_m| \\ &\overset{(a)}{\geq} \frac{|\mathcal{M}(P)| \exp(nH(V|P))}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} - \sum_{m\in\mathcal{M}(P)} |T_V(\phi(m)) \cap \mathcal{D}_m|. \qquad \text{(A.3)}\end{aligned}$$

In the above line, (a) follows by standard properties of V-shells. At this point, note that for all $m \in \mathcal{M}(P)$, $T_V(\phi(m)) \cap \mathcal{D}_m$ are disjoint sets because $\mathcal{D}_m$ are disjoint. Furthermore, if $m \in \mathcal{M}(P)$, $y^n \in T_V(\phi(m))$ implies that $y^n \in T_{PV}$ also, where T_{PV} is the type class of y^n vectors that have type PV. Therefore, we have

$$\begin{aligned}\sum_{m\in\mathcal{M}(P)} |T_V(\phi(m)) \cap \mathcal{D}_m| &= \left| \bigcup_{m\in\mathcal{M}(P)} T_V(\phi(m)) \cap \mathcal{D}_m \right| \\ &\leq |T_{PV}| \\ &\leq \exp(nH(PV)).\end{aligned}$$

Hence, plugging the above inequality into (A.3), we have

$$\begin{aligned}\sum_{m\in\mathcal{M}(P)} |T_V(\phi(m)) \cap \mathcal{D}_m^c| &\geq \frac{|\mathcal{M}(P)| \exp(nH(V|P))}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} - \exp(nH(PV)) \\ &\overset{(a)}{\geq} \frac{\exp(n(R + H(V|P)))}{(n+1)^{|\mathcal{X}|(1+|\mathcal{Y}|)}} - \exp(nH(PV)) \\ &\overset{(b)}{=} \exp\left(n\left(R + H(V|P) - \frac{|\mathcal{X}|(1+|\mathcal{Y}|)}{n}\log(n+1)\right)\right) \times \\ &\quad \left[1 - \exp\left(-n\left(R - I(P,V) - \frac{|\mathcal{X}|(1+|\mathcal{Y}|)}{n}\log(n+1)\right)\right)\right].\end{aligned}$$

In the above, (a) follows from the fact that $|\mathcal{M}(P)|$ is almost as large as $|\mathcal{M}|$ (at least exponentially), and (b) follows by the fact that $I(P,V) = H(PV) - H(V|P)$. We want the term in brackets above to be close to 1, so we restrict the choice of V to those V that have

$I(P,V) \leq R - \delta$ for some small $\delta > 0$. Plugging this into (A.2) gives

$$\begin{aligned} P_e(W) &\geq \frac{1}{|\mathcal{M}|} \sum_{V \in \mathcal{V}_n(P): I(P,V) \leq R-\delta} \exp\left(-n\left(D(V||W|P) - R + \frac{|\mathcal{X}|(1+|\mathcal{Y}|)}{n}\log(n+1)\right)\right) \times \\ &\qquad \left[1 - \exp\left(-n\left[\delta - \frac{|\mathcal{X}|(1+|\mathcal{Y}|)}{n}\log(n+1)\right]\right)\right] \\ &= \sum_{V \in \mathcal{V}_n(P): I(P,V) \leq R-\delta} \exp\left(-n\left(D(V||W|P) + \frac{|\mathcal{X}|(1+|\mathcal{Y}|)}{n}\log(n+1)\right)\right) \times \\ &\qquad \left[1 - \exp\left(-n\left[\delta - \frac{|\mathcal{X}|(1+|\mathcal{Y}|)}{n}\log(n+1)\right]\right)\right]. \qquad (A.4) \end{aligned}$$

Now, for n large enough, depending on $|\mathcal{X}|, |\mathcal{Y}|$ and δ, we have that

$$(n+1)^{-|\mathcal{X}|(1+|\mathcal{Y}|)} \left[1 - \exp\left(-n\left[\delta - \frac{|\mathcal{X}|(1+|\mathcal{Y}|)}{n}\log(n+1)\right]\right)\right] \geq \exp(-n\delta),$$

so plugging into (A.4), we have that for n large enough,

$$\begin{aligned} P_e(W) &\geq \exp\left(-n\left[\min_{V \in \mathcal{V}_n(P): I(P,V) \leq R-\delta} D(V||W|P) + \delta\right]\right) \\ &\overset{(a)}{\geq} \exp\left(-n\left[\max_{P \in \mathcal{P}_n} \min_{V \in \mathcal{V}_n(P): I(P,V) \leq R-\delta} D(V||W|P) + \delta\right]\right). \end{aligned}$$

In the above, (a) follows because we do not know the most populous codeword type *a priori*, so we take the most optimistic one in terms of the exponent. Now, we have a term that depends on n that looks like the sphere-packing exponent. Lemma 6 of Appendix A.12.2 tells us that

$$\begin{aligned} \max_{P \in \mathcal{P}_n} \min_{V \in \mathcal{V}_n(P): I(P,V) \leq R-\delta} D(V||W|P) &\leq \max_{P \in \mathcal{P}} \min_{V \in \mathcal{W}: I(P,V) \leq R - \delta - \frac{2|\mathcal{X}||\mathcal{Y}|}{n}\log(n)} D(V||W|P) \\ &\qquad + \kappa_W \frac{|\mathcal{X}||\mathcal{Y}|}{n} + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log\left(\frac{n}{|\mathcal{X}|}\right) \\ &= E_{sp}\left(R - \delta - \frac{2|\mathcal{X}||\mathcal{Y}|}{n}\log n\right) + \\ &\qquad \kappa_W \frac{|\mathcal{X}||\mathcal{Y}|}{n} + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log\left(\frac{n}{|\mathcal{X}|}\right), \end{aligned}$$

where

$$\kappa_W = \max_{x,y: W(y|x) > 0} \log \frac{1}{W(y|x)}.$$

Hence, for n large enough depending[6] on κ_W, $|\mathcal{X}|, |\mathcal{Y}|$ and δ, we have

$$P_e(W) \geq \exp\left(-n\left[E_{sp}(R-2\delta)+2\delta\right]\right).$$

A.3 Sphere-packing without feedback via change of measure

Theorem 1 (second part) of Section 2.3: For any $\delta > 0$, there exists a finite $n_{sp}(W, \delta)$ such that any block code without feedback of rate at least R and length $n \geq n_{sp}(W, \delta)$ has

$$-\frac{1}{n}\log P_e(W) \leq E_{sp}\left(R-\delta\right)+\delta.$$

Proof: We will use a change-of-measure approach to prove that a high error probability under channel V implies an error probability under channel W involving a divergence term that is exponentially decaying with the sphere-packing exponent after proper selection of V. To do so, we will use Lemma 12 in this Appendix and a lemma about what happens to the error probability when we change measures. First, restrict attention to the largest fixed-type subcode of the code being bounded. That is, assume that there is a $P \in \mathcal{P}_n$ such that $\phi(m) \in T_P$ for all $m \in \mathcal{M}$. We know from Lemma 12 that if all codewords have the same type P and $R - I(P, V) \geq \delta > 0$,

$$P_c(V) \leq \gamma_{SC}(n, \delta, |\mathcal{X}|, |\mathcal{Y}|)$$

with γ_{SC} being defined in Lemma 12. The critical feature of γ_{SC} is that $\gamma_{SC}(n, \delta, |\mathcal{X}|, |\mathcal{Y}|) \to 0$ as $n \to \infty$. Now we can apply Lemma 18 of Appendix A.14 to get (since all codewords are assumed to be of type $P \in \mathcal{P}_n$),

$$P_e(W) \geq P_e(V)\exp\left(-n\left[D(V||W|P)+\frac{\max\{\kappa_V,\kappa_W\}}{P_e(V)}\beta(n,|\mathcal{X}|,|\mathcal{Y}|)\right]\right),$$

where

$$\beta(n,|\mathcal{X}|,|\mathcal{Y}|) = \inf_{\epsilon>0}\epsilon+(n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-n\frac{\epsilon^2}{2}\right).$$

Therefore,

$$\begin{aligned} P_e(W) \geq{}& (1-\gamma_{SC}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|))\times \\ &\exp\left(-n\left[D(V||W|P)+\frac{\max\{\kappa_V,\kappa_W\}}{1-\gamma_{SC}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|)}\beta(n,|\mathcal{X}|,|\mathcal{Y}|)\right]\right). \end{aligned}$$

[6] Note that $|\mathcal{X}|$, $|\mathcal{Y}|$ and κ_W are all quantities that can be derived from W.

At this point, we remove the restriction that all message codewords have to be of the same type. Since there are at most $(n+1)^{|\mathcal{X}|}$ types of length n for $\mathcal{X}^n$, there is at least one type P that has at least $|\mathcal{M}|/(n+1)^{|\mathcal{X}|}$ codewords. We can apply the previous argument to lower bound $P_e(W)$ for this subset of messages. This adds a penalty of $|\mathcal{X}|\log(n+1)/n$ to the probability term (to account for the thinning of the code in order to make the preceding argument). Furthermore, we must assume that $R - |\mathcal{X}|\log(n+1)/n - I(P,V) = \delta > 0$ in order to apply Lemma 12. In the limit as $n \to \infty$, however, these two penalties can be absorbed into slack parameters in the exponent. Taking these two points into consideration yields

$$P_e(W) \geq \frac{(1-\gamma_{SC}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|))}{(n+1)^{|\mathcal{X}|}} \times \\ \exp\left(-n\left[D(V||W|P) + \frac{\max\{\kappa_V,\kappa_W\}}{1-\gamma_{SC}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|)}\beta(n,|\mathcal{X}|,|\mathcal{Y}|)\right]\right),$$

where implicitly, we are now restricted to choices of V such that $I(P,V) \leq R - \delta - |\mathcal{X}|\log(n+1)/n$. Optimizing over such V (i.e. setting V to be the sphere-packing optimizing V^* for type P at rate $R - \delta - \frac{|\mathcal{X}|}{n}\log(n+1)$) yields

$$P_e(W) \geq \exp\left(-n\left[E_{sp}\left(R - \delta - \frac{|\mathcal{X}|}{n}\log(n+1)\right) + \frac{\max\{\kappa_{V^*},\kappa_W\}}{1-\gamma_{SC}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|)}\beta(n,|\mathcal{X}|,|\mathcal{Y}|)\right]\right) \times \\ \frac{(1-\gamma_{SC}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|))}{(n+1)^{|\mathcal{X}|}}.$$

Lemma 1 shows that $\kappa_{V^*} \leq \kappa_W + \log|\mathcal{Y}|$, and we know that $\beta(n,|\mathcal{X}|,|\mathcal{Y}|) = O\left(\sqrt{\frac{\log n}{n}}\right)$, so

$$P_e(W) \geq \exp\left(-n\left[E_{sp}\left(R - \delta - \frac{|\mathcal{X}|}{n}\log(n+1)\right) + \frac{\kappa_W + \log|\mathcal{Y}|}{1-\gamma_{SC}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|)}O\left(\sqrt{\frac{\log n}{n}}\right)\right]\right) \times \\ \frac{(1-\gamma_{SC}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|))}{(n+1)^{|\mathcal{X}|}}.$$

Since $\frac{1}{n}\log(n+1) \to 0$ and $\gamma_{SC}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|) \to 0$, it follows that for n large enough, depending on κ_W, $|\mathcal{X}|$, $|\mathcal{Y}|$ and δ,

$$-\frac{1}{n}\log P_e(W) \leq E_{sp}(R-2\delta) + 2\delta.$$

Lemma 12 (Strong converse without feedback)**.** *Fix a block code without feedback of length n. Suppose there is a $P \in \mathcal{P}_n$ such that $\forall\ m \in \mathcal{M}$, $\phi(m) \in T_P$. Fix an $\epsilon \in (0, 1/2)$. Then,* *for any* $W \in \mathcal{W}$*,*

$$P_c(W) \leq \exp\left(-n\left[\frac{\epsilon^2}{2} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) + \\ \exp\left(-n\left[R - I(P,W) - 2\epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} - \frac{|\mathcal{Y}|}{n}\log(n+1)\right]\right).$$

Therefore, we have

$$P_c(W) \leq \gamma_{SC}(n, R - I(P,W), |\mathcal{X}|, |\mathcal{Y}|),$$

where

$$\gamma_{SC}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|) = \inf_{\epsilon\in(0,1/2)}\left[\exp\left(-n\left[\frac{\epsilon^2}{2} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) + \\ \exp\left(-n\left[\delta - 2\epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} - \frac{|\mathcal{Y}|}{n}\log(n+1)\right]\right)\right].$$

Note that $\gamma_{SC}(n, R - I(P,W), |\mathcal{X}|, |\mathcal{Y}|) \to 0$ *as* $n \to \infty$ *for a fixed* $R - I(P,W) > 0$*.*

Proof: Fix an $\epsilon \in (0, 1/2)$. For each m, define a typical set

$$A_{m,\epsilon}(W) = \{y^n : y^n \in T_V(\phi(m)), V \in \mathcal{V}_n(P), (P,V) \in \mathcal{J}_\epsilon(W)\}$$
$$J_\epsilon(W) = \left\{(P,V) \in \mathcal{P}\times\mathcal{W} : \sum_x P(x)\sum_y |V(y|x) - W(y|x)| \leq \epsilon\right\}.$$

By Lemma 19 of Appendix A.14, we know that for each m,

$$\mathbb{P}_W(Y^n \notin A_{m,\epsilon}(W)|X^n = \phi(m)) \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp(-n\epsilon^2/2).$$

Now,

$$\begin{aligned}
P_c(W) &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m | X^n = \phi(m)) \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W) | X^n = \phi(m)) + \\
&\qquad \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W)^c | X^n = \phi(m)) \\
&\leq \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W) | X^n = \phi(m)) + \\
&\qquad \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \in A_{m,\epsilon}(W)^c | X^n = \phi(m)) \\
&\leq \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W) | X^n = \phi(m)) + \\
&\qquad (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n\frac{\epsilon^2}{2}\right). \qquad \text{(A.5)}
\end{aligned}$$

Considering just the first term in the bound above, we have

$$\begin{aligned}
T_1 &\triangleq \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W) | X^n = \phi(m)) \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{V\in\mathcal{V}_n(P):(P,V)\in\mathcal{J}_\epsilon(W)} \sum_{y^n\in\mathcal{D}_m\cap T_V(\phi(m))} \mathbb{P}_W(Y^n = y^n | X^n = \phi(m)) \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{V\in\mathcal{V}_n(P):(P,V)\in\mathcal{J}_\epsilon(W)} \sum_{y^n\in\mathcal{D}_m\cap T_V(\phi(m))} \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{V\in\mathcal{V}_n(P):(P,V)\in\mathcal{J}_\epsilon(W)} |\mathcal{D}_m \cap T_V(\phi(m))| \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \\
&= \sum_{V\in\mathcal{V}_n(P):(P,V)\in\mathcal{J}_\epsilon(W)} \exp\left(-n\left[D(V||W|P) + H(V|P) + R\right]\right) \sum_{m\in\mathcal{M}} |\mathcal{D}_m \cap T_V(\phi(m))| \\
&\leq \sum_{V\in\mathcal{V}_n(P):(P,V)\in\mathcal{J}_\epsilon(W)} \exp\left(-n\left[H(V|P) + R\right]\right) \sum_{m\in\mathcal{M}} |\mathcal{D}_m \cap T_V(\phi(m))|.
\end{aligned}$$

Now, we note that for all V such that $(P,V) \in \mathcal{J}_\epsilon(W)$, $H(V|P)$ is close to $H(W|P)$, as shown in Proposition 15, so we have

$$
\begin{aligned}
T_1 &\le \sum_{V\in\mathcal{V}_n(P):(P,V)\in\mathcal{J}_\epsilon(W)} \exp\left(-n\left[H(W|P) - \epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} + R\right]\right)\sum_{m\in\mathcal{M}}|\mathcal{D}_m\cap T_V(\phi(m))| \\
&= \exp\left(-n\left[H(W|P) - \epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} + R\right]\right)\sum_{V\in\mathcal{V}_n(P):(P,V)\in\mathcal{J}_\epsilon(W)}\sum_{m\in\mathcal{M}}|\mathcal{D}_m\cap T_V(\phi(m))| \\
&= \exp\left(-n\left[H(W|P) - \epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} + R\right]\right)|A_\epsilon(W)|,
\end{aligned}
$$

where we have defined the union of the message-typical sets to be

$$A_\epsilon(W) = \bigcup_{m\in\mathcal{M}} A_{m,\epsilon}(W).$$

From Proposition 7, we know that

$$|A_\epsilon(W)| \le (n+1)^{|\mathcal{Y}|}\exp\left(n\left[H(PW) + \epsilon\log\frac{|\mathcal{Y}|}{\epsilon}\right]\right).$$

Hence,

$$
\begin{aligned}
T_1 &\le (n+1)^{|\mathcal{Y}|}\exp\left(-n\left[H(W|P) - H(PW) - 2\epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} + R\right]\right) \\
&= (n+1)^{|\mathcal{Y}|}\exp\left(-n\left[R - I(P,W) - 2\epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon}\right]\right).
\end{aligned}
$$

Plugging the bound above back into equation (A.5) gives the result of the lemma.

Proposition 7. *Let PW denote the distribution induced on $\mathcal{Y}$ by the distribution P on $\mathcal{X}$ with channel W, that is $(PW)(y) = \sum_x P(x)W(y|x)$. Then,*

$$|A_\epsilon(W)| \le (n+1)^{|\mathcal{Y}|}\exp\left(n\left[H(PW) + \epsilon\log\frac{|\mathcal{Y}|}{\epsilon}\right]\right).$$

Proof: If $y^n \in A_{m,\epsilon}(W)$, then $y^n \in T_V(\phi(m))$ for some $V \in \mathcal{V}_n(P)$ with

$$\sum_x P(x)\sum_y |V(y|x) - W(y|x)| \le \epsilon.$$

This also implies that $y^n \in T_{PV} \subset \mathcal{Y}^n$, where $PV \in \mathcal{Q}_n$. Now, since V is not far from W, we can show that PV is not far from PW.

$$
\begin{aligned}
\sum_y |(PV)(y) - (PW)(y)| &= \sum_y\left|\sum_x P(x)V(y|x) - P(x)W(y|x)\right| \\
&\le \sum_x P(x)\sum_y |V(y|x) - W(y|x)| \\
&\le \epsilon.
\end{aligned}
$$

Since we have assumed that $\epsilon \in (0, 1/2)$, we can apply Lemma 20 of Appendix A.14 to deduce that

$$|H(PV) - H(PW)| \leq \epsilon \log \frac{|\mathcal{Y}|}{\epsilon}.$$

Hence, if $y^n \in A_\epsilon(W)$, we also have $y^n \in T_Q$ for some $Q \in \mathcal{Q}_n$ with $|H(Q) - H(PW)| \leq \epsilon \log \frac{|\mathcal{Y}|}{\epsilon}$. Therefore, by standard properties of types,

$$\begin{aligned} |A_\epsilon(W)| &= \bigcup_{m \in \mathcal{M}} A_{m,\epsilon}(W) \\ &\leq \bigcup_{Q \in \mathcal{Q}_n} |T_Q| \\ &\leq (n+1)^{|\mathcal{Y}|} \exp\left(n \left[H(PW) + \epsilon \log \frac{|\mathcal{Y}|}{\epsilon}\right]\right). \end{aligned}$$

A.4 Haroutunian exponent for fixed-length codes with feedback

Theorem 2 of Section 2.3: For any $\delta > 0$, there exists a finite $n_h(W, R, \delta)$ such that any fixed-length code with feedback of rate R and length $n \geq n_h(W, R, \delta)$ has

$$-\frac{1}{n} \log P_e(W) \leq E_h(R - \delta) + \delta.$$

Proof: We will use Lemma 13 of this Appendix to show that if the capacity of a test channel V is too small, the error probability will be high. Then, we will use a change-of-measure argument to show that the error probability under channel W will be the error probability under V multiplied by an exponentially decaying divergence term that corresponds to the Haroutunian exponent. Fix a test channel V. We know from Lemma 13 that if $R - C(V) = \delta > 0$,

$$P_c(V) \leq \gamma_{SC,fb}(n, \delta, |\mathcal{X}|, |\mathcal{Y}|)$$

with $\gamma_{SC,fb}$ being defined in Lemma 13. The critical feature of $\gamma_{SC,fb}$ is that $\gamma_{SC,fb}(n, \delta, |\mathcal{X}|, |\mathcal{Y}|) \to 0$ as $n \to \infty$. At this point, we apply Lemma 18 to get

$$P_e(W) \leq P_e(V) \exp\left(-n \left[\max_P D(V||W|P) + \frac{\max\{\kappa_V, \kappa_W\}}{P_e(V)} \beta(n, |\mathcal{X}, |\mathcal{Y}|)\right]\right),$$

where

$$\beta(n, |\mathcal{X}|, |\mathcal{Y}|) = \inf_{\epsilon > 0} \epsilon + (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n \frac{\epsilon^2}{2}\right).$$

If we let V be V^*, where

$$V^* \in \arg \min_{V: C(V) \leq R-\delta} \max_{P} D(V||W|P),$$

then

$$P_e(W) \geq \exp\left(-n\left[E_h(R-\delta) + \frac{\max\{\kappa_{V^*}, \kappa_W\}}{1-\gamma_{SC,fb}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|)}\beta(n,|\mathcal{X},|\mathcal{Y}|)\right]\right) \times \\ (1-\gamma_{SC,fb}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|)).$$

Now, $\beta(n,|\mathcal{X}|,|\mathcal{Y}|) = O\left(\sqrt{\frac{\log n}{n}}\right)$, and $\gamma_{SC,fb}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|) \to 0$, so for n large enough depending on W, R and δ,

$$-\frac{1}{n}\log P_e(W) \leq E_h(R-\delta) + \delta.$$

Lemma 13. *Let $C(W) = \max_{P\in\mathcal{P}} I(P,W)$ be the capacity of W. Then, for a fixed-length block code used with feedback of length n and rate R, and any $\epsilon \in (0,1/2)$,*

$$P_c(w) \leq \exp\left(-n\left[\frac{\epsilon^2}{2} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) + \\ \exp\left(-n\left[R - C(W) - 2\epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right).$$

Therefore,

$$P_c(W) \leq \gamma_{SC,fb}(n, R-C(W), |\mathcal{X}|, |\mathcal{Y}|),$$

where

$$\gamma_{SC,fb}(n,\delta,|\mathcal{X}|,|\mathcal{Y}|) \triangleq \inf_{\epsilon\in(0,1/2)} \exp\left(-n\left[\frac{\epsilon^2}{2} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) + \\ \exp\left(-n\left[\delta - 2\epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right).$$

Note that $\gamma_{SC,fb}(n, R-C(W), |\mathcal{X}|, |\mathcal{Y}|) \to 0$ as $n \to \infty$ if $R - C(W) > 0$ is fixed, and hence the probability of correct reception goes to 0 if $R < C(W)$.

Proof: Fix an $\epsilon \in (0,1/2)$. Define for each m, a typical set for the encoding tree of message m,

$$A_{m,\epsilon}(W) \triangleq \{y^n : (P(m,y^n), V(m,y^n)) \in \mathcal{J}_\epsilon(W)\}$$

$$\mathcal{J}_\epsilon(W) \triangleq \left\{(P,V) \in \mathcal{P}\times\mathcal{W} : \sum_{x\in\mathcal{X}} P(x)\sum_{y\in\mathcal{Y}} |V(y|x) - W(y|x)| \leq \epsilon\right\}.$$

Recall that $P(m, y^n) \in \mathcal{P}_n$ is the type of input for message m along the output sequence[7] y^n. That is, $P(m, y^n)$ is the type of $x^n(m, y^n) \triangleq (x_1(m), x_2(m, y^1), x_3(m, y^2), \ldots, x_n(m, y^{n-1}))$, where y^k denotes the first k entries of the vector y^n. Accordingly, $V(m, y^n)$ is the conditional shell that y^n lies in when viewed from the input $x^n(m, y^n)$, i.e. $y^n \in T_{V(m,y^n)}(x^n(m, y^n))$. Now,

$$\begin{aligned}
P_c(W) &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m | M = m) \\
&= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W) | M = m) \\
&\qquad + \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W)^c | M = m) \\
&\leq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W) | M = m) \\
&\qquad + \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_W(Y^n \in A_{m,\epsilon}(W)^c | M = m) \\
&\leq \left[\frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W) | M = m) \right] + (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n \frac{\epsilon^2}{2} \right).
\end{aligned}$$

where the last line is arrived at by Lemma 19. Restricting attention to the first term above,

$$\begin{aligned}
T_1 &\triangleq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W) | M = m) \\
&= \exp(-nR) \sum_{m \in \mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W) | M = m). \qquad \text{(A.6)}
\end{aligned}$$

For each m and $P \in \mathcal{P}_n$, $V \in \mathcal{V}_n(P)$, recall the 'conditional shell with feedback'

$$B(m, P, V) \triangleq \{y^n : P(m, y^n) = P, V(m, y^n) = V\}.$$

[7] When feedback is available, the channel input codeword and hence the channel input type depends on both the message and the received channel symbols.

Bounding further using this definition,

$$\begin{aligned}
T_2 &\triangleq \sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m \cap A_{m,\epsilon}(W) | M = m) \\
&= \sum_{m\in\mathcal{M}} \sum_{(P,V)\in\mathcal{J}_\epsilon(W)} \mathbb{P}_W(Y^n \in B(m,P,V) \cap \mathcal{D}_m | M = m) \\
&\overset{(a)}{=} \sum_{m\in\mathcal{M}} \sum_{(P,V)\in\mathcal{J}_\epsilon(W)} |B(m,P,V) \cap \mathcal{D}_m| \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \\
&= \sum_{(P,V)\in\mathcal{J}_\epsilon(W)} \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \sum_{m\in\mathcal{M}} |B(m,P,V) \cap \mathcal{D}_m| \\
&\overset{(b)}{\leq} \sum_{(P,V)\in\mathcal{J}_\epsilon(W)} \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \exp(nH(PV)) \\
&\overset{(c)}{\leq} \sum_{(P,V)\in\mathcal{J}_\epsilon(W)} \exp(nI(P,V)), \qquad \text{(A.7)}
\end{aligned}$$

where in the above, (a) follows from Proposition 14b, (b) follows from Proposition 14a and the fact that the decoding regions are disjoint, and (c) follows from divergence being non-negative and $I(P,V) = H(PV) - H(V|P)$. From Proposition 15, we know that $I(P,V) \leq I(P,W) + 2\epsilon \log(|\mathcal{X}||\mathcal{Y}|/\epsilon)$ for $(P,V) \in \mathcal{J}_\epsilon(W)$, so plugging (A.7) into (A.6), and taking the max over $P \in \mathcal{P}$ yields

$$\begin{aligned}
T_1 &\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n\left[R - \max_{P\in\mathcal{P}} I(P,W) - 2\epsilon \log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon}\right]\right) \\
&= \exp\left(-n\left[R - C(W) - 2\epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right).
\end{aligned}$$

So, for any $\epsilon \in (0, 1/2)$, we have shown that

$$\begin{aligned}
P_c(W) \leq \exp&\left(-n\left[\frac{\epsilon^2}{2} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) + \\
&\exp\left(-n\left[R - C(W) - 2\epsilon\log\frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right).
\end{aligned}$$

Optimizing over $\epsilon \in (0, 1/2)$ yields the desired result.

A.5 A special family of test channels

Lemma 1 of Section 2.4:

Let $P(V)$ be any continuous function from $\mathcal{W}$ to $\mathcal{P}$, and let for any $V \in \mathcal{W}$,

$$\kappa_V = \max_{x,y:V(y|x)>0} \log(1/V(y|x)).$$

Then, for all $R > 0, \epsilon > 0$,

$$\inf_{V:I(P(V),V)\leq R} D(V||W|P(V)) + \epsilon \max\{\kappa_V, \kappa_W\} \leq E_{sp}(R) + \epsilon(\kappa_W + \log|\mathcal{Y}|).$$

Proof: This result is due to Baris Nakiboglu. It can be used to fill in an apparent gap in the proof of Sheverdyaev's result (although there is another issue with that paper, as noted in Appendix A.7). Recall that the sphere-packing exponent has two forms ([18], [47]):

$$\begin{aligned}
E_{sp}(R) &= \max_{P\in\mathcal{P}} \min_{V:I(P,V)\leq R} D(V||W|P) \\
E_{sp}(R) &= \sup_{\rho\geq 0} E_0(\rho) - \rho R \\
E_0(\rho) &\triangleq \max_{P\in\mathcal{P}} - \log \sum_y \left(\sum_x P(x) W(y|x)^{1/(1+\rho)} \right)^{1+\rho} \\
&= -\log \min_{P\in\mathcal{P}} \sum_y \left(\sum_x P(x) W(y|x)^{1/(1+\rho)} \right)^{1+\rho} \\
&= -\log \min_{P\in\mathcal{P}} f(P,\rho) \\
f(P,\rho) &\triangleq \sum_y \left(\sum_x P(x) W(y|x)^{1/(1+\rho)} \right)^{1+\rho}.
\end{aligned}$$

Let

$$\mathcal{P}^*_\rho \triangleq \arg\min_{P\in\mathcal{P}} f(P,\rho).$$

It is known that $f(P,\rho)$ is convex in $\mathcal{P}$ [16] and therefore, for each $\rho \geq 0$, $\mathcal{P}^*_\rho$ is either a unique distribution in $\mathcal{P}$ or a convex region of distributions in $\mathcal{P}$. Further, for every ρ, it can be shown ([47], Problem 2.5.23(iv)) that $P \in \mathcal{P}^*_\rho$ if and only if, for all $x \in \mathcal{X}$,

$$\sum_y W(y|x)^{1/(1+\rho)} \left[\sum_{x'} P(x') W(y|x')^{1/(1+\rho)} \right]^\rho \geq \sum_y \left(\sum_{x'} P(x') W(y|x')^{1/(1+\rho)} \right)^{1+\rho}, \quad \text{(A.8)}$$

with equality for all x such that $P(x) > 0$.

Proposition 8. *There exists a parametrized family of distributions*

$$\{P_\delta\}_{\delta\geq 0} \subset \mathcal{P}$$

and a continuous, monotone nondecreasing map $g : [0,\infty) \to [0,\infty)$ *such that* *(a)* P_δ *is continuous as a function of* δ

(b) P_δ *satisfies the condition in (A.8) for* $\rho = g(\delta)$

(c) $g(0) = 0$*, and* $\lim_{\delta\to\infty} g(\delta) = \infty$

The idea is that for each ρ, there is either a unique optimizer or a convex set of optimizers for $f(P, \rho)$. The problem is to make sure that we get a continuous family of P_δ that are each optimizers for a ρ. This is a subtle issue because a priori we have no information about how many ρ have optimizers which are actually sets and not unique elements. We will see that this distinction is equivalent to the fact that $E_0(\rho)$ may be nondifferentiable at some points.

First, define for $\rho, \epsilon \geq 0$,

$$\mathcal{P}_{[\rho,\epsilon]} \triangleq \left\{(P, s) \in \mathcal{P} \times [0, \infty) : |f(P, s) - E_0(\rho)| < \epsilon, \exists\ P' \in \mathcal{P}^*_\rho \ s.t.\ ||(P, s) - (P', \rho)||_1 < \epsilon\right\}.$$

Although $f(P, \rho)$ is convex in P and ρ, it is not necessarily convex in both at the same time. Therefore, when taking the minimum over P, it may not be differentiable in ρ. Therefore, even though $E_0(\rho)$ is continuous, monotone nondecreasing and concave [16], it can be nondifferentiable in places. However, there is a way to bridge optimizers when this happens. Define the sets

$$\mathcal{P}^-_{[\rho,\epsilon],\gamma} \triangleq \mathcal{P}^*_\rho \bigcap \left(\bigcap_{s\in(\rho-\gamma,\rho)} \mathcal{P}_{[s,\epsilon]} \right)$$

$$\mathcal{P}^-_\rho \triangleq \lim_{\epsilon\to 0} \lim_{\gamma\to 0} \mathcal{P}^-_{[\rho,\epsilon],\gamma}.$$

Similarly, for the other side, define $\mathcal{P}^+_\rho$. By continuity of $f(P, \rho)$, we know that both $\mathcal{P}^+_\rho$ and $\mathcal{P}^-_\rho$ are non empty. If their intersection is nonempty, we will choose one element from the intersection and call it P_δ for some δ that will be clear with $g(\delta) = \rho$. If not, we will take one element from each, call them P^+_ρ and P^-_ρ and construct a convex combination $\beta P^-_\rho + (1 - \beta)P^+_\rho = P_{\rho,\beta}$. Note that by convexity, $P_{\rho,\beta} \in \mathcal{P}_\rho$. Also, when the intersection between $\mathcal{P}^+_\rho$ and $\mathcal{P}^-_\rho$ is nonempty, P_ρ is continuous, and when the intersection is empty, we can construct an arbitrary positive length continuous segment going from P^-_ρ to P^+_ρ. We use the monotonically nondecreasing function g to account for these gaps.

Now, the only thing that remains to be checked is that the number of points at which $E_0(\rho)$ is nondifferentiable is countable so that such a function g exists. To see this, we recall that $E_0(\rho)$ is continuous and concave. Therefore, its derivative is monotonic and nonincreasing where it exists (assign it to be the upper derivative at points of discontinuity). It is easy to see however, that the set of discontinuities for a monotonic function is countable (otherwise the jumps add up too fast). Therefore, a function g as the proposition requires exists to map the δ to ρ parametrically and in a monotonic way. Note that g will not be one-to-one if $E_0(\rho)$ is ever nondifferentiable.

From now on, for a given $\delta \geq 0$, let $\rho = g(\delta)$ be parametrically defined through δ and g. Define, for each $\delta > 0$, a channel

$$\begin{aligned} V_\delta(y|x) &\triangleq \frac{W(y|x)^{1/(1+\rho)} \left(\sum_{x'} P_\delta(x')W(y|x')^{1/(1+\rho)}\right)^\rho}{\sum_{y'} W(y'|x)^{1/(1+\rho)} \left(\sum_{x'} P_\delta(x')W(y'|x')^{1/(1+\rho)}\right)^\rho} \qquad \text{(A.9)} \\ r(x,\delta) &\triangleq \sum_{y'} W(y'|x)^{1/(1+\rho)} \left(\sum_{x'} P_\delta(x')W(y'|x')^{1/(1+\rho)}\right)^\rho . \end{aligned}$$

Note, by (A.8), we have for all x, δ,

$$\begin{aligned} r(x,\delta) &\geq \sum_y \left(\sum_x P_\delta(x')W(y|x)^{1/(1+\rho)}\right)^{1+\rho} \\ &= \exp\left(-E_0(\rho)\right) \qquad \text{(A.10)} \end{aligned}$$

where the last line follows because P_δ is a member of $\mathcal{P}^*_\rho$. Also, $r(x,\delta) = \exp(-E_0(\rho))$ for x such that $P_\delta(x) > 0$. Now, fix a $\delta \geq 0$ and consider $D(V_\delta||W|P)$ for an arbitrary $P \in \mathcal{P}$.

$$\begin{aligned} D(V_\delta||W|P) &= \sum_{x,y} P(x)V_\delta(y|x) \log \frac{V_\delta(y|x)}{W(y|x)} \\ &\overset{(a)}{=} \sum_{x,y} P(x)V_\delta(y|x) \log \frac{V_\delta(y|x)}{\frac{V_\delta(y|x)^{1+\rho} \exp(-(1+\rho)E_0(\rho))}{\left(\sum_x' P_\delta(x')W(y|x')^{1/(1+\rho)}\right)^{\rho(1+\rho)}}} \\ &\overset{(b)}{\leq} \sum_{x,y} P(x)V_\delta(y|x) \log \left[\left(\frac{\left(\sum_{x'} P_\delta(x')W(y|x')^{1/(1+\rho)}\right)^{1+\rho}}{V_\delta(y|x)e^{-E_0(\rho)}}\right)^\rho e^{E_0(\rho)}\right], \end{aligned}$$

where (a) follows by inversion of (A.9) and (b) from (A.10).

$$\begin{aligned} D(V_\delta||W|P) &\leq E_0(\rho) + \rho \sum_{x,y} P(x)V_\delta(y|x) \log \frac{\left(\sum_{x'} P_\delta(x')W(y|x')^{1/(1+\rho)}\right)^{1+\rho}}{V_\delta(y|x)e^{-E_0(\rho)}} \\ &= E_0(\rho) - \rho \sum_{x,y} P(x)V_\delta(y|x) \log \left[\frac{V_\delta(y|x)}{(PV_\delta)(y)} \times \frac{(PV_\delta)(y)e^{-E_0(\rho)}}{\left(\sum_{x'} P_\delta(x')W(y|x')^{1/(1+\rho)}\right)^{1+\rho}}\right] \\ &= E_0(\rho) - \rho I(P, V_\delta) - \\ &\qquad \rho \sum_{x,y} P(x)V_\delta(y|x) \log \left(\frac{(PV_\delta)(y)}{\left(\sum_{x'} P_\delta(x')W(y|x')^{1/(1+\rho)}\right)^{1+\rho} / e^{-E_0(\rho)}}\right). \qquad \text{(A.11)} \end{aligned}$$

Now, let $Q_\delta(x|y)$ be the conditional distribution defined by

$$Q_\delta(x|y) = \frac{P(x)V_\delta(y|x)}{(PV_\delta)(y)}.$$

Then, picking up from (A.11),

$$\begin{aligned} D(V_\delta||W|P) \leq E_0(\rho) - \rho(I(P, V_\delta)) - \\ \rho \sum_{x,y} P(x)V_\delta(y|x) \log \left(\frac{Q_\delta(x|y)(PV_\delta)(y)}{Q_\delta(x|y) \left(\sum_{x'} P_\delta(x')W(y|x')^{1/(1+\rho)}\right)^{1+\rho} / e^{-E_0(\rho)}} \right). \end{aligned}$$

Now, note that

$$\sum_y \left(\sum_x P_\delta(x)W(y|x)^{1/(1+\rho)} \right)^{1+\rho} = e^{-E_0(\rho)}$$

Therefore,

$$\frac{\left(\sum_{x'} P_\delta(x')W(y|x')^{1/(1+\rho)}\right)^{1+\rho}}{e^{-E_0(\rho)}}$$

is a distribution on $\mathcal{Y}$ since for each y, it is nonnegative and sums to one over $\mathcal{Y}$. Call this distribution $Q_\delta \in \mathcal{Q}$. Then, we have

$$\begin{aligned} D(V_\delta||W|P) &\leq E_0(\rho) - \rho(I(P, V_\delta)) - \\ &\qquad \rho \sum_{x,y} P(x)V_\delta(y|x) \log \left(\frac{Q_\delta(x|y)(PV_\delta)(y)}{Q_\delta(x|y)Q_\delta(y)} \right) \\ &= E_0(\rho) - \rho(I(P, V_\delta)) - \\ &\qquad \rho \sum_{x,y} P(x)V_\delta(y|x) \log \left(\frac{P(x)V_\delta(y|x)}{Q_\delta(x|y)Q_\delta(y)} \right) \\ &\overset{(c)}{\leq} E_0(\rho) - \rho I(P, V_\delta) \\ &\leq \sup_{\rho \geq 0} E_0(\rho) - \rho I(P, V_\delta) \\ &= E_{sp}(I(P, V_\delta)), \end{aligned} \tag{A.12}$$

where (c) follows because the term that was eliminated was a divergence and hence nonnegative. Now,

$$\begin{aligned}\inf_{V:I(P_V,V)\leq R} D(V||W|P_V) &\leq \inf_{V_\delta,\delta\geq 0:I(P_{V_\delta},V_\delta)\leq R} D(V_\delta||W|P_{V_\delta})\\ &\overset{(d)}{\leq} \inf_{V_\delta,\delta\geq 0:I(P_{V_\delta},V_\delta)\leq R} E_{sp}(I(P_{V_\delta},V_\delta))\\ &\overset{(e)}{\leq} E_{sp}\left(\sup_{V_\delta,\delta\geq 0:I(P_{V_\delta},V_\delta)\leq R} I(P_{V_\delta},V_\delta)\right),\end{aligned}$$

where (d) is true because of (A.12) and (e) follows because $E_{sp}(R)$ is a monotonically non-increasing function of R. Now, V_δ is continuous in δ by Proposition 8 and (A.9). By assumption, P_V is continuous in V, so P_{V_δ} is continuous in δ. Further, $I(P,V)$ is continuous in the pair P,V, so $I(P_{V_\delta},V_\delta)$ is continuous in δ. Note that $V_0 = W$ by definition. As $\delta\to\infty$, it follows that (because $\rho = g(\delta)\to\infty$, $V_\delta\to V^*$ where for each x,

$$V^*(y|x) = \frac{1}{|\{y : W(y|x) > 0\}|}.$$

We claim that

$$\begin{aligned}\lim_{\delta\to\infty} C(V_\delta) &= C(V^*)\\ &= \inf\{R : E_{sp}(R) < \infty\}.\end{aligned}$$

That the limit of capacities is equal to the capacity of the limit is clear because capacity is a continuous function of the channel and V_δ converges to V^* for a suitable norm (such as $\mathcal{L}_1$ norm). Also, note that these V_δ are the channels that are sufficient to upper bound for the sphere-packing exponent ([47], Problem 2.5.23a), so $\lim_{\delta\to\infty} C(V_\delta) \leq \inf\{R : E_{sp}(R) < \infty\}$. Conversely, if $E_{sp}(R) = \infty$, since $D(V^*||W|P) < \infty$ for all $P\in\mathcal{P}$, it follows that $C(V_\delta) > R$. Therefore $\lim_{\delta\to\infty} C(V_\delta) = \inf\{R : E_{sp}(R) < \infty\}$.

Now, returning to the quantity of interest $\sup_{\delta\geq 0} I(P_{V_\delta},V_\delta)$, by the intermediate value theorem applied to the continuous function $I(P_{V_\delta},V_\delta)$, whose endpoints are at $I(P,W)$ and $\inf\{R : E_{sp}(R) < \infty\}$, we must have one of two things.

(i) $I(P_{V_\delta},V_\delta) < R$ for all $\delta > 0$

(ii) There is a δ^* such that $I(P_{V_{\delta^*}},V_{\delta^*}) = R$

In case (i), it is true that $I(P,W) < R$, so by Fano's inequality, the error exponent is as small as desired for large enough n. In case (ii), we have

$$\inf_{V:I(P_V,V)\leq R} D(V||W|P_V) \leq E_{sp}(R) + \epsilon\max\left\{\kappa_{V_{\delta^*}},\kappa_W\right\}.$$

Now, note that if $\theta_W = \min_{x,y:W(y|x)>0} W(y|x)$, and x, y is such that $W(y|x) > 0$, we have for any $\delta \geq 0$

$$\begin{aligned} V_\delta(y|x) &\triangleq \frac{W(y|x)^{1/(1+\rho)} \left(\sum_{x'} P_\delta(x') W(y|x')^{1/(1+\rho)}\right)^\rho}{\sum_{y'} W(y'|x)^{1/(1+\rho)} \left(\sum_{x'} P_\delta(x') W(y'|x')^{1/(1+\rho)}\right)^\rho} \\ &\geq \frac{\theta_W \theta_W^{\rho/(1+\rho)}}{\sum_y' 1} \\ &= \frac{\theta_W}{|\mathcal{Y}|} \\ \kappa_{V_\delta} &\leq \kappa_W + \log |\mathcal{Y}|. \end{aligned}$$

which concludes the proof of the lemma. .

A.6 Failed Approaches to proving the sphere packing bound for codes with feedback

A.6.1 A first attempt via Fano's inequality

Lemma 2 of Section 2.4.1: Fix an $R > 0$ and $\delta \in (1/n, R)$. For any block code with feedback of rate R and length n,

$$\begin{aligned} -\frac{1}{n} \log P_e(W) \leq \inf_{V:I(P_V,V)\leq R-\delta} & D(V||W|P_{I,V}) + \\ & \frac{2\max\{\kappa_V, \kappa_W\}}{\frac{1}{R}\left(\delta - \frac{1}{n}\right)} \beta(n, |\mathcal{X}|, |\mathcal{Y}|) + \frac{1}{n} \log\left(\frac{1}{R}\left(\delta - \frac{1}{n}\right)\right), \end{aligned}$$

where

$$\beta(n, |\mathcal{X}|, |\mathcal{Y}|) = \inf_{\epsilon > 0} \epsilon + (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n\frac{\epsilon^2}{2}\right).$$

Proof: Fix a V such that $I(P_V, V) \leq R - \delta$, with $\delta \in (1/n, R)$. From Proposition 9, we know that

$$P_e(V) \geq \frac{1}{R}\left(\delta - \frac{1}{n}\right). \tag{A.13}$$

Then, from Lemma 18 in Appendix A.14,

$$P_e(W) \geq P_e(V) \exp\left(-nd(V, W)\right),$$

where

$$d(V,W) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{y^n \in \mathcal{D}_m^c} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)} \frac{1}{n} \sum_{i=1}^{n} \log \frac{V(y_i | x_i(m, y^{i-1}))}{W(y_i | x_i(m, y^{i-1}))}.$$

Now, recalling that $B(m,P,U) = \{y^n : P(m,y^n) = P, V(m,y^n) = U\}$, we have

$$\begin{aligned} d(V,W) &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}, P \in \mathcal{P}_n, U \in \mathcal{V}_n(P)} \sum_{y^n \in \mathcal{D}_m^c \cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)} \times \\ &\qquad \sum_{x,y} P(x) U(y|x) \log \frac{V(y|x)}{W(y|x)} \\ &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}, P \in \mathcal{P}_n, U \in \mathcal{V}_n(P)} \sum_{y^n \in \mathcal{D}_m^c \cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)} \times \\ &\qquad \left[D(V||W|P) + \sum_{x,y} P(x)(U(y|x) - V(y|x)) \log \frac{V(y|x)}{W(y|x)} \right]. \end{aligned}$$

A little bit of algebra and linearity of $D(V||W|P)$ in P shows that

$$\frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}, P \in \mathcal{P}_n, U \in \mathcal{V}_n(P)} \sum_{y^n \in \mathcal{D}_m^c \cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)} D(V||W|P) = D(V||W|P_{I,V}),$$

so we need to control the difference term. For $\epsilon > 0$, let

$$\mathcal{J}_\epsilon(V) = \left\{ (P,U) \in \mathcal{P} \times \mathcal{W} : \sum_x P(x) \sum_y |U(y|x) - V(y|x)| \leq \epsilon \right\}.$$

Then, splitting the difference term into 'typical' and 'atypical' parts gives

$$\begin{aligned}
T &\triangleq \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M},P\in\mathcal{P}_n,U\in\mathcal{V}_n(P)}\sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)}\frac{\mathbb{P}_V(Y^n=y^n|M=m)}{P_e(V)}\times \\
&\qquad \sum_{x,y}P(x)(U(y|x)-V(y|x))\log\frac{V(y|x)}{W(y|x)} \\
&= T_1+T_2 \\
T_1 &\triangleq \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M},P\in\mathcal{P}_n,U\in\mathcal{V}_n(P):(P,U)\in\mathcal{J}_\epsilon(V)}\sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)}\frac{\mathbb{P}_V(Y^n=y^n|M=m)}{P_e(V)}\times \\
&\qquad \sum_{x,y}P(x)(U(y|x)-V(y|x))\log\frac{V(y|x)}{W(y|x)} \\
T_2 &\triangleq \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M},P\in\mathcal{P}_n,U\in\mathcal{V}_n(P):(P,J)\notin\mathcal{J}_\epsilon(V)}\sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)}\frac{\mathbb{P}_V(Y^n=y^n|M=m)}{P_e(V)}\times \\
&\qquad \sum_{x,y}P(x)(U(y|x)-V(y|x))\log\frac{V(y|x)}{W(y|x)}.
\end{aligned}$$

Now, by Proposition 13, if $(P,U)\in\mathcal{J}_\epsilon(V)$, we have

$$\sum_{x,y}P(x)(U(y|x)-V(y|x))\log\frac{V(y|x)}{W(y|x)}\le\epsilon\max\{\kappa_V,\kappa_W\},$$

for all the terms in T_1 that contribute something non-zero to the sum. That is, if $\mathbb{P}(Y^n=y^n|M=m)=0$, then it does not matter if the difference is infinite. Hence,

$$T_1\le\epsilon\max\{\kappa_V,\kappa_W\}.$$

As for term T_2, since $D(V||W|P_{I,V})<\infty$ by assumption, $\sum_{x,y}P(x)(U(y|x)-V(y|x))\log\frac{V(y|x)}{W(y|x)}<\infty$ for all terms in the sum that are not 0. By Lemma 19, we know that for each $m\in\mathcal{M}$,

$$\sum_{P\in\mathcal{P}_n,U\in\mathcal{V}_n(P):(P,U)\notin\mathcal{J}_\epsilon(V)}\mathbb{P}_V(Y^n\in B(m,P,U)|M=m)\le(n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-n\frac{\epsilon^2}{2}\right).$$

Therefore, because $\sum_y|U(y|x)-V(y|x)|\le 2$ for all x and $U,V\in\mathcal{W}$,

$$T_2\le\frac{2\max\{\kappa_V,\kappa_W\}}{P_e(V)}\exp\left(-n\left[\frac{\epsilon^2}{2}-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right).$$

Combining the bounds for T_1 and T_2 tells us that

$$T\le\frac{2\max\{\kappa_V,\kappa_W\}}{P_e(V)}\left[\epsilon+(n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-n\frac{\epsilon^2}{2}\right)\right].$$

Optimizing over $\epsilon>0$ gives us the result of the lemma with $\beta(n,|\mathcal{X}|,|\mathcal{Y}|)$.

Proposition 9 (Fano's inequality for codes with feedback)**.** *If for some $V \in \mathcal{W}$, $I(P_V, V) \leq R - \delta$ for a rate R length n block code with feedback, then*

$$P_e(V) \geq \frac{1}{R}\left(\delta - \frac{1}{n}\right).$$

Proof: First, we will show that if $I(P_V, V) \leq R - \delta$, then $\frac{1}{n} I(M; Y^n) \leq R - \delta$ also[8]. For two random variables A, B on the product set $\mathcal{A} \times \mathcal{B}$, recall that the mutual information is

$$\begin{aligned} I(A; B) &= \sum_{(a,b) \in \mathcal{A} \times \mathcal{B}} \mathbb{P}(A = a, B = b) \log \frac{\mathbb{P}(A = a, B = b)}{\mathbb{P}(A = a)\mathbb{P}(B = b)} \\ &= H(A) - H(A|B), \end{aligned}$$

where $H(A)$ denotes the entropy of A and $H(A|B)$ denotes the conditional entropy of A given B. Using standard properties of the mutual information of a collection of random variables (see [21]),

$$\begin{aligned} \frac{1}{n} I(M; Y^n) &= \frac{1}{n}\left(H(Y^n) - H(Y^n|M)\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(H(Y_i|Y^{i-1}) - H(Y_i|M, Y^{i-1})\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} I(M; Y_i|Y^{i-1}) \\ &\overset{(a)}{\leq} \frac{1}{n} \sum_{i=1}^{n} I(X_i; Y_i|Y^{i-1}), \end{aligned}$$

where (a) follows because given Y^{i-1}, M and Y_i are conditionally related through X_i. That is,

$$\begin{aligned} \mathbb{P}_V(M = m, X_i = x, Y_i = y|Y^{i-1} = y^{i-1}) = \mathbb{P}_V(M = m|Y^{i-1} = y^{i-1}) \times \\ \left[1(x_i(m, y^{i-1}) = x_i) V(y_i|x_i)\right]. \end{aligned}$$

[8] The probabilities in this proposition are all with respect to the measure induced by V.

Continuing and expanding the conditional mutual information,

$$\begin{aligned}\frac{1}{n}I(M;Y^n) &\leq \frac{1}{n}\sum_{i=1}^{n} I(X_i;Y_i|Y^{i-1})\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{y^{i-1}} \mathbb{P}_V(y^{i-1})I(X_i;Y_i|Y^{i-1}=y^{i-1})\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{y^{i-1}} \mathbb{P}_V(y^{i-1})I(P_{y^{i-1}},V)\\ \forall\ x\in\mathcal{X},\ P_{y^{i-1}}(x) &\triangleq \sum_{m\in\mathcal{M}} \mathbb{P}_V(M=m|Y^{i-1}=y^{i-1})1(x_i(m,y^{i-1})=x).\end{aligned}$$

At this point, we can use concavity of mutual information in the input distribution to show

$$\begin{aligned}\frac{1}{n}I(M;Y^n) &\leq \frac{1}{n}\sum_{i=1}^{n}\sum_{y^{i-1}} \mathbb{P}_V(y^{i-1})I(P_{y^{i-1}},V)\\ &\leq I\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{y^{i-1}} \mathbb{P}_V(y^{i-1})P_{y^{i-1}},V\right).\end{aligned}$$

However, noting that

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{y^{i-1}} \mathbb{P}_V(y^{i-1})P_{y^{i-1}} = P_V,$$

we have shown that $\frac{1}{n}I(M;Y^n) \leq I(P_V,V) \leq R-\delta$ for a code with feedback. Now, by Fano's inequality (see [21], Theorem 2.11.1), we have

$$\begin{aligned}H(M|Y^n) &\leq h_b(P_e(V)) + \log(|\mathcal{M}|-1)P_e(V)\\ &\leq 1 + nRP_e(V).\end{aligned}$$

Now, we also have that

$$\begin{aligned}H(M|Y^n) &= H(M) - I(M;Y^n)\\ &\geq nR - n(R-\delta) = n\delta.\end{aligned}$$

Therefore,

$$\begin{aligned}n\delta &\leq 1 + nRP_e(V)\\ P_e(V) &\geq \frac{1}{R}\left(\delta - \frac{1}{n}\right).\end{aligned}$$

A.6.2 A refined strong converse

Lemma 3 of Section 2.4.2: If the refined strong converse holds, then for any $R > 0$, $\delta \in (0, R)$, there is a finite $n_{RSC}(W, R, \delta)$ such that for any block code with feedback of rate R and length $n \geq n_{RSC}(W, R, \delta)$,

$$-\frac{1}{n} \log P_e(W) \leq E_{sp}(R - \delta) + \delta.$$

Hence, if the refined strong converse holds, so does the sphere-packing bound.

Proof: Fix a V such that $I(P_V, V) \leq R - \delta$. From the proof of Lemma 2, we know that

$$P_e(W) \geq P_e(V) \exp\left(-n\left[D(V||W|P_{I,V}) + \frac{2\max\{\kappa_V, \kappa_W\}}{P_e(V)}\beta(n, |\mathcal{X}|, |\mathcal{Y}|)\right]\right),$$

where $P_{I,V}$ is defined in (2.15) and

$$\beta(n, |\mathcal{X}|, |\mathcal{Y}|) = \inf_{\epsilon > 0} \epsilon + (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n\frac{\epsilon^2}{2}\right).$$

From the assumption that the refined strong converse holds, we have

$$P_e(V) \geq 1 - \gamma_{RSC}(n, \delta, R, \kappa_V).$$

Now, we claim that since the error event has high probability as $\gamma_{RSC}(n, \delta, R, \kappa_V) \to 0$, it must be that P_V is close to $P_{I,V}$ and hence $D(V||W|P_V)$ is close to $D(V||W|P_{I,V})$.

$$\begin{aligned}
||P_V - P_{I,V}||_1 &= \sum_{x \in \mathcal{X}} |P_V(x) - P_{I,V}(x)| \\
&= \sum_{x \in \mathcal{X}} |(P_c(V)P_{C,V}(x) + P_e(V)P_{I,V}(x)) - P_{I,V}(x)| \\
&= \sum_{x \in \mathcal{X}} |P_c(V)P_{C,V}(x) - P_c(V)P_{I,V}(x)| \\
&= P_c(V) \sum_{x \in \mathcal{X}} |P_{C,V}(x) - P_{I,V}(x)| \\
&\leq P_c(V) \sum_{x \in \mathcal{X}} P_{C,V}(x) + P_{I,V}(x) \\
&\leq 2P_c(V) \leq 2\gamma_{RSC}(n, \delta, R, \kappa_V), \qquad \text{(A.14)}
\end{aligned}$$

where the 'correct input distribution' is

$$P_{C,V}(x) \triangleq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{y^n \in \mathcal{D}_m} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_c(V)} \frac{1}{n} \sum_{i=1}^{n} 1(x_i(m, y^{i-1}) = x). \qquad \text{(A.15)}$$

Now,

$$\begin{aligned} D(V||W|P_{I,V}) - D(V||W|P_V) &= \sum_{x,y} P_{I,V}(x)V(y|x)\log\frac{V(y|x)}{W(y|x)} - P_V(x)V(y|x)\log\frac{V(y|x)}{W(y|x)} \\ &= \sum_{x,y}(P_{I,V}(x) - P_V(x))V(y|x)\log\frac{V(y|x)}{W(y|x)} \\ &\leq \sum_{x,y}|P_{I,V}(x) - P_V(x)|V(y|x)\max\{\kappa_V, \kappa_W\} \\ &= ||P_{I,V} - P_V||_1 \max\{\kappa_V, \kappa_W\} \\ &\leq 2\gamma_{RSC}(n,\delta,R,\kappa_V)\max\{\kappa_V,\kappa_W\} \end{aligned}$$

provided that $V(y|x) = 0$ whenever $W(y|x) = 0$. Therefore,

$$D(V||W|P_{I,V}) \leq D(V||W|P_V) + 2\gamma_{RSC}(n,\delta,R,\kappa_V)\max\{\kappa_V,\kappa_W\}.$$

So, if $I(P_V, V) \leq R - \delta$,

$$P_e(W) \geq \exp\left(-n\left[D(V||W|P) + \frac{2\max\{\kappa_V,\kappa_W\}}{1-\gamma_R SC(n,\delta,R,\kappa_V)}(\beta(n,|\mathcal{X}|,|\mathcal{Y}|) + 2\gamma_{RSC}(n,\delta,R,\kappa_V))\right]\right) \times \\ (1 - \gamma_{RSC}(n,\delta,R,\kappa_V)).$$

Now, we optimize over all V such that $I(P_V, V) \leq R - \delta$. From Lemma 1, we know that for the family of sphere-packing optimizing V's, $\kappa_V \leq \kappa_W + \log|\mathcal{Y}|$, so

$$\begin{aligned} -\frac{1}{n}\log P_e(W) &\leq \inf_{V:I(P_V,V)\leq R-\delta} D(V||W|P_V) + \frac{1}{n}\log(1-\gamma_{RSC}(n,\delta,R,\kappa_V)) + \\ &\qquad \frac{2\max\{\kappa_V,\kappa_W\}}{1-\gamma_{RSC}(n,\delta,R,\kappa_V)}(\beta(n,|\mathcal{X}|,|\mathcal{Y}|) + 2\gamma_{RSC}(n,\delta,R,\kappa_V)) \\ &\leq E_{sp}(R-\delta) + \frac{1}{n}\log(1-\gamma_{RSC}(n,\delta,R,\kappa_W+\log|\mathcal{Y}|)) + \\ &\qquad \frac{2(\kappa_W + \log|\mathcal{Y}|)}{1-\gamma_{RSC}(n,\delta,R,\kappa_W+\log|\mathcal{Y}|)} \times \\ &\qquad (\beta(n,|\mathcal{X}|,|\mathcal{Y}|) + 2\gamma_{RSC}(n,\delta,R,\kappa_W+\log|\mathcal{Y}|)). \end{aligned}$$

Since $\beta(n,|\mathcal{X}|,|\mathcal{Y}|) \to 0$ as $n \to \infty$, and $\gamma_{RSC}(n,\delta,R,\kappa_W + \log|\mathcal{Y}|) \to 0$ as $n \to \infty$ (assuming the refined strong converse holds), it follows that for large enough n (depending on W, R and δ),

$$-\frac{1}{n}\log P_e(W) \leq E_{sp}(R-\delta) + \delta.$$

A.6.3 A test channel with memory

Lemma 4 of Section 2.4.3: Let $\widetilde{V}$ be a measure[9] on $\mathcal{M} \times \mathcal{Y}^n$. Fix a $\delta > 0$. Let[10]

$$A \triangleq \left\{ (m, y^n) : \frac{1}{n} \log \frac{\mathbb{P}_{\widetilde{V}}(Y^n = y^n | M = m)}{\mathbb{P}_{\widetilde{V}}(Y^n = y^n)} \leq R - \delta \right\} \tag{A.16}$$

$$B \triangleq \left\{ (m, y^n) : \frac{1}{n} \log \frac{\mathbb{P}_{\widetilde{V}}(Y^n = y^n | M = m)}{\mathbb{P}_W(Y^n = y^n | M = m)} \leq E_{sp}(R - 2\delta) + \delta \right\} \tag{A.17}$$

$$E \triangleq \{ (m, y^n) : y^n \notin \mathcal{D}_m \}.$$

Then,

$$P_e(W) \geq \exp\left(-n \left[E_{sp}(R - 2\delta) + \delta\right]\right) \left[1 - \exp(-n\delta) - \mathbb{P}_{\widetilde{V}}(B^c) - \mathbb{P}_{\widetilde{V}}(A^c)\right].$$

Hence,

$$P_e(W) \geq \exp(-n[E_{sp}(R - 2\delta) + 2\delta])$$

for large n provided

$$\mathbb{P}_{\widetilde{V}} \left(\frac{1}{n} \log \frac{\mathbb{P}_{\widetilde{V}}(Y^n | M)}{\mathbb{P}_{\widetilde{V}}(Y^n)} \leq R - \delta \right) \to 1, \ n \to \infty \tag{A.18}$$

$$\mathbb{P}_{\widetilde{V}} \left(\frac{1}{n} \log \frac{\mathbb{P}_{\widetilde{V}}(Y^n | M)}{\mathbb{P}_W(Y^n | M)} \leq E_{sp}(R - 2\delta) + \delta \right) \to 1, \ n \to \infty. \tag{A.19}$$

Proof: The proof of this lemma is essentially the same as the proof of the information-spectrum converse. The only difference is the addition of the set B to make sure that the error event under channel $\widetilde{V}$ has a high enough probability under channel W. In the information spectrum vernacular, $\log(\mathbb{P}_{\widetilde{V}}(Y^n|M)/\mathbb{P}_{\widetilde{V}}(Y^n))$ would be the information density random variable and $\log(\mathbb{P}_{\widetilde{V}}(Y^n|M)/\mathbb{P}_W(Y^n|M))$ would be the divergence density random variable. Now,

$$\begin{aligned}
P_e(W) &= \mathbb{P}_W(E) \\
&\geq \mathbb{P}_W(E \cap B) \\
&= \sum_{(m, y^n) \in E \cap B} \frac{1}{|\mathcal{M}|} \mathbb{P}_W \left(Y^n = y^n | M = m\right) \\
&\overset{(a)}{\geq} \sum_{(m, y^n) \in E \cap B} \frac{1}{|\mathcal{M}|} \mathbb{P}_{\widetilde{V}} \left(Y^n = y^n | M = m\right) \exp(-n(E_{sp}(R - 2\delta) + \delta)) \\
&= \exp(-n(E_{sp}(R - \delta) + \delta)) \mathbb{P}_{\widetilde{V}} \left((M, Y^n) \in B \cap E\right).
\end{aligned} \tag{A.20}$$

[9] Of course, this measure is induced from one on $\mathcal{M} \times \mathcal{X}^n \times \mathcal{Y}^n$.

[10] Recall that capital letters are used to denote random variables while lower case vectors denote nonrandom variables.

In the above, (a) follows from the definition of B in (2.19). Now,

$$\begin{aligned}\mathbb{P}_{\widetilde{V}}(B \cap E) &= 1 - \mathbb{P}_{\widetilde{V}}\left((B \cap E)^c\right) \\ &= 1 - \mathbb{P}_{\widetilde{V}}\left(B^c \cup E^c\right) \\ &\geq 1 - \mathbb{P}_{\widetilde{V}}(B^c) - \mathbb{P}_{\widetilde{V}}(E^c). \qquad \text{(A.21)}\end{aligned}$$

Now,

$$\begin{aligned}\mathbb{P}_{\widetilde{V}}(E^c) &= \sum_{(m,y^n) \in E^c} \frac{1}{|\mathcal{M}|} \mathbb{P}_{\widetilde{V}}(y^n|m) \\ &= \sum_{(m,y^n) \in E^c \cap A} \frac{1}{|\mathcal{M}|} \mathbb{P}_{\widetilde{V}}(y^n|m) + \sum_{(m,y^n) \in E^c \cap A^c} \frac{1}{|\mathcal{M}|} \mathbb{P}_{\widetilde{V}}(y^n|m) \\ &\overset{(b)}{\leq} \sum_{(m,y^n) \in E^c \cap A} \frac{\exp(n(R-\delta))}{|\mathcal{M}|} \mathbb{P}_{\widetilde{V}}(y^n) + \sum_{(m,y^n) \in E^c \cap A^c} \frac{1}{|\mathcal{M}|} \mathbb{P}_{\widetilde{V}}(y^n|m) \\ &= \exp(-n\delta) \sum_{(m,y^n) \in E^c \cap A} \mathbb{P}_{\widetilde{V}}(y^n) + \sum_{(m,y^n) \in E^c \cap A^c} \frac{1}{|\mathcal{M}|} \mathbb{P}_{\widetilde{V}}(y^n|m) \\ &\overset{(c)}{\leq} \exp(-n\delta) + \sum_{(m,y^n) \in E^c \cap A^c} \frac{1}{|\mathcal{M}|} \mathbb{P}_{\widetilde{V}}(y^n|m) \\ &\leq \exp(-n\delta) + \mathbb{P}_{\widetilde{V}}(A^c), \qquad \text{(A.22)}\end{aligned}$$

where (b) follows from the definition of A in (A.16) and (c) follows because each y^n appears in at most one $\mathcal{D}_m$ by the definition of decoding regions. Plugging (A.22) into (A.21) and (A.21) into (A.20) completes the proof.

Proposition 1 of Section 2.4.3: With the choice of $\widetilde{V}$ in (2.17) and the definitions of Z_i and $\widetilde{Z}_i$ in (2.22) and (2.23) respectively, we have for $i = 1, \ldots, n$,

$$\mathbb{E}_{\widetilde{V}}[Z_i] = \sum_{y^{i-1} \in \mathcal{Y}^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) I(P_{y^{i-1}}, \widetilde{V}_{y^{i-1}}) \leq R - 2\delta \qquad \text{(A.23)}$$

$$\mathbb{E}_{\widetilde{V}}[\widetilde{Z}_i] = \sum_{y^{i-1} \in \mathcal{Y}^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) E_{sp}(R - 2\delta, P_{y^{i-1}}) \leq E_{sp}(R - 2\delta). \qquad \text{(A.24)}$$

Proof: Starting with the definition of Z_i,

$$\begin{aligned}
\mathbb{E}_{\widetilde{V}}[Z_i] &= \mathbb{E}_{\widetilde{V}}\left[\log \frac{\mathbb{P}_{\widetilde{V}}(Y_i|M, Y^{i-1})}{\mathbb{P}_{\widetilde{V}}(Y_i|Y^{i-1})}\right] \\
&= \sum_{m\in\mathcal{M}, y^{i-1}\in\mathcal{Y}^{i-1}, y_i\in\mathcal{Y}} \frac{1}{|\mathcal{M}|}\mathbb{P}_{\widetilde{V}}(y^{i-1}|m)\mathbb{P}_{\widetilde{V}}(y_i|m, y^{i-1}) \log \frac{\mathbb{P}_{\widetilde{V}}(y_i|m, y^{i-1})}{\mathbb{P}_{\widetilde{V}}(y_i|y^{i-1})} \\
&= \sum_{y^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) \sum_{m, y_i} \mathbb{P}_{\widetilde{V}}(M = m|Y^{i-1} = y^{i-1})\mathbb{P}_{\widetilde{V}}(y_i|m, y^{i-1}) \log \frac{\mathbb{P}_{\widetilde{V}}(y_i|m, y^{i-1})}{\mathbb{P}_{\widetilde{V}}(y_i|y^{i-1})}.
\end{aligned} \tag{A.25}$$

Now,

$$\begin{aligned}
\mathbb{P}_{\widetilde{V}}(y_i|m, y^{i-1}) &= \widetilde{V}_{y^{i-1}}(y_i|x_i(m, y^{i-1})) \\
\mathbb{P}_{\widetilde{V}}(y_i|y^{i-1}) &= \sum_{m\in\mathcal{M}} \mathbb{P}_{\widetilde{V}}(M = m|Y^{i-1} = y^{i-1})\widetilde{V}_{y^{i-1}}(y_i|x_i(m, y^{i-1})) \\
&= \sum_{m\in\mathcal{M}} \sum_{x\in\mathcal{X}} \mathbb{P}_{\widetilde{V}}(M = m|Y^{i-1} = y^{i-1})1(x_i(m, y^{i-1}) = x)\widetilde{V}_{y^{i-1}}(y_i|x) \\
&= \sum_{x\in\mathcal{X}} \widetilde{V}_{y^{i-1}}(y_i|x) \sum_{m\in\mathcal{M}} \mathbb{P}_{\widetilde{V}}(M = m|Y^{i-1} = y^{i-1})1(x_i(m, y^{i-1}) = x) \\
&\overset{(a)}{=} \sum_{x\in\mathcal{X}} \widetilde{V}_{y^{i-1}}(y_i|x)P_{y^{i-1}}(x) \\
&= \left(P_{y^{i-1}}\widetilde{V}_{y^{i-1}}\right)(y).
\end{aligned} \tag{A.26}$$

where (a) follows from the definition of the input distribution conditioned on y^{i-1} in (2.16).

Plugging (A.26) into (A.25) yields

$$\begin{aligned}
\mathbb{E}_{\widetilde{V}}[Z_i] &= \sum_{y^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) \sum_{m,y_i} \mathbb{P}_{\widetilde{V}}(M=m|Y^{i-1}=y^{i-1}) \widetilde{V}_{y^{i-1}}(y_i|x_i(m,y^{i-1})) \log \frac{\widetilde{V}(y_i|x_i(m,y^{i-1}))}{\left(P_{y^{i-1}}\widetilde{V}_{y^{i-1}}\right)(y_i)} \\
&= \sum_{y^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) \sum_{m,y_i} \sum_{x\in\mathcal{X}} 1(x_i(m,y^{i-1})=x) \mathbb{P}_{\widetilde{V}}(M=m|Y^{i-1}=y^{i-1}) \times \\
&\qquad \widetilde{V}_{y^{i-1}}(y_i|x)) \log \frac{\widetilde{V}(y_i|x)}{\left(P_{y^{i-1}}\widetilde{V}_{y^{i-1}}\right)(y_i)} \\
&= \sum_{y^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) \sum_{y_i} \sum_{x\in\mathcal{X}} \widetilde{V}_{y^{i-1}}(y_i|x)) \log \frac{\widetilde{V}(y_i|x)}{\left(P_{y^{i-1}}\widetilde{V}_{y^{i-1}}\right)(y_i)} \times \\
&\qquad \sum_{m\in\mathcal{M}} \mathbb{P}_{\widetilde{V}}(M=m|Y^{i-1}=y^{i-1}) 1(x_i(m,y^{i-1})=x) \\
&\overset{(b)}{=} \sum_{y^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) \sum_{y_i} \sum_{x\in\mathcal{X}} P_{y^{i-1}}(x) \widetilde{V}_{y^{i-1}}(y_i|x)) \log \frac{\widetilde{V}(y_i|x)}{\left(P_{y^{i-1}}\widetilde{V}_{y^{i-1}}\right)(y_i)} \\
&\overset{(c)}{=} \sum_{y^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) I\left(P_{y^{i-1}}, \widetilde{V}_{y^{i-1}}\right) \\
&\overset{(d)}{\leq} \sum_{y^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1})(R-2\delta) \leq R-2\delta,
\end{aligned}$$

where again (b) follows from the definition of the input distribution in (2.16), (c) is from the definition of mutual information and (d) comes from the choice of $\widetilde{V}$ in (2.17). An entirely analogous sequence of equations shows that

$$\begin{aligned}
\mathbb{E}_{\widetilde{V}}\left[\widetilde{Z}_i\right] &= \sum_{y^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) \min_{V\in\mathcal{W}} \left\{D(V||W|P_{y^{i-1}}) : I\left(P_{y^{i-1}}, V\right) \leq R-2\delta\right\} \\
&= \sum_{y^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) E_{sp}\left(R-2\delta, P_{y^{i-1}}\right) \\
&\leq \sum_{y^{i-1}} \mathbb{P}_{\widetilde{V}}(y^{i-1}) E_{sp}(R-2\delta).
\end{aligned}$$

A.7 Sheverdyaev's proof

In 1978, Sheverdyaev submitted a paper to *Problemy Peredachi Informatsii* (Problems of Information Transmission), entitled 'Lower bound for error probability in a discrete memoryless channel with feedback.' The paper was published by PPI in 1982 [30] and a fairly good English translation is available at research university engineering libraries. There are two results claimed in the paper about block codes with feedback:

1. The average probability of *correct reception* is exponentially decaying in the block length n for rates R greater than the capacity $C(W)$. It is shown in this paper that this exponent is the same as the case for block codes without feedback, where there is a matching lower and upper bound. Hence, the 'correct reception' exponent is determined for communication with feedback at all rates above capacity over a DMC (no zero transition probabilities are allowed in W).

2. The error exponent for block codes with feedback at rates R below capacity $C(W)$ is upper bounded by the sphere-packing exponent $E_{sp}(R)$ (for W with no zero transition probabilities), as is the case for block codes without feedback.

Of these two results, the first is certainly true, while the second's proof has serious issues. While we believe that the result is true (i.e. that the sphere-packing exponent is actually an upper bound to the error exponent for block codes with feedback at rates below capacity), we are not convinced the proof in [30] is complete. In this section, we will outline the proof and detail its shortcomings (as we perceive them).

Claim 1. *Fix a DMC W that has no zero probability transitions, i.e. $\tau_W \triangleq \min_{x,y} W(y|x) > 0$. Let*

$$\begin{aligned} \overline{P}_{c,fb}(W,R,n) &\triangleq \sup\{P_c(W): \text{ code is at least rate } R \text{ and blocklength is } n\} \\ \underline{P}_{e,fb}(W,R,n) &\triangleq 1 - \overline{P}_{c,fb}(W,R,n) \end{aligned} \tag{A.27}$$

be the optimal code correct reception and error probabilities for block codes with feedback of length n and rate at least R. Then,

$$\limsup_{n\to\infty} -\frac{1}{n}\log \underline{P}_{e,fb}(W,R,n) \le E_{sp}(R) = \max_P \min_{V:I(P,V)\le R} D(V||W|P),$$

so the sphere-packing exponent is an upper bound to the error exponent for block codes with feedback.

Proof outline: A series of lemmas are proved to come to the conclusion of the claim. We will state these lemmas and arrive at the point in the proof at the end where the justifications of two steps are not obvious. This is not to say that there are counterexamples to these justifications, but rather they are the same sticking points where Haroutunian could not tighten his bound, and the justifications are hand-waving of not-so-obvious points. With that in mind, the first lemma is intuitively straightforward and claims that non-randomized encoders and decoders are optimal.

Lemma 14. *We need only consider deterministic encoders and decoders with feedback. That is, randomized codes are not needed to achieve the performance in (A.27). This lemma is intuitively true*[11] *because of the absence of any adversarial agents such as in arbitrarily varying channel coding [55].*

The next lemma initiates the bounding of the error probability by using a reverse Hölder's inequality to get an inequality involving correct reception and error probabilities under two channels. Fix a test channel $V \in \mathcal{W}$ and also a set of conditional distributions[12]

$$\left\{q(\cdot|y^{i-1}) \in \mathcal{Q} : \forall i = 1, \dots, n, y^{i-1} \in \mathcal{Y}^{i-1}\right\}$$

that determine a measure on $\mathcal{M} \times \mathcal{Y}^n$ by

$$\mathbb{P}_q(M = m, Y^n = y^n) = \frac{1}{|\mathcal{M}|} \prod_{i=1}^{n} q(y_i|y^{i-1}).$$

Similarly,

$$\mathbb{P}_V(M = m, Y^n = y^n) = \frac{1}{|\mathcal{M}|} \mathbb{P}_V(Y^n = y^n | M = m).$$

Lemma 15. *Fix an $\alpha \in \mathbb{R}$, $\alpha \geq 1$. For an arbitrary deterministic encoder and decoder pair of rate at least R and length n,*

$$P_c(V)^\alpha \left(\frac{1}{|\mathcal{M}|}\right)^{1-\alpha} + P_e(V)^\alpha \left(1 - \frac{1}{|\mathcal{M}|}\right)^{1-\alpha} \leq \sum_{m,y^n} \mathbb{P}_V^\alpha(m, y^n) \mathbb{P}_q^{1-\alpha}(m, y^n) \tag{A.28}$$

$$P_c(V)^\alpha P_c(W)^{1-\alpha} + P_e(V)^\alpha P_e(W)^{1-\alpha} \leq \sum_{m,y^n} \mathbb{P}_V^\alpha(m, y^n) \mathbb{P}_W^{1-\alpha}(m, y^n). \tag{A.29}$$

Lemma 15 is proved by Reverse Hölder's inequality.

Proposition 10 (Reverse Hölder's Inequality)**.** *Suppose for each $i = 1, \dots, n$, $a_i \geq 0, b_i > 0, c_i \geq 0$, and $\sum_i c_i > 0$. For $\alpha \geq 1$,*

$$\sum_i a_i^\alpha b_i^{1-\alpha} c_i \geq \left(\sum_i a_i c_i\right)^\alpha \left(\sum_i b_i c_i\right)^{1-\alpha}.$$

[11] It is always good to verify this though given that randomization is generally required to obtain the optimal performance in hypothesis testing.

[12] Equivalently, one can straightaway define a distribution on $\mathcal{Y}^n$ and deduce the conditional distributions.

Reverse Hölder's inequality can be proved by using the usual Hölder's inequality [56]. The restriction that $\tau_W > 0$ is due to the use of Reverse Hölder's inequality where the channel transition probabilities are used as the b_i in Proposition 10. Now, we will define the 'tilted' measures (distributions)

$$\begin{aligned}\mathbb{P}_\alpha(m,y^n) &\triangleq \mathbb{P}_V^\alpha(m,y^n)\mathbb{P}_q^{1-\alpha}(m,y^n)\exp(-\mu(\alpha))\\ \mu(\alpha) &\triangleq \log\sum_{m,y^n}\mathbb{P}_V^\alpha(m,y^n)\mathbb{P}_q^{1-\alpha}(m,y^n)\end{aligned}$$

and

$$\begin{aligned}\widetilde{\mathbb{P}}_\alpha(m,y^n) &\triangleq \mathbb{P}_V^\alpha(m,y^n)\mathbb{P}_W^{1-\alpha}(m,y^n)\exp(-\widetilde{\mu}(\alpha))\\ \widetilde{\mu}(\alpha) &\triangleq \log\sum_{m,y^n}\mathbb{P}_V^\alpha(m,y^n)\mathbb{P}_W^{1-\alpha}(m,y^n).\end{aligned}$$

Lemma 16. *With the definitions above, which are used to continue bounding the bounds in Lemma 15, we have*

$$\mu(\alpha) \le \sum_{m,y^n}\mathbb{P}_\alpha(m,y^n)\sum_{i=1}^n\log\left[\sum_y V(y|x_i(m,y^{i-1}))^\alpha q(y|y^{i-1})^{1-\alpha}\right] \tag{A.30}$$

$$\widetilde{\mu}(\alpha) \le \sum_{m,y^n}\widetilde{\mathbb{P}}_\alpha(m,y^n)\sum_{i=1}^n\log\left[\sum_y V(y|x_i(m,y^{i-1}))^\alpha W(y|x_i(m,y^{i-1}))^{1-\alpha}\right]. \tag{A.31}$$

At this point, Sheverdyaev makes a specific 'optimal' choice in setting $q = q^*$ which induces $\mathbb{P}_\alpha = \mathbb{P}_\alpha^*$. For this choice of q^*, he proves the following lemma.

Lemma 17.

$$\frac{\mu(\alpha)}{\alpha n} \le \lambda(\alpha) \triangleq \log\sum_y\left[\sum_x P_{q^*,\alpha}(x)V^\alpha(y|x)\right]^{1/\alpha} \tag{A.32}$$

$$P_{q^*,\alpha}(x) \triangleq \sum_{m,y^n}\mathbb{P}_\alpha^*(m,y^n)\frac{1}{n}\sum_{i=1}^n 1\left(x_i(m,y^{i-1}) = x\right)$$

$$\frac{\widetilde{\mu}(\alpha)}{n} \le \nu(\alpha) \triangleq \log\sum_x\widetilde{P}_{V,\alpha}(x)\left[\sum_y V^\alpha(y|x)W^{1-\alpha}(y|x)\right] \tag{A.33}$$

$$\widetilde{P}_{V,\alpha}(x) \triangleq \sum_{m,y^n}\widetilde{\mathbb{P}}_\alpha(m,y^n)\frac{1}{n}\sum_{i=1}^n 1\left(x_i(m,y^{i-1}) = x\right).$$

At this point, we start to see the input distribution induced by different channels coming into play, because $P_{q^*,\alpha}$ is the input distribution under the tilted measure $\mathbb{P}_\alpha$ when $q = q^*$ and $\widetilde{P}_{V,\alpha}$ is the input distribution under the tilted measure $\widetilde{\mathbb{P}}_\alpha$. Note that for both cases, as $\alpha \to 1$, the input distributions under the tilted measure converge to P_V, the input distribution under channel V.

Cascading the bounds of (A.28), (A.30) and (A.32) yields

$$\begin{aligned} P_c^\alpha(V)\left(\frac{1}{|\mathcal{M}|}\right)^{1-\alpha} &\leq \exp(\mu(\alpha)) \leq \exp(\alpha n \lambda(\alpha)) \\ P_c(V) &\leq |\mathcal{M}|^{\frac{1-\alpha}{\alpha}} \exp(n\lambda(\alpha)) \\ &\leq \exp\left(-n\left(-\lambda(\alpha) - \frac{1-\alpha}{\alpha}R\right)\right), \end{aligned}$$

where we have used that the rate of the code is at least R. Hence,

$$\begin{aligned} -\frac{1}{n}\log P_c(V) &\geq -\lambda(\alpha) - \frac{1-\alpha}{\alpha}R \qquad \text{(A.34)} \\ &= -\log \sum_y \left(\sum_x P_{q^*,\alpha}(x) V^\alpha(y|x)\right)^{1/\alpha} - \frac{1-\alpha}{\alpha}R. \end{aligned}$$

If we let $\rho = (1-\alpha)/\alpha$, we have

$$\begin{aligned} -\frac{1}{n}\log P_c(V) &\geq -\log \sum_y \left(\sum_x P_{q^*,\alpha}(x) V^{\frac{1}{1+\rho}}(y|x)\right)^{1+\rho} - \rho R \\ &\geq \inf_{P\in\mathcal{P}} -\log \sum_y \left(\sum_x P(x) V^{\frac{1}{1+\rho}}(y|x)\right)^{1+\rho} - \rho R, \ \forall\ \rho \in (-1, 0]. \end{aligned}$$

Hence, if we just let $V = W$, we get the first result given in the paper regarding correct reception for rates above capacity:

$$\begin{aligned} -\frac{1}{n}\log P_c(W) &\geq \sup_{\rho\in(-1,0]} \left[\inf_P E_0(P, W, \rho) - \rho R\right] \\ E_0(P, W, \rho) &= -\log \sum_y \left(\sum_x P(x) W^{\frac{1}{1+\rho}}(y|x)\right)^{1+\rho}. \end{aligned}$$

Hence,

$$\liminf_{n\to\infty} -\frac{1}{n}\log \overline{P}_{c,fb}(W, R, n) \geq \sup_{\rho\in(-1,0]} \left[\inf_P E_0(P, W, \rho) - \rho R\right] \triangleq E_c(R).$$

It can be shown that $E_c(R) > 0$ for all $R > C(W)$ [47], so the strong converse holds with feedback and exponentially decreasing probability of correct reception. Further, this bound is the same as for block codes without feedback operating above capacity ([47], Problem 2.5.16(a,b)).

Now, returning to the case for $R < C(W)$, from (A.34), we know that

$$\begin{aligned} R &\leq \frac{\alpha}{\alpha-1}\left[-\frac{1}{n}\log P_c(V) + \lambda(\alpha)\right] \\ &= \frac{\alpha}{\alpha-1}\left[-\frac{1}{n}\log P_c(V)\right] + \frac{\alpha}{\alpha-1}\left[(\alpha-1)\lambda'(1) + O((\alpha-1)^2)\right], \end{aligned} \tag{A.35}$$

where we have taken a Taylor expansion[13] of $\lambda(\alpha)$ about $\alpha = 1$ and noted that $\lambda(1) = 0$. From (A.29), we have

$$\begin{aligned} P_e^{\alpha}(V)P_e^{1-\alpha}(W) &\leq \exp(n\nu(\alpha)) \\ \alpha\log P_e(V) - (\alpha-1)\log P_e(W) &\leq n\nu(\alpha). \end{aligned}$$

Therefore, a little rearranging gives

$$\begin{aligned} -\frac{1}{n}\log P_e(W) &\leq \frac{\nu(\alpha)}{\alpha-1} + \frac{\alpha}{\alpha-1}\frac{-1}{n}\log P_e(V) \\ &= \frac{1}{\alpha-1}\big((\alpha-1)\nu'(1) + O((\alpha-1)^2)\big) + \frac{\alpha}{\alpha-1}\frac{-1}{n}\log P_e(V), \end{aligned} \tag{A.36}$$

where we have taken a Taylor expansion of $\nu(\alpha)$ about $\alpha = 1$ and noted that $\nu(1) = 0$. A bit of differentiation shows that

$$\begin{aligned} \lambda'(1) &= I(P_V, V) \\ \nu'(1) &= D(V||W|P_V), \end{aligned}$$

where P_V is the input distribution under channel V. Plugging these values into (A.35) and (A.36) yields

$$R \leq \frac{\alpha}{\alpha-1}\left(-\frac{1}{n}\log P_c(V)\right) + \alpha I(P_V, V) + g(V, \alpha-1) \tag{A.37}$$

$$-\frac{1}{n}\log P_e(W) \leq \frac{\alpha}{\alpha-1}\left(-\frac{1}{n}\log P_e(V)\right) + D(V||W|P_V) + \widetilde{g}(V, \alpha-1), \tag{A.38}$$

where the $g(V, \alpha-1)$ and $\widetilde{g}(V, \alpha-1)$ are constants depending on V (through the second-order derivatives of λ and ν respectively). For a fixed V, g and $\widetilde{g}$ can be made arbitrarily

[13] The notation $O((\alpha-1)^2)$ is meant to apply for $\alpha \to 1$.

small if $\alpha - 1$ is arbitrarily small. However, *it is unclear*[14] *if there are uniform (over V and n) constants that bound them.* This is the first main issue with the proof of this claim. Especially for the second derivative of ν, a uniform constant must be justified.

At this point, Sheverdyaev makes a choice of $\alpha = 1 + 1/\sqrt{n}$ and considers only channels V such that $I(P_V, V) \leq R - \epsilon$ for some $\epsilon > 0$. Then, since mutual information is bounded by $\log|\mathcal{Y}|$, (A.37) becomes

$$\begin{aligned}\epsilon - \frac{\log|\mathcal{Y}|}{\sqrt{n}} &\leq R - \left(1 + \frac{1}{\sqrt{n}}\right) I(P_V, V) \\ &\leq (\sqrt{n} + 1)\left(-\frac{1}{n}\log P_c(V)\right) + g(V, 1/\sqrt{n}).\end{aligned}$$

From this, we can deduce that

$$\begin{aligned}&P_c(V) \leq \exp\left(-\frac{n}{\sqrt{n}+1}\left[\epsilon - \frac{\log|\mathcal{Y}|}{n} - g(V, 1/\sqrt{n})\right]\right) \\ &P_e(V) \simeq 1, \text{ for large } n \text{ provided } I(P_V, V) \leq R - \epsilon\end{aligned}$$

Note that *this conclusion is dubious* because we have a counterexample (Example 1) showing that the 'refined strong converse' does not hold, even for codes without feedback. That is, having a mutual information lower than the rate is not enough to force the error probability to 1. Although Fano's inequality can be used, this only shows the probability of error is bounded above some non-vanishing constant. Forging ahead, however, plugging this information into (A.38) gives

$$-\frac{1}{n}\log \underline{P}_{e,fb}(W, R, n) \leq D(V||W|P_V) + f(n, \epsilon, V),$$

where the function f incorporates $\widetilde{g}$ and the $\log P_e(V)$ term in (A.38). By choice of V that have a low mutual information, we also have

$$-\frac{1}{n}\log \underline{P}_{e,fb}(W, R, n) \leq \inf_{V:I(P_V,V)\leq R-\epsilon} D(V||W|P) + \sup_{V:I(P_V,V)\leq R-\epsilon} f(n, \epsilon, V).$$

It is then claimed that $f(n, \epsilon, V)$ can be bounded uniformly over V, but as mentioned earlier, it is unclear how to verify this point. In the next step, he claims that *"a fortiori"* (obviously),

$$\inf_{V:I(P_V,V)\leq R-\epsilon} D(V||W|P_V) \leq \sup_{P\in\mathcal{P}} \inf_{V:I(P,V)\leq R-\epsilon} D(V||W|P). \tag{A.39}$$

While this step might seem obvious, as it apparently did to Sheverdyaev and reviewers of the paper, it suffers from a subtle logical error. One might think that by taking the supremum

[14] This is something that both we along with Giacomo Como and Baris Nakiboglu of MIT could not verify.

over P_V and then infimizing over V, we are only weakening the bound. However, we cannot apply the error probability bound to all pairs $(P, V) \in \mathcal{P} \times \mathcal{W}$ that have $I(P, V) \leq R - \epsilon$. Rather the bound only applies to (P_V, V) that have $I(P_V, V) \leq R - \epsilon$. That is, if there is a V such that $P_V = P$, there is no guarantee that $I(P_V, V) \leq R - \epsilon$. So the infimizing channel V for a given P is not necessarily available to use for bounding purposes. Hence, the RHS of (A.39) *contains more channels in the infimum than it should.* There is a potential fix to this issue as developed by Baris Nakiboglu (and reproduced in Lemma 1), but the necessity of this bridge in the proof was not anticipated by Sheverdyaev, at least as can be reasonably concluded from the paper.

A.8 Augustin's manuscript

Augustin [34] claims that the sphere-packing bound holds for fixed-length block codes with feedback. His Habilitationschrift, entitled *Noisy Channels*, from Universiät Erlangen-Nürenberg was submitted to Springer Lecture Notes in 1978, but appears to be unpublished as it seems to be unavailable in the literature[15].

Our brief inspection of the contents of the manuscript indicate that it is a rigorous treatment of point-to-point channel coding for channels with input and output spaces that are not required to be finite or real-valued. Measurability is the main requirement for many of the results, while stationarity of channels is sometimes assumed. One result of note is an extension of the sphere-packing bound to these abstract input-output space channels, inspired by the original proof of Shannon, Gallager and Berlekamp [18], without requiring a fixed composition or type assumption. Of particular importance to us is one result located towards the end of the manuscript. It is a result that claims that the sphere-packing exponent is an upper bound on the error exponent for fixed-length block codes with feedback used over finite-alphabet DMCs. In this section, we will discuss the result in question.

Unfortunately, we are not able to ascertain that the result was correctly proved nor say for sure that the result has not been rigorously proved due to two reasons. First, the original manuscript was written in German and translated to English. However, the phrasing used in the translation makes it difficult to follow in many places. The second reason is that the proof of the result is not explained in sufficient detail in several places where it is different from the proof of sphere-packing for block codes without feedback (as done in [34]). With that said, our observations in this section are much less formal than those made of the Sheverdyaev paper.

Claim 2 (Theorem 41.7 of [34]). *We are restricting here to what the theorem says about stationary DMCs. The actual theorem applies in a more general form to finite input, stationary channels and list codes with feedback. For a DMC W, and length n, rate R code with*

[15] We received a photocopy of the Springer Lecture Notes submission in Summer 2010 from Professor Fady Alajaji of Queens University, Canada after being informed by Professor Imre Csiszár of its existence.

feedback,

$$-\frac{1}{n}\log P_e(W) \leq \sup_{\rho \geq 0} E_0(\rho) - \rho\left(R - O\left(\frac{\log(n)}{n^{1/3}}\right)\right) + O\left(\frac{\log(n)}{n^{1/3}}\right)$$
$$= E_{sp}\left(R - O\left(\frac{\log(n)}{n^{1/3}}\right)\right) + O\left(\frac{\log(n)}{n^{1/3}}\right).$$

Note that the convergence of the error exponent involves terms of the order of $\log(n)/n^{1/3}$. This is much slower than the convergence of the error exponent for block codes without feedback, where the 'slack' terms are of the order of $\log(n)/n$ for the argument within the sphere-packing exponent and $1/\sqrt{n}$ for the additive term outside.

Augustin's proof, as might be expected considering the history of this problem, is primarily concerned with 'mutual information' terms of the form

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{y} V_{i,\rho}(y|x_i(m,y^{i-1}))\log\frac{V_{i,\rho}(y|x_i(m,y^{i-1}))}{Q_\rho(y)} \tag{A.40}$$

and 'divergence' terms of the form

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{y} V_{i,\rho}(y|x_i(m,y^{i-1}))\log\frac{V_{i,\rho}(y|x_i(m,y^{i-1}))}{W_\rho(y|x_i(m,y^{i-1}))}, \tag{A.41}$$

for appropriately chosen 'tilted' channels and distributions $V_{i,\rho}$ and Q_ρ. His proof of the sphere-packing bound for codes without feedback proves probabilistic statements about terms like (A.40) and (A.41) when there is a measure over the tilting parameter ρ. He then uses a pigeon-hole argument to claim that those probability statements also hold for an exponentially-equivalent-rate subcode, for a small range of ρ of size $\Theta(1/n^2)$, with a slightly smaller probability. Then, he proves some continuity statements to show that since the range of ρ is now very small, we get terms close to $E_0(\rho)$ from (A.41). Since we don't know *a priori* which ρ will come out of the pigeon-hole argument, a sup over all $\rho \geq 0$ is taken.

For the proof with feedback, he notes that he will choose ρ depending not only on the message, but also the input letter at a given time. He claims that a selection process to refine down to a small range of ρ after this yields a probability that decays faster than exponentially (as the ρ depend on too many factors now). The claimed fix to this problem is to group instants of time into blocks of length $k = O(n^{1/3})$, but not $k = o(n^{1/3})$. The ρ are then selected based on the message and received symbols up to those in the last block of length k. Augustin claims that the resulting probability bounds, after thinning out to get to a small range of ρ, are not sub-exponential anymore, but some bounding power is lost in this process, resulting in the $\log(n)/n^{1/3}$ terms.

Unfortunately, this proof is only sketched and it is unclear how allowing the ρ to depend on blocks of length k received symbols leads to a pigeon hole argument over a sub-exponential number of holes (as we understand it, this still leads to a super-exponential number of possible holes). If the ρ's do not depend on past histories but only the number of the sub-block in the whole block, it is unclear how the probability statements are being made. Hence, we can neither verify, nor refute this claim and proof.

A.9 Sphere packing holds when the fixed type encoding tree condition holds

A.9.1 Proof of fixed type proposition

Proposition 2 of Section 2.5: Suppose an encoding tree for a message $m \in \mathcal{M}$ satisfies the fixed type encoding tree condition with type $P \in \mathcal{P}_n$. Then, for all $V \in \mathcal{V}_n(P)$,

$$\begin{aligned} |B(m,P,V)| &= |T_V(x^n)| \\ &\geq \frac{\exp(nH(V|P))}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}}, \end{aligned}$$

where $x^n \in T_P$ is arbitrary because $|T_V(x^n)|$ depends only on the type of x^n.

Proof:

Definition 9 (Canonical sequence for type P)**.** *For an integer $n \geq 1$ and a type $P \in \mathcal{P}_n$, we define the* **canonical sequence** *of type P (and implicitly of length n) as follows. First, without loss of generality, assume that $\mathcal{X} = \{1, 2, \ldots, |\mathcal{X}|\}$. The canonical sequence of type P is denoted $\overline{x}_P^n$ and is the 'lexicographically' ordered sequence in type P. That is, it is a string with $nP(1)$ 1's, followed by $nP(2)$ 2's, followed by $nP(3)$ 3's, and so on until it is ended by $nP(|\mathcal{X}|)$ $|\mathcal{X}|$'s.*

For example, if $\mathcal{X} = \{1, 2, 3, 4, 5\}, n = 10$ and $P = (2/10, 3/10, 0/10, 4/10, 1/10)$, then

$$\overline{x}_P^n = (1, 1, 2, 2, 2, 4, 4, 4, 4, 5).$$

To warm up, we recall the reason that $|T_V(\overline{x}_P^n)| = |T_V(x^n)|$ for all $x^n \in T_P$. Because x^n has type P, there is a permutation σ that rearranges $\overline{x}_P^n$ to result in x^n. That is,

$$\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$$

is a permutation (so it is one-to-one and onto) such that

$$\overline{x}_{P,i} = x_{\sigma(i)}.$$

The fact that σ is a permutation means that the type of x^n is the same as the type of $\overline{x}_P^n$. Now, fix a $V \in \mathcal{V}_n(P)$. We want to show that $|T_V(x^n)| = |T_V(\overline{x}_P^n)|$, and we can do so by coming up with a one-to-one map from $\mathcal{Y}^n$ to $\mathcal{Y}^n$ that preserves the input-output pair counts. Since we have a σ that maps indices and rearranges x^n to get $\overline{x}_P^n$, it turns out doing the same for y^n sequences preserves the conditional shell relationship. More rigorously, let $\tau : \mathcal{Y}^n \to \mathcal{Y}^n$ be defined by

$$\tau(y^n) = \tau\left((y_1, \dots, y_n)\right) = (y_{\sigma(1)}, y_{\sigma(2)}, \dots y_{\sigma(n)}).$$

Clearly τ is a permutation map of $\mathcal{Y}^n$. All that needs to be checked is that it preserves the (x, y) counts between sequences. Fix a $y^n \in T_V(x^n)$ and any $x \in \mathcal{X}, y \in \mathcal{Y}$. Because $y^n \in T_V(x^n)$,

$$\begin{aligned} nP(x)V(y|x) &= \sum_{i=1}^{n} 1(x_i = x, y_i = y) \\ &= \sum_{j=1}^{n} \sum_{i=\sigma(j)} 1(x_i = x, y_i = y) \\ &= \sum_{i=1}^{n} 1(x_{\sigma(i)} = x, y_{\sigma(i)} = y) \\ &= \sum_{i=1}^{n} 1(\overline{x}_{P,i} = x, y_{\sigma(i)} = y). \end{aligned}$$

Therefore, we have that if $y^n \in T_V(x^n)$, $\tau(y^n) \in T_V(\overline{x}_P^n)$. Also, if $y^n \notin T_V(x^n)$, $y^n \in T_{V'}(x^n)$ for some other $V' \in \mathcal{V}_n(P)$ and hence $\tau(y^n) \in T_{V'}(\overline{x}_P^n)$ and $\tau(y^n) \notin T_V(\overline{x}_P^n)$. Therefore, $|T_V(x^n)| = |T_V(\overline{x}_P^n)|$.

Now, if the code has feedback, a simple permutation of the indices does not work as a map to show that $|B(m, P, V)| = T_V(\overline{x}_P^n)$. A bit more work needs to be done, and we start with several definitions.

Definition 10 (Encoding tree (with feedback))**.** *An* **encoding tree** *of length n, $\mathcal{E}$, is a set of channel input maps with feedback (with a dummy message index m),*

$$\mathcal{E} = \left\{x_i(m, y^{i-1}) \in \mathcal{X} : i = 1, \dots, n, y^{i-1} \in \mathcal{Y}^{i-1}\right\}.$$

The tree is visualized with nodes being labeled with channel inputs and edges being labeled with channel outputs as in Figure 2.9.

Definition 11 (Canonical encoding tree)**.** *For a length n and type $P \in \mathcal{P}_n$, the* **canonical encoding tree** *of type P (and implicitly of length n), denoted by $\mathcal{E}_P$ is*

$$\mathcal{E}_P \triangleq \{x_i(m, y^{i-1}) = \overline{x}_{P,i} : i = 1, \dots, n, y^{i-1} \in \mathcal{Y}^{i-1}\}.$$

The canonical encoding tree uses the feedback trivially and uses the canonical sequence of type P as the input.

The result that we wish to prove in this proposition is that if $\mathcal{E}$ is an encoding tree that satisfies the fixed-type encoding tree condition with type P, then for any $V \in \mathcal{V}_n(P)$, $|B(m,P,V)| = |T_V(\overline{x}_P^n)|$. We will prove this proposition by starting with the canonical encoding tree for type P and performing a finite number of changes to the tree (in labeling of the inputs at the nodes) to arrive at the given encoding tree $\mathcal{E}$. Each of these changes will be shown to keep $|B(m,P,V)| = |T_V(\overline{x}_P^n)|$. We will be working with several encoding trees, so to avoid confusion, let $B_{\mathcal{E}}(P,V)$ denote the output sequences (leaves in the tree) that have conditional type V when seen from the input sequence for encoding tree $\mathcal{E}$.

So, assume that $\mathcal{E}$ satisfies the fixed type encoding tree condition with type P. Start with the canonical encoding tree $\mathcal{E}_P$. If $x_1(m) = \overline{x}_{P,1}$, then $\mathcal{E}$ and $\mathcal{E}_P$ agree at the first letter and we don't need to modify $\mathcal{E}_P$. Suppose, however that $x_1(m) \neq \overline{x}_{P,1}$ so that $\mathcal{E}$ and $\mathcal{E}_P$ disagree on the symbol at the root node of the tree. We claim that we can modify $\mathcal{E}_P$ to get a new encoding tree $\mathcal{E}' = \{x'(m, y^{i-1}) : i = 1, \ldots, n, y^{i-1} \in \mathcal{Y}^{i-1}\}$ such that $x'_1(m) = x_1(m)$ and for each $y \in \mathcal{Y}$, the encoding tree of length $n-1$ that is the child of $y^1 = y$ is the canonical encoding tree $\mathcal{E}_{P'}$ for the type P' with

$$P'(x) = \begin{cases} nP(x)/(n-1) & \text{if } x \neq x_1(m) \\ (nP(x)-1)/(n-1) & \text{if } x = x_1(m) \end{cases}$$

That is, P' is the type for length $n-1$ one gets by taking type P for length n and removing one use of the symbol $x_1(m)$. This is done by using Proposition 11 repeatedly. The proposition says that the input symbol at a node and the input symbol of its immediate children can be interchanged without affecting the number of sequences in $B_{\mathcal{E}_P}(P,V)$ (although the actual sequences themselves will change) provided that the input symbols for children further down do not depend on any more received symbols (only the depth in the tree). Therefore, we can modify $\mathcal{E}_P$ to $\mathcal{E}'$ by repeatedly applying Proposition 11. We simply slide the first occurrence of symbol $x_1(m)$ in the canonical encoding tree (such an occurrence must exist because of the fixed type encoding tree assumption) back to the root node (interchanging symbols as we go) and the children of the root node are all canonical trees of type P' and length $n-1$. Proposition 11 ensures that for each $V \in \mathcal{V}_n(P)$, $|B_{\mathcal{E}_P}(P,V)| = |B_{\mathcal{E}'}(P,V)|$.

Now, $\mathcal{E}$ can use the feedback from y^1 nontrivially, so $x(m,y^1)$ can be different for each $y^1 \in \mathcal{Y}$. This is fine, however, because we can repeat the process for each child of the root node separately to get another encoding tree $\widetilde{\mathcal{E}}$ of length n such that $\mathcal{E}$ and $\widetilde{\mathcal{E}}$ agree for the first two levels and the children after two levels are all canonical encoding trees of length $n-2$ (however, the types of canonical trees will be different if the encoding tree uses the feedback at time 2 nontrivially), and $|B_{\widetilde{\mathcal{E}}}(P,V)| = |B_{\mathcal{E}_P}(P,V)|$ for all $V \in \mathcal{V}_n(P)$. The process can be continued until we have modified $\mathcal{E}_P$ to get $\mathcal{E}$. We crucially use the fixed type encoding tree condition because we require that along every sequence y^n the number of x's is the same for each $x \in \mathcal{X}$.

To finish off the proof, we need only recognize that $|B_{\mathcal{E}_P}(P,V)| = |T_V(\overline{x}_P^n)|$ since the canonical encoding tree does not use feedback.

Proposition 11. *Let $\mathcal{E}$ be an encoding tree of length n and suppose there is a $k \in \{0, 1, \ldots, n-2\}$ and $\overline{y}^k \in \mathcal{Y}^k$ such that*[16] *for all $i \in \{k+1, \ldots, n\}$, for all $y^{i-1-k}, \widetilde{y}^{i-1-k} \in \mathcal{Y}^{i-1-k}$,*

$$x_i(m, (\overline{y}^k, y^{i-1-k})) = x_i(m, (\overline{y}^k, y^{i-1-k})).$$

That is, if the received sequence up through time k was $\overline{y}^k$, the input symbol is independent on any of the received symbols after time k. Consider the following modified encoding tree $\widetilde{\mathcal{E}} = \{\widetilde{x}_i(m, y^{i-1}) : i = 1, \ldots, n, y^{i-1} \in \mathcal{Y}^{i-1}\}$,

$$\widetilde{x}_i(m, y^{i-1}) = \left\{ \begin{array}{cc} x_i(m, y^{i-1}), & \textit{if } i \leq k \\ x_i(m, y^{i-1}), & \textit{if } i > k, y^k \neq \overline{y}^k \\ x_{k+1}(m, \overline{y}^k), & \textit{if } i = k+2, y^k = \overline{y}^k \\ x_{k+2}(m, (\overline{y}^k, 1)) & \textit{if } i = k+1, y^k = \overline{y}^k \end{array} \right.$$

$\widetilde{\mathcal{E}}$ is the encoding tree one gets if we interchange the symbols of the input for encoding tree $\mathcal{E}$ at time $k+1$ and $k+2$ when the received sequence after time k is $\overline{y}^k$. By the assumption that feedback is not used after time k when the received sequences is $\overline{y}^k$, $x_{k+2}(m, (\overline{y}^k, y))$ is independent of $y \in \mathcal{Y}$, so in the above definition of $\widetilde{x}$, we have used the dummy symbol $1 \in \mathcal{Y}$. Then, for all $P \in \mathcal{P}_n$ and $V \in \mathcal{V}_n(P)$,

$$|B_{\mathcal{E}}(P, V)| = |B_{\widetilde{\mathcal{E}}}(P, V)|.$$

Proof: Figure A.1 depicts how $\mathcal{E}$ is modified to get $\widetilde{\mathcal{E}}$. In order to prove this proposition, we will construct a mapping $\tau : \mathcal{Y}^n \to \mathcal{Y}^n$ that is one-to-one and onto (i.e. a permutation) that preserves the (x, y) counts for the $\mathcal{E}$ after the modification. Let

$$\tau(y^n) = \left\{ \begin{array}{cl} y^n & \text{if } y^k \neq \overline{y}^k \\ (y_1, \ldots, y_k, y_{k+2}, y_{k+1}, y_{k+3}, \ldots, y_n) & \text{if } y^k = \overline{y}^k. \end{array} \right.$$

Hence, τ performs the same interchanging to y^n as was done to modify $\mathcal{E}$. It is clear that τ is a permutation of $\mathcal{Y}^n$ because it is one-to-one and onto (the inverse permutation is actually τ as well). Now, fix a $P \in \mathcal{P}_n, V \in \mathcal{V}_n(P)$, and $y^n \in B_{\mathcal{E}}(P, V)$. We wish to show that $\tau(y^n) \in B_{\widetilde{\mathcal{E}}}(P, V)$. If $y^k \neq \overline{y}^k$, it is clear that $y^n = \tau(y^n) \in B_{\widetilde{\mathcal{E}}}(P, V)$ still because the

[16] If $k = 0$, $\mathcal{Y}^k = \emptyset$ and y^k is the null string.

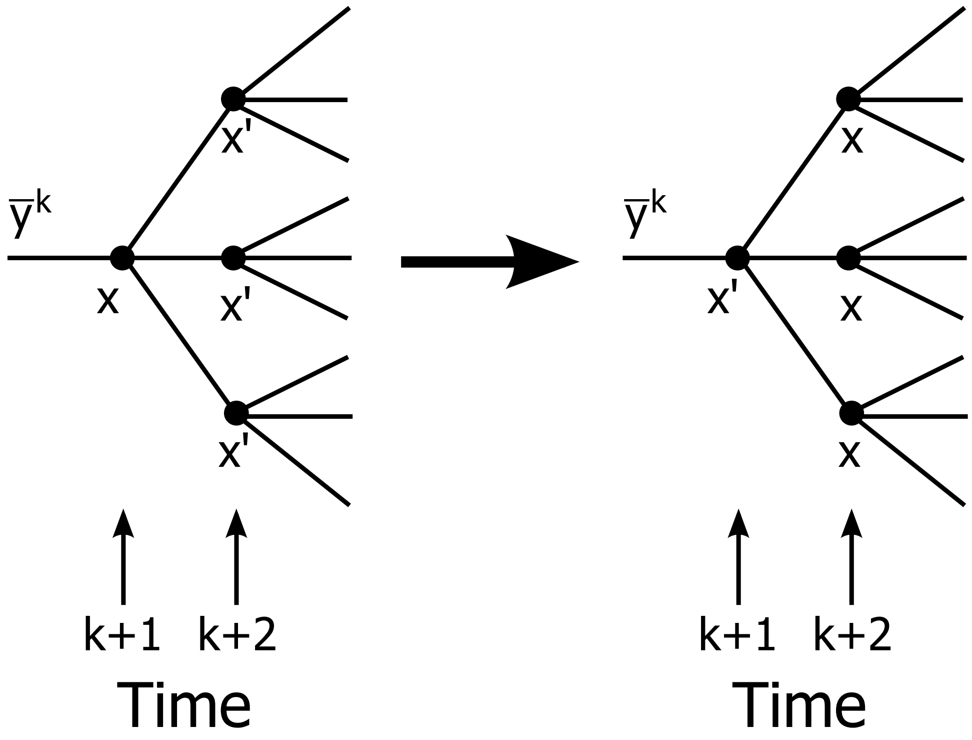

Figure A.1: The interchange of inputs for times $k+1$ and $k+2$ as described in Proposition 11. The encoding tree is assumed to have inputs that only depend on the time i after time k when the received sequence is $\overline{y}^k$. The original tree is modified by interchanging the input at time $k+1$ with the input at time $k+2$, only for the nodes along the branch $\overline{y}^k$. All parts of the encoding tree that are not shown remain unchanged. The resulting modified tree has the same number of sequences in its conditional shells with feedback as the original tree.

encoding tree along y^n has not changed. So suppose that $y^k = \overline{y}^k$, and let $z^n = \tau(y^n)$, so

$$\begin{aligned}
nP(x)V(y|x) &= \sum_{i=1}^{n} 1(x_i(m, y^{i-1}) = x, y_i = y) \\
&= \sum_{i=1}^{k} 1(x_i(m, y^{i-1}) = x, y_i = y) + 1(x_{k+1}(m, y^k) = x, y_{k+1} = y) + \\
&\qquad 1(x_{k+2}(m, y^{k+1}) = x, y_{k+2} = y) + \sum_{i=k+3}^{n} 1(x_i(m, y^{i-1}) = x, y_i = y) \\
&\overset{(a)}{=} \sum_{i=1}^{k} 1(\widetilde{x}_i(m, y^{i-1}) = x, y_i = y) + 1(x_{k+1}(m, y^k) = x, y_{k+1} = y) + \\
&\qquad 1(x_{k+2}(m, y^{k+1}) = x, y_{k+2} = y) + \sum_{i=k+3}^{n} 1(\widetilde{x}_i(m, y^{i-1}) = x, y_i = y) \\
&\overset{(b)}{=} \sum_{i=1}^{k} 1(\widetilde{x}_i(m, y^{i-1}) = x, y_i = y) + 1(\widetilde{x}_{k+1}(m, y^k) = x, y_{k+2} = y) + \\
&\qquad 1(\widetilde{x}_{k+2}(m, y^{k+1}) = x, y_{k+1} = y) + \sum_{i=k+3}^{n} 1(\widetilde{x}_i(m, y^{i-1}) = x, y_i = y) \\
&= \sum_{i=1}^{n} 1(\widetilde{x}_i(m, z^{i-1}) = x, z_i = y),
\end{aligned}$$

where in (a) we have used the fact that $\widetilde{x}_i$ does not change for times other than $i = k+1, k+2$ and in (b), we have interchanged both the x_i and y_i for $i = k+1, k+2$. Hence, $z^n = \tau(y^n) \in B_{\widetilde{\mathcal{E}}}(P, V)$. Hence, τ maps $B_{\mathcal{E}}(P, V)$ into $B_{\widetilde{\mathcal{E}}}(P, V)$ for each $P \in \mathcal{P}_n, V \in \mathcal{V}_n(P)$. Therefore, since these are disjoint sets, $|B_{\mathcal{E}}(P, V)| = |B_{\widetilde{\mathcal{E}}}(P, V)|$ for each $P \in \mathcal{P}_n, V \in \mathcal{V}_n(P)$.

A.9.2 Proof of theorem

Theorem 3 of Section 2.5: Fix a $\delta > 0, R > 0$. There exists a finite $n_{FT}(W, R, \delta)$ such that for any fixed-length code with feedback of length $n \geq n_{FT}(W, R, \delta)$ and rate R with encoding trees for all messages satisfying the fixed type encoding tree condition,

$$-\frac{1}{n} \log P_e(W) \leq E_{sp}(R - \delta) + \delta.$$

Proof: For now, assume that the fixed type encoding tree condition holds for all messages in the code with the same $P \in \mathcal{P}_n$, so $\forall\ y^n \in \mathcal{Y}^n, P(m, y^n) = P$. With that in mind,

$$\begin{aligned}
P_e(W) &= \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \notin \mathcal{D}_m | M = m) \\
&= \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \sum_{V\in\mathcal{V}_n(P)} \mathbb{P}_W(Y^n \in \mathcal{D}_m^c \cap B(m, P, V) | M = m) \\
&=\frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}(P)} \sum_{V\in\mathcal{V}_n(P)} \sum_{y^n\in B(m,P,V)\cap\mathcal{D}_m^c} \mathbb{P}_W(Y^n = y^n | M = m) \\
&= \sum_{V\in\mathcal{V}_n(P)} \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} |\mathcal{D}_m^c \cap B(m, P, V)| \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \\
&= \sum_{V\in\mathcal{V}_n(P)} \exp\left(-n\left[D(V||W|P) + H(V|P) + R\right]\right) \sum_{m\in\mathcal{M}} |B(m, P, V) \cap \mathcal{D}_m^c| \\
&= \sum_{V\in\mathcal{V}_n(P)} \exp\left(-n\left[D(V||W|P) + H(V|P) + R\right]\right) \times \\
&\qquad \left[\sum_{m\in\mathcal{M}} |B(m, P, V)| - |B(m, P, V) \cap \mathcal{D}_m|\right] \\
&\overset{(a)}{\geq} \sum_{V\in\mathcal{V}_n(P)} \exp\left(-n\left[D(V||W|P) + H(V|P) + R\right]\right) \left[\left(\sum_{m\in\mathcal{M}} |B(m, P, V)|\right) - |T_{PV}|\right] \\
&\geq \sum_{V\in\mathcal{V}_n(P)} \exp\left(-n\left[D(V||W|P) + H(V|P) + R\right]\right) \times \\
&\qquad \left[\left(\sum_{m\in\mathcal{M}} |B(m, P, V)|\right) - \exp(nH(PV))\right],
\end{aligned}$$

where (a) follows from Proposition 14. From Proposition 2, we know that $|B(m, P, V)| \geq$

$\exp(nH(V|P))/(n+1)^{|\mathcal{X}||\mathcal{Y}|}$ for all $m \in \mathcal{M}$, so for any $\delta > 0$, we have

$$\begin{aligned}
P_e(W) &\geq \sum_{V \in \mathcal{V}_n(P)} \exp\left(-n\left[D(V||W|P) + H(V|P) + R\right]\right) \times \\
&\qquad \left[\exp\left(n\left[H(V|P) + R - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) - \exp(nH(PV))\right] \\
&\geq \sum_{V \in \mathcal{V}_n(P)} \exp\left(-n\left[D(V||W|P) + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) \times \\
&\qquad \left[1 - \exp\left(-n\left[R - H(PV) + H(V|P) - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right)\right] \\
&\geq \sum_{V \in \mathcal{V}_n(P): I(P,V) \leq R - \delta} \exp\left(-n\left[D(V||W|P) + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) \times \\
&\qquad \left[1 - \exp\left(-n\left[R - H(PV) + H(V|P) - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right)\right] \\
&\geq \exp\left(-n\left[\min_{V \in \mathcal{V}_n(P): I(P,V) \leq R - \delta} D(V||W|P) + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1) + \tau(n, \delta, |\mathcal{X}|, |\mathcal{Y}|)\right]\right),
\end{aligned}$$

where

$$\tau(n, \delta, |\mathcal{X}|, |\mathcal{Y}|) \triangleq -\frac{1}{n}\log\left(1 - \exp\left(-n\left[\delta - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right)\right).$$

Note that $\tau(n, \delta, |\mathcal{X}|, |\mathcal{Y}|) \to 0$ as $n \to \infty$ and is $O(1/n)$. From Lemma 6, we know that

$$\min_{V \in \mathcal{V}_n(P): I(P,V) \leq R - \delta} D(V||W|P) \leq E_{sp}\left(R - \delta - 2\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log n\right) + \kappa_W \frac{|\mathcal{X}||\mathcal{Y}|}{n} + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log\frac{n}{|\mathcal{X}|}.$$

Therefore,

$$\begin{aligned}
-\frac{1}{n}\log P_e(W) \leq E_{sp}\left(R - \delta - 2\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log n\right) + \kappa_W \frac{|\mathcal{X}||\mathcal{Y}|}{n} + \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log\frac{n}{|\mathcal{X}|} + \\
\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1) + \tau(n, \delta, |\mathcal{X}|, |\mathcal{Y}|).
\end{aligned}$$

Now, the above is true when we assume that the fixed type encoding tree condition holds with the same P for all messages. Removing the requirement that the condition holds with the same P for all messages means we need to thin the code before applying the argument that got us to the point above. Since there are at most $(n+1)^{|\mathcal{X}|}$ types P that the condition can hold with for each message, we have

$$-\frac{1}{n}\log P_e(W) \leq E_{sp}\left(R-\delta-2\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log n-\frac{|\mathcal{X}|}{n}\log(n+1)\right)+\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log\frac{n}{|\mathcal{X}|}+$$
$$\kappa\frac{|\mathcal{X}||\mathcal{Y}|}{n}+\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)+\tau(n,\delta,|\mathcal{X}|,|\mathcal{Y}|)+\frac{|\mathcal{X}|}{n}\log(n+1)$$

Since $\log(n)/n \to 0$ as $n \to \infty$, we have for large enough n depending on δ, R and W that

$$-\frac{1}{n}\log P_e(W) \leq E_{sp}(R-2\delta)+2\delta.$$

A.10 Limited memory at the encoder means the sphere packing bound holds

Proposition 5 of Section 2.9: The following relationship between $E_{sp}(R)$ and $E'_{sp}(R)$ holds for all $k \geq 2$ (i.e. all interesting k):

$$E'_{sp}(R) = kE_{sp}\left(\frac{R}{k}\right).$$

This result is due to Baris Nakiboglu and Giacomo Como [33].

Proof: We will show this for the case of $k=2$ and argue the rest by induction. First, let $\mathcal{S}'$ be the set of test channels $V' \in \mathcal{W}'$ of the form

$$V'(y_1,y_2|x_1,f_2) = V_1(y_1|x_1)V_2(y_2|f_2(y_1))$$

for some $V_1, V_2 \in \mathcal{W}$. So, V' is a succession of two possibly different DMCs. Then, for $V' \in \mathcal{S}' \subset \mathcal{W}'$,

$$\begin{aligned}
D(V'||W'|P') &= \sum_{x_1,f_2}\sum_{y_1,y_2} P'(x_1,f_2)V'(y_1,y_2|x_1,f_2)\log\frac{V'(y_1,y_2|x_1,f_2)}{W'(y_1,y_2|x_1,f_2)} \\
&= \sum_{x_1,f_2}\sum_{y_1,y_2} P'(x_1,f_2)V_1(y_1|x_1)V_2(y_2|f_2(y_1))\log\frac{V_1(y_1|x_1)V_2(y_2|f_2(y_1))}{W(y_1|x_1)W(y_2|f_2(y_1))} \\
&= \sum_{x_1,f_2}\sum_{y_1,y_2} P'(x_1,f_2)V_1(y_1|x_1)V_2(y_2|f_2(y_1))\log\frac{V_1(y_1|x_1)}{W(y_1|x_1)}+ \\
&\quad \sum_{x_1,f_2}\sum_{y_1,y_2} P'(x_1,f_2)V_1(y_1|x_1)V_2(y_2|f_2(y_1))\log\frac{V_2(y_2|f_2(y_1))}{W(y_2|f_2(y_1))}.
\end{aligned}$$

Now, for the first term above,

$$
\begin{aligned}
T_1 &\triangleq \sum_{x_1,f_2}\sum_{y_1,y_2} P'(x_1,f_2)V_1(y_1|x_1)V_2(y_2|f_2(y_1))\log\frac{V_1(y_1|x_1)}{W(y_1|x_1)} \\
&= \sum_{x_1,y_1}\left[\sum_{f_2} P'(x_1,f_2)\sum_{y_2} V_2(y_2|f_2(y_1))\right] V_1(y_1|x_1)\log\frac{V_1(y_1|x_1)}{W(y_1|x_1)} \\
&\overset{(a)}{=} \sum_{x_1,y_1}\left[\sum_{f_2} P'(x_1,f_2)\right] V_1(y_1|x_1)\log\frac{V_1(y_1|x_1)}{W(y_1|x_1)} \\
&= D(V_1||W|P_{X_1}) \qquad \text{(A.42)} \\
P_{X_1} &\triangleq \sum_{f_2} P'(x_1,f_2),
\end{aligned}
$$

where (a) follows because for each f_d, y_1, $\sum_{y_2} V(y_2|f_2(y_1)) = 1$. Note that P_{X_1} depends only on P'. For the second term,

$$
\begin{aligned}
T_2 &\triangleq \sum_{x_1,f_2}\sum_{y_1,y_2} P'(x_1,f_2)V_1(y_1|x_1)V_2(y_2|f_2(y_1))\log\frac{V_2(y_2|f_2(y_1))}{W(y_2|f_2(y_1))} \\
&= \sum_{x_1,f_2}\sum_{y_1,y_2}\sum_{x_2} 1(f_2(y_1)=x_2)P'(x_1,f_2)V_1(y_1|x_1)V_2(y_2|f_2(y_1))\log\frac{V_2(y_2|x_2)}{W(y_2|x_2)} \\
&= \sum_{x_2,y_2}\left[\sum_{x_1,y_1,f_2:f_2(y_1)=x_2} P'(x_1,f_2)V_1(y_1|x_1)\right] V_2(y_2|f_2(y_1))\log\frac{V_2(y_2|x_2)}{W(y_2|x_2)} \\
&= \sum_{x_2,y_2} P_{X_2}(x_2)V_2(y_2|f_2(y_1))\log\frac{V_2(y_2|x_2)}{W(y_2|x_2)} \\
&= D(V_2||W|P_{X_2}) \qquad \text{(A.43)} \\
P_{X_2}(x_2) &\triangleq \sum_{x_1,y_1,f_2:f_2(y_1)=x_2} P'(x_1,f_2)V_1(y_1|x_1),
\end{aligned}
$$

where we note that $P_{X_2}(\cdot)$ is a distribution because it is nonnegative and sums to 1. Note that P_{X_2} depends on P' and V_1 but not on V_2. Combining (A.42) and (A.43) yields

$$
D(V'||W'|P') = D(V_1||W|P_{X_1}) + D(V_2||W|P_{X_2}). \qquad \text{(A.44)}
$$

Now, it is well known [47] that for $P \in \mathcal{P}, V \in \mathcal{V}$,

$$
\begin{aligned}
I(P,V) &= D(V||(PV)|P) \\
&= \min_{Q\in\mathcal{Q}} D(V||Q|P),
\end{aligned}
$$

where the divergence between a conditional distribution V and a distribution Q on $\mathcal{Y}$ conditioned on P is

$$D(V||Q|P) = \sum_{x,y} P(x)V(y|x) \log \frac{V(y|x)}{Q(y)}.$$

With this property in mind,

$$\begin{aligned}
I(P', V') &= \min_{Q' \in \mathcal{Q}'} \sum_{x_1, f_2, y_1, y_2} P'(x_1, f_2) V_1(y_1|x_1) V_2(y_2|f_2(y_1)) \log \frac{V_1(y_1|x_1)V_2(y_2|f_2(y_1))}{Q(y_1, y_2)} \\
&\leq \sum_{x_1, f_2, y_1, y_2} P'(x_1, f_2) V_1(y_1|x_1) V_2(y_2|f_2(y_1)) \log \frac{V_1(y_1|x_1)V_2(y_2|f_2(y_1))}{(P_{X_1}V_1)(y_1)(P_{X_2}V_2)(y_2)} \\
&= \sum_{x_1, f_2, y_1, y_2} P'(x_1, f_2) V_1(y_1|x_1) V_2(y_2|f_2(y_1)) \log \frac{V_1(y_1|x_1)}{(P_{X_1}V_1)(y_1)} + \\
&\quad\ \sum_{x_1, f_2, y_1, y_2} P'(x_1, f_2) V_1(y_1|x_1) V_2(y_2|f_2(y_1)) \log \frac{V_2(y_2|f_2(y_1))}{(P_{X_2}V_2)(y_2)} \\
&= \sum_{x_1, y_1} \left[\sum_{f_2, y_2} P'(x_1, f_2) V_2(y_2|f_2(y_1)) \right] V_1(y_1|x_1) \log \frac{V_1(y_1|x_1)}{(P_{X_1}V_1)(y_1)} + \\
&\quad\ \sum_{x_1, f_2, y_1, y_2} P'(x_1, f_2) V_1(y_1|x_1) V_2(y_2|f_2(y_1)) \log \frac{V_2(y_2|f_2(y_1))}{(P_{X_2}V_2)(y_2)} \\
&= \sum_{x_1, y_1} P_{X_1}(x_1) V_1(y_1|x_1) \log \frac{V_1(y_1|x_1)}{(P_{X_1}V_1)(y_1)} + \\
&\quad\ \sum_{x_1, f_2, y_1, y_2} P'(x_1, f_2) V_1(y_1|x_1) V_2(y_2|f_2(y_1)) \log \frac{V_2(y_2|f_2(y_1))}{(P_{X_2}V_2)(y_2)} \\
&= I(P_{X_1}, V_1) + \sum_{x_1, f_2, y_1, y_2} P'(x_1, f_2) V_1(y_1|x_1) V_2(y_2|f_2(y_1)) \log \frac{V_2(y_2|f_2(y_1))}{(P_{X_2}V_2)(y_2)} \\
&= I(P_{X_1}, V_1) + \sum_{y_2} \sum_{x_2} \left[\sum_{x_1, y_1, f_1 : f_2(y_1) = x_2} P'(x_1, f_2) V_1(y_1|x_1) \right] V_2(y_2|x_2) \log \frac{V_2(y_2|x_2)}{(P_{X_2}V_2)(y_2)} \\
&= I(P_{X_1}, V_1) + \sum_{y_2} \sum_{x_2} P_{X_2}(x_2) V_2(y_2|x_2) \log \frac{V_2(y_2|x_2)}{(P_{X_2}V_2)(y_2)}
\end{aligned}$$

$$I(P', V') \leq I(P_{X_1}, V_1) + I(P_{X_2}, V_2). \tag{A.45}$$

Using the above bound for $V' \in \mathcal{S}'$, we can upper bound $E'_{sp}(R)$ as

$$\begin{aligned}
E'_{sp}(R) &= \max_{P' \in \mathcal{P}'} \min_{V' \in \mathcal{W}' : I(P',V') \leq R} D(V'||W'|P') \\
&\leq \max_{P' \in \mathcal{P}'} \min_{V' \in \mathcal{S}' : I(P',V') \leq R} D(V'||W'|P') \\
&\overset{(a)}{\leq} \max_{P' \in \mathcal{P}'} \min_{V' \in \mathcal{S}' : I(P_{X_1},V_1) + I(P_{X_2},V_2) \leq R} D(V'||W'|P') \\
&\overset{(b)}{\leq} \max_{P' \in \mathcal{P}'} \min_{V' \in \mathcal{S}' : I(P_{X_1},V_1) \leq R/2,\ I(P_{X_2},V_2) \leq R/2} D(V'||W'|P') \qquad \text{(A.46)} \\
&\overset{(c)}{=} \max_{P' \in \mathcal{P}'} \min_{V' \in \mathcal{S}' : I(P_{X_1},V_1) \leq R/2,\ I(P_{X_2},V_2) \leq R/2} D(V_1||W|P_{X_1}) + D(V_2||W|P_{X_2}),
\end{aligned}$$

where (a) is due to reducing the set of admissible V', (b) follows from (A.45) and (c) follows from (A.44). Now, we have

$$\begin{aligned}
E'_{sp}(R) &\leq \max_{P' \in \mathcal{P}'} \min_{V_1,V_2 : I(P_{X_1},V_1) \leq R/2,\ I(P_{X_2},V_2) \leq R/2} D(V_1||W|P_{X_1}) + D(V_2||W|P_{X_2}) \\
&\overset{(d)}{\leq} \max_{P_{X_1},P_{X_2}} \min_{V_1 : I(P_{X_1},V_1) \leq R/2} D(V_1||W|P_{X_1}) + \min_{V_2 : I(P_{X_2},V_2) \leq R/2} D(V_2||W|P_{X_2}) \\
&= 2E_{sp}(R/2),
\end{aligned}$$

where (d) follows because while P_{X_2} is induced by P' and V_1, allowing it to be arbitrary can only make the max larger.

The above proof can be extended to k symbols as follows. First, note that we never used anywhere explicitly that the first and second symbol within the supersymbol are from the same alphabet, so for example we can have $\mathcal{X}' = \mathcal{X} \times \mathcal{X}^{k-1}$ and similarly for $\mathcal{Y}'$ and $\mathcal{W}'$. Secondly, the bound from the line above (A.46) to (A.46) does not have to be an $R/2$, $R/2$ split in the mutual information. We can just as well require $I(P_{X_1}, V_1) \leq R/k$ and $I(P_{X_2}, V_2) \leq (k-1)R/k$. In the end, we would have

$$E'_{sp}(R) \leq E_{sp,1}(R/k) + E_{sp,2}((k-1)R/k), \qquad \text{(A.47)}$$

where $E_{sp,1}$ denotes the sphere packing bound of the first channel and $E_{sp,2}$ denotes the sphere packing bound for the second channel. So to prove the proposition by induction, assume that we have shown for all $l \leq k$ (we have already done the base case of $k = 2$),

$$E^{(l)}_{sp}(R) \leq lE_{sp}(R/l),$$

where $E^{(l)}_{sp}(R)$ denotes the sphere packing bound of the superchannel composed of l symbols from $\mathcal{X}$. Then, using (A.47),

$$\begin{aligned} E_{sp}^{(k+1)}(R) &\leq E_{sp}(R/(k+1)) + E_{sp}^{(k)}(kR/(k+1)) \\ &\overset{(e)}{\leq} E_{sp}(R/(k+1)) + kE_{sp}(R/(k+1)) \\ &= (k+1)E_{sp}(R/(k+1)), \end{aligned}$$

where (e) follows from the inductive hypothesis. As for the lower bound that $E'_{sp}(R) \geq kE_{sp}(R/k)$, it follows obviously from the fact that the codes do not have to use feedback at all, that is P' can be a product distribution on $\mathcal{X}^k$, in which case the optimization turns into k independent optimizations over test channels. Using the fact that both mutual information and divergence are convex in the channel V yields the desired lower bound. We only outline this portion because we actually only need the upper bound for our purposes.

A.11 Equivalent statements

A.11.1 S-P encoding tree condition implies SP holds with feedback

Proposition 3 of Section 2.6: If the sphere-packing encoding tree condition holds, then sphere-packing for fixed-length codes with feedback holds.

Proof: Fix a $\delta > 0$, and block code with feedback of rate R and length $n \geq n_{ET,fb}(\delta, R, W)$. If the sphere-packing encoding tree condition holds, we know that for each message encoding tree, there exists a $P \in \mathcal{P}_n, V \in \mathcal{V}_n(P)$ such that (2.26), (2.27) and (2.28) hold. Now, the (P, V) can be different for different messages, but there are at most $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$ joint types (P, V), so there must be at least one (P, V) such that (2.26), (2.27) and (2.28) hold for at least $|\mathcal{M}|/(n+1)^{|\mathcal{X}||\mathcal{Y}|}$ of the messages. Fixing this P, V, and letting $\mathcal{M}' \subset \mathcal{M}$ be the set of messages for which (2.26), (2.27) and (2.28) hold, we have

$$\begin{aligned} P_e(W) &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m^c | M = m) \\ &\geq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{P}_W(Y^n \in B(m, P, V) \cap \mathcal{D}_m^c | M = m) \\ &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} |B(m, P, V) \cap \mathcal{D}_m^c| \exp(-n(D(V||W|P) + H(V|P))) \\ &\geq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}'} |B(m, P, V) \cap \mathcal{D}_m^c| \exp(-n(D(V||W|P) + H(V|P))) \\ &= \exp(-n(R + D(V||W|P) + H(V|P)) \sum_{m \in \mathcal{M}'} |B(m, P, V) \cap \mathcal{D}_m^c|. \qquad \text{(A.48)} \end{aligned}$$

At this point, we can use the properties of $|B(m,P,V)|, D(V||W|P)$ and $I(P,V)$ that are provided by the sphere-packing encoding tree condition.

$$\begin{aligned}
\sum_{m\in\mathcal{M}'} |B(m,P,V)\cap\mathcal{D}_m^c| &= \sum_{m\in\mathcal{M}'} |B(m,P,V)| - |B(m,P,V)\cap\mathcal{D}_m| \\
&\geq \sum_{m\in\mathcal{M}'} |B(m,P,V)| - |T_{PV}| \\
&\geq \sum_{m\in\mathcal{M}'} |B(m,P,V)| - \exp(nH(PV)) \\
&\geq \frac{|\mathcal{M}|}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}} \exp(n(H(V|P)-\delta)) - \exp(nH(PV)) \\
&= \exp\left(n\left[H(V|P)+R-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)-\delta\right]\right)\times \\
&\qquad \left[1-\exp\left(-n\left[H(V|P)-H(PV)+R-\delta-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right)\right] \\
&= \exp\left(n\left[H(V|P)+R-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)-\delta\right]\right)\times \\
&\qquad \left[1-\exp\left(-n\left[R-I(P,V)-\delta-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right)\right] \\
&\geq \exp\left(n\left[H(V|P)+R-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)-\delta\right]\right)\times \\
&\qquad \left[1-\exp\left(-n\left[2\delta-\delta-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right)\right] \\
&= \exp\left(n\left[H(V|P)+R-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)-\delta\right]\right)\times \\
&\qquad \left[1-\exp\left(-n\left[\delta-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right)\right] \\
&\geq \frac{1}{2}\exp\left(n\left[H(V|P)+R-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)-\delta\right]\right), \qquad \text{(A.49)}
\end{aligned}$$

where the last line holds for large enough n depending on $\delta, |\mathcal{X}|$ and $|\mathcal{Y}|$. Plugging the inequality of (A.49) into (A.48) yields that for large enough n,

$$\begin{aligned}
P_e(W) &\geq \frac{1}{2}\exp\left(-n\left[D(V||W|P)+\delta+\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) \\
&\geq \frac{1}{2}\exp\left(-n\left[E_{sp}(R-2\delta)+\delta+\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right).
\end{aligned}$$

Since $\delta > 0$ can be made arbitrarily small and $\log(n+1)/n \to 0$ as $n \to \infty$, sphere-packing then holds with feedback.

A.11.2 Intermediate S-P condition and S-P holding with feedback are equivalent

Proposition 4 of Section 2.6: If SP holds with feedback, then the intermediate SP condition holds. Conversely, if the intermediate SP condition holds, SP holds with feedback.

Proof: First, assume that SP holds with feedback. Fix a $\delta > 0$ and a block code with feedback of rate R and length $n \geq n_{SP,fb}(\delta, R, W)$. Now, because SP holds with feedback,

$$P_e(W) \geq \exp\left(-n(E_{sp}(R-\delta)+\delta)\right).$$

Expanding the error probability,

$$\begin{aligned}
P_e(W) &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \notin \mathcal{D}_m | M = m) \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\sum_{P\in\mathcal{P}_n, V\in\mathcal{V}_n(P)} \mathbb{P}_W(Y^n \in B(m,P,V)\cap\mathcal{D}_m^c | M = m) \\
&\overset{(a)}{=} \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\sum_{P\in\mathcal{P}_n, V\in\mathcal{V}_n(P)} |B(m,P,V)\cap\mathcal{D}_m^c| \exp(-n(D(V||W|P) + H(V|P))) \\
&= \sum_{P\in\mathcal{P}_n, V\in\mathcal{V}_n(P)} \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} |B(m,P,V)\cap\mathcal{D}_m^c| \exp(-n(D(V||W|P) + H(V|P))) \\
&= \sum_{P\in\mathcal{P}_n, V\in\mathcal{V}_n(P)} K(P,V) \\
K(P,V) &\triangleq \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} |B(m,P,V)\cap\mathcal{D}_m^c| \exp(-n(D(V||W|P) + H(V|P))), \qquad \text{(A.50)}
\end{aligned}$$

where in the above, (a) follows from Proposition 14. Note that all the steps above are equalities. Hence, we know that

$$\sum_{P\in\mathcal{P}_n, V\in\mathcal{V}_n(P)} K(P,V) \geq \exp(-n(E_{sp}(R-\delta)+\delta)).$$

Since all the $K(P,V)$ are non-negative, and there are at most $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$ joint types of length n, it follows by the pigeonhole principle that there is at least one $P \in \mathcal{P}_n$ and $V \in \mathcal{V}_n(P)$

such that

$$
\begin{aligned}
K(P,V) &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} |B(m,P,V)\cap\mathcal{D}_m^c|\exp(-n(D(V||W|P)+H(V|P))) \\
&\geq \frac{1}{(n+1)^{|\mathcal{X}||\mathcal{Y}|}}\exp(-n(E_{sp}(R-\delta)+\delta)) \\
\frac{1}{n}\log K(P,V) &= -D(V||W|P)-H(V|P)+\frac{1}{n}\log\left[\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}|B(m,P,V)\cap\mathcal{D}_m^c|\right] \\
&\geq -E_{sp}(R-\delta)-\delta-\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1).
\end{aligned}
$$

Rearranging yields

$$
\begin{aligned}
\zeta &\triangleq \frac{1}{n}\log\left[\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}|B(m,P,V)\cap\mathcal{D}_m^c|\right] && \text{(A.51)} \\
&\geq H(V|P)-\left[E_{sp}(R-\delta)-D(V||W|P)+\delta+\frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right] \\
&\geq H(V|P)-[E_{sp}(R-\delta)-D(V||W|P)+2\delta] && \text{(A.52)}
\end{aligned}
$$

for n large enough, as $\log(n+1)/n$ goes to 0 as $n\to\infty$. Now,

$$
\begin{aligned}
\frac{1}{n}\log\left[\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}|B(m,P,V)\cap\mathcal{D}_m^c|\right] &\leq \frac{1}{n}\log\left[\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}|B(m,P,V)|\right] \\
&\overset{(b)}{\leq} \frac{1}{n}\log\left[\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\exp(nH(V|P))\right] \\
&= H(V|P), && \text{(A.53)}
\end{aligned}
$$

where (b) comes from Proposition 14. Plugging the inequality of (A.53) into (A.52) implies that additionally,

$$
D(V||W|P)\leq E_{sp}(R-\delta)+2\delta.
$$

Hence, the intermediate sphere packing condition holds.

Now, suppose the intermediate sphere packing condition holds. That is, for $n\geq n_{SPI,fb}(\delta,R,W)$ there exists a $P\in\mathcal{P}_n, V\in\mathcal{V}_n(P)$ such that (2.30) and (2.31) hold. Then, plugging (2.31) into (A.50) and working backwards through the inequalities gives us that the sphere-packing bound holds with feedback.

A.12 Delayed feedback

A.12.1 Error probability bound

Lemma 5 of Section 2.7: Define a channel independent constant

$$\alpha(T) \triangleq \frac{|\mathcal{X}|(2+|\mathcal{Y}|)\log(T+1)}{T}.$$

Fix an $\epsilon > \alpha(T)$. Then, for any block length $n = NT$ (with $N \geq 1$) rate R coding system with a type 2 encoder,

$$-\frac{1}{NT}\log P_e(NT,\mathcal{E},\mathcal{D}) \leq E_{sp,T}(R-\epsilon) + \alpha(T) + \gamma(N,T,\epsilon),$$

where

$$E_{sp,T}(R) \triangleq \max_{P\in\mathcal{P}_T}\ \min_{V\in\mathcal{V}_T(P):I(P,V)\leq R} D(V||W|P), \tag{A.54}$$

and

$$\gamma(N,T,\epsilon) = \frac{1}{NT}\log\frac{1}{1-\exp(-NT(\epsilon-\alpha(T)))}.$$

Proof: The argument begins, as with the sphere-packing proof for codes without feedback, by showing that there is a subcode of approximate rate R for which most codewords have the same input types. This is normally done by whittling down the messages to those whose codewords belong to the largest common type, by message population. The challenge is that now we have feedback every T symbols, so we need to carefully show in what sense messages have the same input types. We do this by induction on N, the total number of T-length blocks.

First, for $N = 1$, there has been no feedback. Let $P_1(m)$ denote the type of $x^T(m) \triangleq (\phi_1(m),\ldots,\phi_T(m))$. Now, group messages according to their type $P_1(m)$. Since $|\mathcal{P}_T|$ is at most $(T+1)^{|\mathcal{X}|}$, there exists a $P_1 \in \mathcal{P}_T$ such that

$$|\{m : P_1(m) = P_1\}| \geq \frac{2^{nR}}{(T+1)^{|\mathcal{X}|}}.$$

This is the usual argument for fixing the composition of a high-rate subcode in the proof of the sphere-packing bound for codes used without feedback. After this, we choose a $V_1 \in \mathcal{V}_T(P_1)$ such that $I(P_1,V_1) \leq R-\epsilon$. The choice of V_1 is made so as to minimize $D(V_1||W|P_1)$ amongst those $V_1 \in \mathcal{V}_T(P_1)$ that have $I(P_1,V_1) \leq R-\epsilon$. The existence of a V_1 such that $I(P_1,V_1) \leq R-\epsilon$ is not immediately obvious for channels in which $E_{sp}(R)$ can be infinite, even if $E_{sp}(R)$ is not infinite for the given R. If no such V_1 exists, the result of the lemma is meaningless if we take the convention that min over the null set is ∞. Hence, if the optimization in the right hand side of (2.32) evaluates to something finite, we can safely assume the existence, for each P_1, of a V_1 with $I(P_1, V_1) \leq R - \epsilon$. For the rest of the proof, we assume we are in this case.

Without feedback, the selection at this point would be enough to show that for a high-rate subcode with the same type, a substantial portion of the selected V_1-shells (as defined in [47]) around the codewords for these messages overlap to cause a significant error. However, we now have feedback every T symbols, so we will iterate this selection process N times. We will prove a claim showing that the messages are not thinned too much and there are many y^n sequences which must 'overlap'. We will do this by selecting N input types $P_1, \ldots, P_N$ and N channel shells $V_1, \ldots, V_N$ sequentially by induction. Let, for $1 \leq k \leq N$,

$$B^{(k)}(m) \triangleq \left\{ y^{kT} : \begin{array}{c} y^{(k-1)T} \in B^{(k-1)}(m), P_k(m, y^{(k-1)T}) = P_k, \\ y^{kT}_{(k-1)T+1} \in T_{V_k}\left(x^{kT}_{(k-1)T+1}(m, y^{(k-1)T})\right) \end{array} \right\},$$

where $P_k(m, y^{(k-1)T})$ is the type of $x^{kT}_{(k-1)T+1}(m, y^{(k-1)T})$, which is itself defined as

$$\left(\phi_{(k-1)T+1}(m, y^{(k-1)T}), \ldots, \phi_{kT}(m, y^{(k-1)T})\right).$$

Now, let for $1 \leq k \leq N$,

$$A^{(k)} \triangleq \left\{ m : |B^{(k)}(m)| \geq \frac{\exp_2(T \sum_{i=1}^{k} H(V_i|P_i))}{(T+1)^{k(|\mathcal{X}|+|\mathcal{X}||\mathcal{Y}|)}} \right\},$$

where $H(V|p) = \sum_{x,y} p(x)V(y|x)\log(1/V(y|x))$ denotes conditional entropy. Note the dependence of both the $A^{(k)}$ and $B^{(k)}(m)$ sets on P_i and V_i. We drop the dependence in the notation for convenience. In words, we are breaking the transmission length into blocks of length T, and then coming up with a set of messages that have common types within each block as long as the output sequence being fed back lies exactly in a certain V-shell. The set $B^{(k)}(m)$ keeps track of how many of these output sequences lead to the common type for each message and the set $A^{(k)}$ keeps track of how many of the messages have the desired lower bound on size for $B^{(k)}(m)$.

Claim 3. *For a block length $n = NT$, rate R type 2 encoder, there exist $P_1, \ldots, P_N \in \mathcal{P}_T$ and $V_1, \ldots, V_N$, with $V_i \in \arg\min_{V \in \mathcal{V}_T(P_i): I(P_i,V) \leq R-\epsilon} D(V||W|P_i)$ such that*

$$|A^{(N)}| \geq \frac{2^{nR}}{(T+1)^{N|\mathcal{X}|}}.$$

Proof: We proceed by induction with the base case of $N = k = 1$. As we have seen there is a $P_1 \in \mathcal{P}_T$ such that at least $2^{nR}/(T+1)^{|\mathcal{X}|}$ messages have $P_1(m) = P_1$. We then choose $V_1 \in \mathcal{V}_T(P_1)$ such that $I(P_1, V_1) \leq R - \epsilon$. It is clear that for those m with $P_1(m) = P_1$,

$$|B^{(1)}(m)| = |T_{V_1}(P_1)| \geq \frac{1}{(T+1)^{|\mathcal{X}||\mathcal{Y}|}} \exp_2(TH(V_1|P_1)),$$

where $|T_V(p)|$ denotes the number of vectors in a V-shell around a vector of type p (i.e. $|T_V(x^T)|$ if x^T is of type p). Hence, the claim is true for $N = 1$.

Now for $N = k > 1$, assume the claim is true for $k-1$. For each $m \in A^{(k-1)}$, group the $y^{(k-1)T} \in B^{(k-1)}(m)$ according to $P_k(m, y^{(k-1)T})$. At least $|B^{(k-1)}(m)|/(T+1)^{|\mathcal{X}|}$ of the $y^{(k-1)T} \in B^{(k-1)}(m)$ have a common type $P_k(m, y^{(k-1)T}) = P_k(m)$. Now, group the messages in $A^{(k-1)}$ according to $P_k(m)$. At least $|A^{(k-1)}|/(T+1)^{|\mathcal{X}|}$ have a common type $P_k(m) = P_k$. Now select a $V_k \in \mathcal{V}_T(P_k)$ such that $I(P_k, V_k) \leq R - \epsilon$. It is now readily seen that for the $m \in A^{(k-1)}$ with $P_k(m) = P_k$,

$$\begin{aligned} |B^{(k)}(m)| &\geq \frac{|B^{(k-1)}(m)|}{(T+1)^{|\mathcal{X}|}} |T_{V_k}(P_k)| \\ &\geq \frac{\exp_2\left(T \sum_{i=1}^{k} H(V_i|P_i)\right)}{(T+1)^{k(|\mathcal{X}|+|\mathcal{X}||\mathcal{Y}|)}}. \end{aligned}$$

This holds for at least $|A^{(k-1)}|/(T+1)^{|\mathcal{X}|} \geq 2^{nR}/(T+1)^{k|\mathcal{X}|}$ messages, hence the claim is true.

Now, note that for all $y^n \in B^{(N)}(m)$ with $m \in A^{(N)}$, we also have $y_{(i-1)T+1}^{iT} \in T_{P_i V_i}$ for all $1 \leq i \leq N$. Hence,

$$\begin{aligned} \forall\, m,\ B^{(N)}(m) &\subset T_{P_1 V_1} \times \cdots \times T_{P_N V_N} \\ \therefore |\cup_{m \in A^{(N)}} B^{(N)}(m)| &\leq |T_{P_1 V_1}| \times \cdots \times |T_{P_N V_N}| \\ &\leq \exp_2\left(T \sum_{i=1}^{N} H(P_i V_i)\right), \end{aligned} \tag{A.55}$$

where $H(PV)$ is the entropy of the distribution $(PV)(y) = \sum_x p(x) V(y|x)$. Focusing our attention on these output sequences,

$$\begin{aligned} P_e(n, \mathcal{E}, \mathcal{D}) &= \frac{1}{2^{nR}} \sum_{m=1}^{2^{nR}} \sum_{y^n \notin \psi^{-1}(m)} \prod_{i=1}^{n} W\left(y_i \middle| \phi_i\left(m, y^{\lfloor i/T \rfloor T}\right)\right) \\ &\geq \frac{1}{2^{nR}} \sum_{m \in A^{(N)}} \sum_{y^n \in \overline{\psi^{-1}(m)} \cap B^{(N)}(m)} P_W(y^n | M = m) \\ &\overset{(a)}{=} \frac{\exp_2\left(-T \sum_{i=1}^{N} (D(V_i || W | P_i) + H(V_i | P_i))\right)}{2^{nR}} \sum_{m \in A^{(N)}} |\overline{\psi^{-1}(m)} \cap B^{(N)}(m)| \\ &= \frac{\exp_2\left(-T \sum_{i=1}^{N} (D(V_i || W | P_i) + H(V_i | P_i))\right)}{2^{nR}} \times \\ &\qquad \sum_{m \in A^{(N)}} (|B^{(N)}(m)| - |B^{(N)}(m) \cap \psi^{-1}(m)|), \end{aligned}$$

where in (a), we have noted that the probability for all $y^n \in B^{(N)}(m)$ is equal to

$$\exp_2\left(-T\sum_{i=1}^{N}(D(V_i||W|P_i)+H(V_i|P_i))\right).$$

Continuing, we have

$$\begin{aligned}
P_e(n,\mathcal{E},\mathcal{D}) &\geq \frac{\exp_2\left(-T\sum_{i=1}^{N}(D(V_i||W|P_i)+H(V_i|P_i))\right)}{2^{nR}}\times \\
&\qquad \left[\left(\sum_{m\in A^{(N)}}|B^{(N)}(m)|\right)-\left|\bigcup_{m\in A^{(N)}}B^{(N)}(m)\right|\right] \\
&\geq \frac{\exp_2\left(-T\sum_{i=1}^{N}(D(V_i||W|P_i)+H(V_i|P_i))\right)}{2^{nR}}\times \\
&\qquad \left(\frac{|A^{(N)}|2^{T\sum_{i=1}^{N}H(V_i|P_i)}}{(T+1)^{N(|\mathcal{X}|+|\mathcal{X}||\mathcal{Y}|)}}-\left|\bigcup_{m\in A^{(N)}}B^{(N)}(m)\right|\right) \\
&\overset{(b)}{\geq} \frac{\exp_2(-T\sum_{i=1}^{N}D(V_i||W|P_i))}{2^{nR}}\left[\frac{|A^{(N)}|}{(T+1)^{N(|\mathcal{X}|+|\mathcal{X}||\mathcal{Y}|)}}-\exp_2\left(T\sum_{i=1}^{N}I(P_i,V_i)\right)\right] \\
&\overset{(c)}{\geq} \exp_2(-T\sum_{i=1}^{N}D(V_i||W|P_i))\left[\frac{1}{(T+1)^{N(2+|\mathcal{Y}|)|\mathcal{X}|}}-\exp_2\left(-NT\epsilon\right)\right] \\
&= \frac{\exp_2\left(-T\sum_{i=1}^{N}D(V_i||W|P_i)\right)}{\exp_2(NT\alpha(T))}\left[1-2^{-NT(\epsilon-\alpha(T))}\right].
\end{aligned}$$

In inequality (b), we have used the inequality of equation (A.55). Claim 3 is used in inequality (c). In the selection process of the claim, for each $P_i \in \mathcal{P}_T$, we choose a $V_i \in \mathcal{V}_T(p)$ with $I(P_i,V_i) \leq R-\epsilon$ that minimizes the average divergence. Then, since we can't say anything about the P_i, we bound by the worst-case P to take a max over all $P \in \mathcal{P}_T$. Taking logs and dividing by NT gives the result of the lemma.

A.12.2 Length T sphere-packing exponent

Lemma 6 of Section 2.7: For any $T \geq 2|\mathcal{X}||\mathcal{Y}|$, for all $P \in \mathcal{P}_T$,

$$\begin{aligned}
\min_{U\in\mathcal{V}_T(P):I(P,U)\leq R} D(U||W|P) \leq E_{sp}\left(R-\frac{2|\mathcal{X}||\mathcal{Y}|\log T}{T},P\right)+\frac{\kappa_W|\mathcal{X}||\mathcal{Y}|}{T}+ \\
\frac{|\mathcal{X}||\mathcal{Y}|\log(T/|\mathcal{X}|)}{T},
\end{aligned}$$

where

$$E_{sp}(R,P) \triangleq \min_{V:I(P,V)\leq R} D(V||W|P),$$
$$\kappa_W \triangleq \max_{x,y:W(y|x)>0} \log \frac{1}{W(y|x)}.$$

Proof: First, write $P \in \mathcal{P}_T$ as $P(x) = k_x/T$ where k_x are nonnegative integers that sum to T.

Claim 4. *Let U be an arbitrary channel for which $|U(y|x) - V(y|x)| \leq 1/k_x$ for all x, y, and $U(y|x) = V(y|x) = 0$ when $W(y|x) = 0$. Then,*

$$|D(U||W|P) - D(V||W|P)| \leq |\mathcal{X}||\mathcal{Y}| \frac{\kappa_W + \log(T/|\mathcal{X}|)}{T}.$$

Proof: First, note that $|r \log r - s \log s| \leq -|r-s| \log |r-s|$ whenever $r, s \in [0,1]$. This can be seen by noting that the function $f(r) = -r \log r, r \in [0,1]$ is concave and maximal absolute slope at $r = 0$, where the derivative is unbounded above. Hence, $|f(r) - f(0)| = -r \log r, r \in [0,1]$ is a bound to the difference between two points on the curve at distance r. Now, keeping in mind that $U(y|x) = V(y|x) = 0$ whenever $W(y|x) = 0$, a little algebra shows that

$$\begin{aligned}
|D(U||W|P) - D(V||W|P)| &\leq \sum_x P(x) \sum_y \left| U(y|x) \log \frac{U(y|x)}{W(y|x)} - V(y|x) \log \frac{V(y|x)}{W(y|x)} \right| \\
&\leq \sum_x \frac{k_x}{T} \sum_y |U(y|x) \log U(y|x) - V(y|x) \log V(y|x)| + \\
&\qquad \sum_x \frac{k_x}{T} \sum_{y:W(y|x)>0} |U(y|x) - V(y|x)| \log \frac{1}{W(y|x)} \\
&\leq \sum_x \frac{k_x}{T} \sum_y \left[\frac{1}{k_x} \log k_x + \frac{1}{k_x} \kappa_W \right] \\
&\leq \frac{|\mathcal{X}||\mathcal{Y}|\kappa_W}{T} + \frac{|\mathcal{Y}|}{T} \sum_x \log k_x \\
&\overset{(a)}{\leq} \frac{|\mathcal{X}||\mathcal{Y}|\kappa_W}{T} + \frac{|\mathcal{X}||\mathcal{Y}| \log T/|\mathcal{X}|}{T}.
\end{aligned}$$

In (a), we are also using the fact that since log is a symmetric, concave ($\cap$) function, so

$$\max_{k_x \in \mathbb{N}: \sum_x k_x = T} \sum_x \log k_x \leq |\mathcal{X}| \log \frac{T}{|\mathcal{X}|}.$$

Now, for an $\epsilon > 0$, pick V to be in $\arg\min_{V' : I(P,V') \leq R - \epsilon} D(V' || W|P)$. We will find a $U \in \mathcal{V}_T(P)$ such that $I(P,U) \leq I(P,V) + \epsilon \leq R$. First, we show that there exists a $U \in \mathcal{V}_T(P)$ such that $|U(y|x) - V(y|x)| \leq \frac{1}{k_x}$ for all x, y.

For each x, y, let $\widetilde{U}(y|x) = \lfloor k_x V(y|x) \rfloor / k_x$. Note that $\widetilde{U}$ is missing some mass to be a transition matrix if the entries of $V(\cdot|x)$ are not multiples of $1/k_x$. The missing mass can be bounded, for a fixed x,

$$\begin{aligned} 1 - \sum_y \widetilde{U}(y|x) &= \sum_y V(y|x) - \widetilde{U}(y|x) \\ &\leq |\{y : k_x V(y|x) \notin \mathbb{Z}\}| / k_x. \end{aligned}$$

Now, the missing mass must be a multiple of $1/k_x$ for each x because $\widetilde{U}(\cdot|x)$ has terms that are multiples of $1/k_x$. Therefore, the missing mass can be distributed amongst the y that have $k_x V(y|x) \notin \mathbb{Z}$ in multiples of $1/k_x$ in such a way so that no y has more than $1/k_x$ mass added to it. We let $U(y|x)$ be the resulting transition matrix. Since $\widetilde{U}(y|x) = \lfloor k_x V(y|x) \rfloor / k_x$ and either 0 or $1/k_x$ is added to get to $U(y|x)$, it follows that $|V(y|x) - U(y|x)| \leq 1/k_x$. Also, $U(y|x) = 0$ when $V(y|x) = 0$.

Note that

$$\sum_x p(x) \sum_y |U(y|x) - V(y|x)| \leq \sum_x \frac{k_x}{T} \sum_y \frac{1}{k_x} \leq \frac{|\mathcal{X}||\mathcal{Y}|}{T}.$$

If $T \geq 2|\mathcal{X}||\mathcal{Y}|$, we can use the continuity lemma for entropy (Lemma 20). By using this lemma twice after expanding mutual information, we get

$$\begin{aligned} |I(P,U) - I(P,V)| &\leq |H(PU) - H(PV)| + |H(P,U) - H(P,V)| \\ &\leq \frac{2|\mathcal{X}||\mathcal{Y}|}{T} \log T. \end{aligned}$$

Hence,

$$I(P,U) \leq R - \epsilon + \frac{2|\mathcal{X}||\mathcal{Y}|}{T} \log T \leq R$$

provided

$$\frac{2|\mathcal{X}||\mathcal{Y}|}{T} \log T \leq \epsilon.$$

Therefore, there exists a $U \in \mathcal{V}_T(P)$, with $I(P,U) \leq R$ such that

$$\begin{aligned} D(U||W|P) &= E_{sp}(R - \epsilon, P) + D(U||W|P) - D(V||W|P) \\ &\leq E_{sp}\left(R - \frac{2|\mathcal{X}||\mathcal{Y}|}{T} \log T, P\right) + \frac{\kappa_W |\mathcal{X}||\mathcal{Y}|}{T} + \frac{|\mathcal{X}||\mathcal{Y}|}{T} \log \frac{T}{|\mathcal{X}|}. \end{aligned}$$

A.13 The Haroutunian exponent of a parallel channel

Lemma 7 of Section 2.8: Fix a $W \in \mathcal{W}$ and integer $L \geq 1$. Let $W^{(L)}$ denote the probability transition matrix from $\mathcal{X}^L$ to $\mathcal{Y}^L$ obtained by using L independent copies of W. That is, for $(x_1, \ldots, x_L) = x^L \in \mathcal{X}^L$ and $(y_1, \ldots, y_L) = y^L \in \mathcal{Y}^L$,

$$W^{(L)}(y^L|x^L) = \prod_{i=1}^{L} W(y_i|x_i).$$

Let $E_h(LR; W^{(L)})$ denote the Haroutunian exponent for $W^{(L)}$ at rate LR and let $E_{sp}(R; W)$ denote the sphere-packing exponent for W at rate R. Then,

$$E_h(LR; W^{(L)}) \leq LE_{sp}\left(R - \frac{|\mathcal{X}|}{L}\log(L+1); W\right).$$

Proof: First note that $C(W^{(L)}) = LC(W)$ because the copies of W in $W^{(L)}$ are independent. Let $\mathcal{W}^{(L)}$ denote the set of probability transition matrices from $\mathcal{X}^L$ to $\mathcal{Y}^L$. Then,

$$E_h(LR; W^{(L)}) = \min_{V \in \mathcal{W}^{(L)}: C(V) \leq LR} \max_{x^L} D(V(\cdot|x^L)||W^{(L)}(\cdot|x^L)). \tag{A.56}$$

For each length L type $P \in \mathcal{P}_L$, fix $\epsilon_L = \frac{|\mathcal{X}|}{L}\log(L+1)$ and

$$U_P \in \arg \min_{V' \in \mathcal{W}: I(P, V') \leq R - \epsilon_L} D(V'||W|P). \tag{A.57}$$

Then, for each $x^L \in T_P, y^L \in \mathcal{Y}^L$, define

$$V(y^L|x^L) = \prod_{i=1}^{L} U_P(y_i|x_i). \tag{A.58}$$

V is a legitimate probability transition matrix in $\mathcal{W}^{(L)}$ because it is clearly nonnegative and for each fixed $x^L \in \mathcal{X}^L$,

$$\begin{aligned}
\sum_{y^L} V(y^L|x^L) &= \sum_{y^L} \prod_{i=1}^{L} U_P(y_i|x_i) \\
&= \sum_{y_1} \cdots \sum_{y_L} \prod_{i=1}^{L} U_P(y_i|x_i) \\
&= \prod_{i=1}^{L} \sum_{y_i \in \mathcal{Y}} U_P(y_i|x_i) \\
&= \prod_{i=1}^{L} 1 = 1,
\end{aligned}$$

where P is the type of x^L. It is shown in Proposition 12 that $C(V) \leq LR$. Therefore, V is included in the minimization for $E_h(LR; W^{(L)})$. Now, fix an x^L and let P be the type of x^L, so $x^L \in T_P$. Then,

$$\begin{aligned}
D(V(\cdot|x^L)||W^{(L)}(\cdot|x^L)) &= \sum_{y^L} V(y^L|x^L) \log \frac{V(y^L|x^L)}{W^{(L)}(y^L|x^L)} \\
&= \sum_{y^L} \left[\prod_{i=1}^{L} U_P(y_i|x_i)\right] \log \frac{\prod_{j=1}^{L} U_P(y_j|x_j)}{\prod_{j=1}^{L} W(y_j|x_j)} \\
&= \sum_{y^L} \left[\prod_{i=1}^{L} U_P(y_i|x_i)\right] \left[\sum_{j=1}^{L} \log \frac{U_P(y_j|x_j)}{W(y_j|x_j)}\right] \\
&= \sum_{j=1}^{L} \sum_{y^L} \left[\prod_{i=1}^{L} U_P(y_i|x_i)\right] \log \frac{U_P(y_j|x_j)}{W(y_j|x_j)} \\
&= \sum_{j=1}^{L} \left(\sum_{y_i, i\neq j} \left[\prod_{i\neq j} U_P(y_i|x_i)\right]\right) \sum_{y_j} U_P(y_j|x_j) \log \frac{U_P(y_j|x_j)}{W(y_j|x_j)} \\
&= \sum_{j=1}^{L} \left(\sum_{y_i, i\neq j} \left[\prod_{i\neq j} U_P(y_i|x_i)\right]\right) D(U_P(\cdot|x_j)||W(\cdot|x_j)) \\
&\overset{(a)}{=} \sum_{j=1}^{L} D(U_P(\cdot|x_j)||W(\cdot|x_j)) \\
&\overset{(b)}{=} LD(U_P||W|P) \\
&\overset{(c)}{\leq} LE_{sp}(R - \epsilon_L; W),
\end{aligned}$$

where (a) follows because the term that disappears is 1, (b) follows because the type of x^L is P and (c) follows by the choice of U_P as a sphere-packing optimizing channel in (A.57). This upper bound is independent of x^L, therefore

$$\max_{x^L} D(V(\cdot|x^L)||W^{(L)}(\cdot|x^L)) \leq LE_{sp}(R - \epsilon_L; W),$$

which proves the lemma because V is included in the minimization of (A.56).

Proposition 12. *For the channel V defined in (A.57) and (A.58), $C(V) \leq LR$.*

Proof: Note that by definition[17], $C(V) = \max_{P^{(L)}(X^L)} I(X^L; Y^L)$ where $P^{(L)}(X^L)$ is a probability mass function on $\mathcal{X}^L$ and $\mathbb{P}(Y^L|X^L) = V(Y^L|X^L)$. Let Z be the random variable

[17] Recall that in our notation capital letters such as X^L denote random variables.

which takes values in $\mathcal{P}_L$ which takes the value of the type of X_L. That is, if $X_L \in T_P$, $Z = P$. Note that Z is a deterministic function of X^L. Now, for any distribution $P^{(L)}(X^L)$,

$$\begin{aligned} I(X^L, Z; Y^L) &= I(X^L; Y^L) + I(Z; Y^L | X^L) \\ &= I(X^L; Y^L), \end{aligned}$$

where the first equality follows by chain rule for mutual information and the second equality follows because Z is a function of X^L and therefore independent of Y^L given X^L. Now, if we expand the other way,

$$\begin{aligned} I(X^L, Z; Y^L) &= I(Z; Y^L) + I(X^L; Y^L | Z) \\ &\le H(Z) + I(X^L; Y^L | Z) \\ &\le |\mathcal{X}| \log(L + 1) + I(X^L; Y^L | Z), \end{aligned}$$

where the last line follows by the fact that Z takes its values in the set of types and $|\mathcal{P}_L| \le (L + 1)^{|\mathcal{X}|}$. By definition,

$$\begin{aligned} I(X^L; Y^L | Z) &= \sum_{P \in \mathcal{P}_L} \mathbb{P}(Z = P) I(X^L; Y^L | Z = P) \\ I(X^L; Y^L | Z = P) &= H(Y^L | Z = P) - H(Y^L | Z = P, X^L). \end{aligned}$$

Now, because $Z = P$ specifies that $X^L \in T_P$,

$$H(Y^L | Z = P, X^L) = \sum_{x^L \in T_P} \mathbb{P}(X^L = x^L | Z = P) H(Y^L | X^L = x^L, Z = P).$$

For any $x^L \in T_P$, because we have the Markov property $Z - X^L - Y^L$,

$$\begin{aligned}
H(Y^L|X^L = x^L, Z = P) &= \sum_{y^L} V(y^L|x^L) \log \frac{1}{V(y^L|x^L)} \\
&\overset{(a)}{=} \sum_{y^L} \left[\prod_{i=1}^{L} U_P(y_i|x_i)\right] \log \frac{1}{\prod_{j=1}^{L} U_P(y_j|x_j)} \\
&= \sum_{y^L} \left[\prod_{i=1}^{L} U_P(y_i|x_i)\right] \left[\sum_{j=1}^{L} \log \frac{1}{U_P(y_j|x_j)}\right] \\
&\overset{(b)}{=} \sum_{j=1}^{L} \sum_{y^L} \left[\prod_{i=1}^{L} U_P(y_i|x_i)\right] \log \frac{1}{U_P(y_j|x_j)} \\
&= \sum_{j=1}^{L} \sum_{y_j} \left[U_P(y_j|x_j) \log \frac{1}{U_P(y_j|x_j)}\right] \sum_{y_i, i\neq j} \prod_{i \neq j} U_P(y_i|x_i) \\
&= \sum_{j=1}^{L} \sum_{y} \left[U_P(y_j|x_j) \log \frac{1}{U_P(y_j|x_j)}\right] \\
&\overset{(c)}{=} L \sum_{x} P(x) \sum_{y} U_P(y|x) \log \frac{1}{U_P(y|x)} \\
&= LH(U_P|P),
\end{aligned}$$

where (a) follows by the definition of V for $x^L \in T_P$, (b) follows by the interchange of sums and products, and (c) follows by the fact that $x^L \in T_P$ and y_j is just a dummy variable for $y \in \mathcal{Y}$. Now, by chain rule for entropy,

$$\begin{aligned}
H(Y^L|Z = P) &= \sum_{i=1}^{L} H(Y_i|Z = P, Y_1, \ldots, Y_{i-1}) \\
&\leq \sum_{i=1}^{L} H(Y_i|Z = P)
\end{aligned}$$

because conditioning can only reduce entropy. For this fixed P, let P_i be the distribution of X_i conditioned on $Z = P$. That is $P_i(x) = \mathbb{P}(X_i = x|Z = P)$. Because we have conditioned

$Z = P$, it follows that $\frac{1}{L}\sum_{i=1}^{L} P_i = P$. Now, by the Markov property $Z - X_i - Y_i$,

$$\begin{aligned} H(Y_i|Z = P) &= \sum_y \mathbb{P}(Y_i = y|Z = P) \log \frac{1}{\mathbb{P}(Y_i = y|Z = P)} \\ &= \sum_y \left[\sum_x P_i(x)U_P(y|x)\right] \log \frac{1}{\sum_x P_i(x)U_P(y|x)} \\ &= H(P_iU_P). \end{aligned}$$

Hence,

$$\begin{aligned} H(Y^L|Z = P) &\leq \sum_{i=1}^{L} H(Y_i|Z = P) \\ &= L\sum_{i=1}^{L} \frac{1}{L} H(P_iU_P) \\ &\overset{(d)}{\leq} LH\left(\frac{1}{L}\sum_{i=1}^{L}(P_iU_P)\right) \\ &\overset{(e)}{=} LH(PU_P), \end{aligned}$$

where (d) follows by concavity $(\cap)$ of entropy and (e) follows by linearity and the fact that $\frac{1}{L}\sum_{i=1}^{L} P_i = P$. Therefore,

$$\begin{aligned} I(X^L; Y^L|Z = P) &= H(Y^L|Z = P) - H(Y^L|Z = P, X^L) \\ &\leq LH(PU_P) - LH(U_P|P) \\ &= LI(P, U_P) \\ &\leq L(R - \epsilon_L), \end{aligned}$$

where the last inequality follows by the definition of U_P in (A.57). Hence, for any distribution $P^{(L)}$ on X^L,

$$\begin{aligned} I(X^L; Y^L) &\leq I(X^L; Y^L|Z) + |\mathcal{X}| \log(L + 1) \\ &\leq L(R - \epsilon_L) + |\mathcal{X}| \log(L + 1) \\ &= LR \end{aligned}$$

because $L\epsilon_L = |\mathcal{X}| \log(L + 1)$. Therefore, $C(V) \leq LR$.

A.14 Lemmas for block codes

The lemmas in this appendix are used in results elsewhere in the thesis. They are lemmas on typicality as well as continuity of information measures.

Lemma 18. *Fix a block code of length n, either with or without feedback. Then,*

$$P_e(W) \geq P_e(V) \exp(-nd(V,W)),$$

where

$$d(V,W) \triangleq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{y^n \in \mathcal{D}_m^c} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)} \frac{1}{n} \sum_{i=1}^{n} \log \frac{V(y_i | x_i(m, y^{i-1}))}{W(y_i | x_i(m, y^{i-1}))}.$$

(a) If the code is a block code without feedback and there is a $P \in \mathcal{P}_n$ such that $\forall\ m \in \mathcal{M}$, $\phi(m) \in T_P$, then

$$d(V,W) \leq D(V||W|P) + \frac{\max\{\kappa_V, \kappa_W\}}{P_e(V)} \beta(n, |\mathcal{X}|, |\mathcal{Y}|),$$

where

$$\beta(n, |\mathcal{X}|, |\mathcal{Y}|) \triangleq \inf_{\epsilon > 0} \epsilon + (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n\frac{\epsilon^2}{2}\right).$$

(b) If the code is an arbitrary block code with feedback,

$$d(V,W) \leq \max_P D(V||W|P) + \frac{\max\{\kappa_V, \kappa_W\}}{P_e(V)} \beta(n, |\mathcal{X}|, |\mathcal{Y}|),$$

where the constant $\beta(n, |\mathcal{X}|, |\mathcal{Y}|)$ is defined above.

Note that by setting $\epsilon^2 = 2\frac{|\mathcal{X}||\mathcal{Y}|+1}{n} \log(n+1)$, we immediately have

$$\beta(n, |\mathcal{X}|, |\mathcal{Y}|) = O\left(\sqrt{\frac{\log(n)}{n}}\right).$$

Proof: Assume for now that the code has feedback. In the case that it does not (i.e. part (a) of the lemma), we will specialize the result along the way. Also assume that $D(V||W|P) < \infty$ in the non-feedback case and $\max_P D(V||W|P) < \infty$ in the case with feedback, otherwise there is nothing to prove.

From the definition of $P_e(W)$, we have

$$\begin{aligned}
P_e(W) &= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbb{P}_W(Y^n \in \mathcal{D}_m^c | M = m) \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P)} \sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)} \mathbb{P}_W(Y^n = y^n | M = m) \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P)} \sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)} \mathbb{P}_V(Y^n = y^n | M = m)\frac{\mathbb{P}_W(Y^n = y^n | M = m)}{\mathbb{P}_V(Y^n = y^n | M = m)} \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P)} \sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)} \mathbb{P}_V(Y^n = y^n | M = m)\prod_{i=1}^n \frac{W(y_i|x_i(m, y^{i-1}))}{V(y_i|x_i(m, y^{i-1}))} \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P)} \sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)} \mathbb{P}_V(Y^n = y^n | M = m)\prod_{x,y}\left[\frac{W(y|x)}{V(y|x)}\right]^{nP(x)U(y|x)},
\end{aligned}$$

where the last line follows from the definition of $B(m, P, U)$. Note that in the above, by the finite conditional divergence assumptions, if $\mathbb{P}_W(Y^n = y^n|M = m) = 0$, then $\mathbb{P}_V(Y^n = y^n|M = m) = 0$ also. Continuing, and multiplying and dividing by $P_e(V)$, we have

$$\begin{aligned}
\frac{P_e(W)}{P_e(V)} &\geq \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P)} \sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)}\prod_{x,y}\left[\frac{W(y|x)}{V(y|x)}\right]^{nP(x)U(y|x)} \\
&= \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P)} \sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)}\times \\
&\qquad \exp\left(-n\left[\sum_{x,y} P(x)U(y|x)\log\frac{V(y|x)}{W(y|x)}\right]\right) \\
&\overset{(A)}{\geq} \exp\Bigg(- n\Bigg[\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P)} \sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)}\times \\
&\qquad \sum_{x,y} P(x)U(y|x)\log\frac{V(y|x)}{W(y|x)}\Bigg]\Bigg) \\
&= \exp\Bigg(- n\Bigg[\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P)} \sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)}\frac{1}{n}\times \\
&\qquad \sum_{i=1}^n \log\frac{V(y_i|x_i(m, y^{i-1}))}{W(y_i|x_i(m, y^{i-1}))}\Bigg]\Bigg),
\end{aligned}$$

where (A) follows by Jensen's inequality applied to the function $f(t) = e^t$ and the last line follows by the definitive property of $y^n \in B(m, P, U)$. Fix an $\epsilon > 0$. We can focus our attention on $d(V, W)$ by splitting it into two terms,

$$\begin{aligned}
d(V,W) &= \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n, U \in \mathcal{V}_n(P)} \sum_{y^n \in \mathcal{D}_m^c \cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)} \times \\
&\qquad \sum_{x,y} P(x) U(y|x) \log \frac{V(y|x)}{W(y|x)} \\
&= T_1 + T_2 \\
T_1 &\triangleq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n, U \in \mathcal{V}_n(P) : (P,U) \in \mathcal{J}_\epsilon(V)} \sum_{y^n \in \mathcal{D}_m^c \cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)} \times \\
&\qquad \sum_{x,y} P(x) U(y|x) \log \frac{V(y|x)}{W(y|x)} \\
T_2 &\triangleq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n, U \in \mathcal{V}_n(P) : (P,U) \notin \mathcal{J}_\epsilon(V)} \sum_{y^n \in \mathcal{D}_m^c \cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)} \times \\
&\qquad \sum_{x,y} P(x) U(y|x) \log \frac{V(y|x)}{W(y|x)},
\end{aligned}$$

where

$$\mathcal{J}_\epsilon(V) = \left\{ (P,U) \in \mathcal{P} \times \mathcal{W} : \sum_{x,y} P(x) |U(y|x) - V(y|x)| \leq \epsilon \right\}.$$

For the $(P,U) \in \mathcal{J}_\epsilon(V)$ in T_1, Proposition 13 tells us that

$$\sum_{x,y} P(x) U(y|x) \log \frac{V(y|x)}{W(y|x)} \leq D(V||W|P) + \epsilon \max\{\kappa_V, \kappa_W\}.$$

For T_2, Lemma 19 tells us that for all $m \in \mathcal{M}$,

$$\sum_{P \in \mathcal{P}_n, U \in \mathcal{V}_n(P) : (P,U) \notin \mathcal{J}_\epsilon(V)} \mathbb{P}_V \left(Y^n \in B(m,P,U) | M = m \right) \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n \frac{\epsilon^2}{2} \right). \quad \text{(A.59)}$$

Now, if $W(y|x) = 0$, and $U(y|x) > 0$ for some y^n in the sum for T_2, then we are also guaranteed that $\mathbb{P}_V(Y^n = y^n | M = m) = 0$, so that y^n does not change the sum. Hence, for all y^n that have non-zero terms in the sum for T_2,

$$\sum_{x,y} P(x) U(y|x) \log \frac{V(y|x)}{W(y|x)} \leq \kappa_W.$$

So, we have

$$\begin{aligned}
T_1 \leq \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P):(P,U)\in\mathcal{J}_\epsilon(V)} \sum_{y^n\in\mathcal{D}_m^c\cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n=y^n|M=m)}{P_e(V)} \times \\
[D(V||W|P)+\epsilon\max\{\kappa_V,\kappa_W\}] \qquad \text{(A.60)}
\end{aligned}$$

$$\begin{aligned}
T_2 &\leq \frac{\kappa_W}{P_e(V)} \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P):(P,U)\notin\mathcal{J}_\epsilon(V)} \mathbb{P}_V(Y^n\in B(m,P,U)\cap\mathcal{D}_m^c|M=m) \\
&\leq \frac{\kappa_W}{P_e(V)} \frac{1}{|\mathcal{M}|} \sum_{m\in\mathcal{M}} \sum_{P\in\mathcal{P}_n, U\in\mathcal{V}_n(P):(P,U)\notin\mathcal{J}_\epsilon(V)} \mathbb{P}_V(Y^n\in B(m,P,U)|M=m) \\
&\leq \frac{\kappa_W}{P_e(V)}(n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-n\frac{\epsilon^2}{2}\right),
\end{aligned}$$

where the last line follows from averaging over the bound in (A.59), which holds uniformly for all $m\in\mathcal{M}$. Now, if the code does not have feedback and all codewords have type P as is assumed in part (a) of the lemma, then

$$\begin{aligned}
T_1 &\leq [D(V||W|P)+\epsilon\max\{\kappa_V,\kappa_W\}]\times \\
&\qquad \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \sum_{U\in\mathcal{V}_n(P):(P,U)\in\mathcal{J}_\epsilon(V)} \frac{\mathbb{P}_V(Y^n\in\mathcal{D}_m^c\cap B(m,P,U)|M=m)}{P_e(V)} \\
&\leq [D(V||W|P)+\epsilon\max\{\kappa_V,\kappa_W\}]\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\sum_{U\in\mathcal{V}_n(P)} \frac{\mathbb{P}_V(Y^n\in\mathcal{D}_m^c\cap B(m,P,U)|M=m)}{P_e(V)} \\
&= [D(V||W|P)+\epsilon\max\{\kappa_V,\kappa_W\}]\frac{1}{|\mathcal{M}|}\frac{\mathbb{P}_V(Y^n\in\mathcal{D}_m^c|M=m)}{P_e(V)} \\
&= [D(V||W|P)+\epsilon\max\{\kappa_V,\kappa_W\}].
\end{aligned}$$

Therefore, for part (a) of the lemma,

$$\begin{aligned}
d(V,W) &\leq T_1+T_2 \\
&\leq [D(V||W|P)+\epsilon\max\{\kappa_V,\kappa_W\}] + \frac{\kappa_W}{P_e(V)}(n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-n\frac{\epsilon^2}{2}\right) \\
&\leq D(V||W|P) + \frac{\max\{\kappa_V,\kappa_W\}}{P_e(V)}\left[\epsilon+(n+1)^{|\mathcal{X}||\mathcal{Y}|}\exp\left(-n\frac{\epsilon^2}{2}\right)\right].
\end{aligned}$$

Optimizing over $\epsilon>0$ gives the result of part (a) of the lemma. As for part (b), if the code does have feedback, we can continue bounding T_1 in (A.60) below to get

$$\begin{aligned}
T_1 &\leq \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n, U \in \mathcal{V}_n(P) : (P,U) \in \mathcal{J}_\epsilon(V)} \sum_{y^n \in \mathcal{D}_m^c \cap B(m,P,U)} \frac{\mathbb{P}_V(Y^n = y^n | M = m)}{P_e(V)} \times \\
&\qquad\qquad [D(V||W|P) + \epsilon \max\{\kappa_V, \kappa_W\}] \\
&\leq \left[\max_{P \in \mathcal{P}} D(V||W|P) + \epsilon \max\{\kappa_V, \kappa_W\} \right] \times \\
&\qquad \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n, U \in \mathcal{V}_n(P) : (P,U) \in \mathcal{J}_\epsilon(V)} \frac{\mathbb{P}_V(Y^n \in B(m,P,U) \cap \mathcal{D}_m^c | M = m)}{P_e(V)} \\
&\leq \left[\max_{P \in \mathcal{P}} D(V||W|P) + \epsilon \max\{\kappa_V, \kappa_W\} \right] \times \\
&\qquad \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{P \in \mathcal{P}_n, U \in \mathcal{V}_n(P)} \frac{\mathbb{P}_V(Y^n \in B(m,P,U) \cap \mathcal{D}_m^c | M = m)}{P_e(V)} \\
&= \left[\max_{P \in \mathcal{P}} D(V||W|P) + \epsilon \max\{\kappa_V, \kappa_W\} \right] \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{\mathbb{P}_V(Y^n \in \cap \mathcal{D}_m^c | M = m)}{P_e(V)} \\
&= \left[\max_{P \in \mathcal{P}} D(V||W|P) + \epsilon \max\{\kappa_V, \kappa_W\} \right].
\end{aligned}$$

Combining this with the bound on T_2, we have for part (b),

$$\begin{aligned}
d(V, W) &\leq T_1 + T_2 \\
&\leq \left[\max_{P \in \mathcal{P}} D(V||W|P) + \epsilon \max\{\kappa_V, \kappa_W\} \right] + \frac{\kappa_W}{P_e(V)} (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n \frac{\epsilon^2}{2} \right) \\
&\leq \max_{P \in \mathcal{P}} D(V||W|P) + \frac{\max\{\kappa_V, \kappa_W\}}{P_e(V)} \left[\epsilon + (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n \frac{\epsilon^2}{2} \right) \right].
\end{aligned}$$

Once again, optimizing over $\epsilon > 0$ gives the result of part (b).

Lemma 19. *Fix an $\epsilon > 0$ and a block code of length n either with or without feedback. For a given $m \in \mathcal{M}, V \in \mathcal{W}$, define the typical set under channel V,*

$$A_{m,\epsilon}(V) \triangleq \{ y^n : (P(m, y^n), V(m, y^n)) \in \mathcal{J}_\epsilon(V) \},$$

where

$$\mathcal{J}_\epsilon(V) \triangleq \left\{ (P, U) \in \mathcal{P} \times \mathcal{W} : \sum_{x \in \mathcal{X}} P(x) \sum_{y \in \mathcal{Y}} |U(y|x) - V(y|x)| \leq \epsilon \right\}.$$

Then,

$$\mathbb{P}_V(Y^n \notin A_{m,\epsilon}(V)|M=m) \leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n\frac{\epsilon^2}{2}\right).$$

Note that for a block code that does not have feedback, if $x^n(m) \in T_P$ for some type $P \in \mathcal{P}_n$,

$$A_{m,\epsilon}(V) = \left\{y^n : y^n \in T_U(x^n(m)), U \in \mathcal{V}_n(P), (P,U) \in \mathcal{J}_\epsilon(V)\right\}.$$

Proof: For now, assume that the code has feedback. Any code without feedback can be thought of as having feedback and not using it.

For a $(P,U) \notin \mathcal{J}_\epsilon(V)$,

$$\begin{aligned} D(U||V|P) &\overset{(a)}{=} \sum_x P(x) D(U(\cdot|x)||V(\cdot|x)) \\ &\overset{(b)}{\geq} \sum_x P(x) \frac{1}{2}\left(\sum_y |U(y|x) - V(y|x)|\right)^2 \\ &\overset{(c)}{\geq} \frac{1}{2}\left(\sum_x P(x) \sum_y |U(y|x) - V(y|x)|\right)^2 \\ &\overset{(d)}{\geq} \frac{\epsilon^2}{2}. \end{aligned}$$

In the above, (a) follows by the definition of conditional divergence, (b) follows by Pinsker's inequality [47], (c) from an application of Jensen's inequality to the function $f(t) = t^2$ and (d) follows from the assumption that $(P,U) \in \mathcal{J}_\epsilon(V)$.

Recall the definition of $B(m,P,U)$ for $P \in \mathcal{P}_n$, $U \in \mathcal{V}_n(P)$,

$$B(m,P,U) = \{y^n : P(m,y^n) = P, V(m,y^n) = U\}.$$

Using the fact that there are no more than $(n+1)^{|\mathcal{X}||\mathcal{Y}|}$ joint types (P,U), we have

$$\begin{aligned} \mathbb{P}_V(Y^n \notin A_{m,\epsilon}(V)|M=m) &= \sum_{(P,U)\notin\mathcal{J}_\epsilon(V)} \mathbb{P}_V(Y^n \in B(m,P,U)|M=m) \\ &\overset{(a)}{\leq} \sum_{(P,U)\notin\mathcal{J}_\epsilon(V)} \exp(-nD(V||W|P)) \\ &\leq (n+1)^{|\mathcal{X}||\mathcal{Y}|} \exp\left(-n\frac{\epsilon^2}{2}\right), \end{aligned}$$

where (a) follows by Proposition 14.

For a block code without feedback, it is clear that $y^n \in B(m,P,U)$ if and only if $y^n \in T_U(x^n(m))$, hence

$$A_{m,\epsilon}(V) = \left\{y^n : y^n \in T_U(x^n(m)), U \in \mathcal{V}_n(P), (P,U) \in \mathcal{J}_\epsilon(V)\right\}.$$

Proposition 13. *Let* $P \in \mathcal{P}, U, V, W \in \mathcal{W}$ *and recall that* $\kappa_V = \max_{x,y:V(y|x)>0} \log 1/V(y|x)$ *(similarly for* κ_W*). Assume that* $V(y|x) = U(y|x) = 0$ *whenever* $W(y|x) = 0$*. If for some* $\epsilon > 0$*,* $(P, U) \in \mathcal{J}_\epsilon(V)$*, that is,*

$$\sum_x P(x) \sum_y |U(y|x) - V(y|x)| \leq \epsilon,$$

then

$$\sum_x P(x) \sum_y U(y|x) \log \frac{V(y|x)}{W(y|x)} \leq D(V||W|P) + \epsilon \max\left\{\kappa_V, \kappa_W\right\}.$$

Proof: Forgive the excessively slow pace of this proof, but several times I thought that I had proved that the discrepancy from $D(V||W|P)$ was bounded by $\kappa_W \epsilon$, when that is not the case. First,

$$\begin{aligned}\sum_x P(x) \sum_y U(y|x) \log \frac{V(y|x)}{W(y|x)} &= \sum_x P(x) \sum_y V(y|x) \log \frac{V(y|x)}{W(y|x)} + \\ &\qquad \sum_x P(x) \sum_y (U(y|x) - V(y|x)) \log \frac{V(y|x)}{W(y|x)} \\ &= D(V||W|P) + \sum_x P(x) \sum_y (U(y|x) - V(y|x)) \log \frac{V(y|x)}{W(y|x)}.\end{aligned}$$

Now, splitting the 'extra' term above into different cases yields

$$\begin{aligned}
T &\triangleq \sum_x P(x) \sum_y (U(y|x) - V(y|x)) \log \frac{V(y|x)}{W(y|x)} \\
&= \sum_{x,y:U(y|x)>V(y|x)} P(x)(U(y|x) - V(y|x)) \log \frac{V(y|x)}{W(y|x)} + \\
&\qquad \sum_{x,y:U(y|x)<V(y|x)} P(x)(U(y|x) - V(y|x)) \log \frac{V(y|x)}{W(y|x)} \\
&\overset{(a)}{\leq} \sum_{x,y:U(y|x)>V(y|x)} P(x)(U(y|x) - V(y|x))\kappa_W + \\
&\qquad \sum_{x,y:U(y|x)<V(y|x)} P(x)(U(y|x) - V(y|x)) \log \frac{V(y|x)}{W(y|x)} \\
&= \sum_{x,y:U(y|x)>V(y|x)} P(x)(U(y|x) - V(y|x))\kappa_W + \\
&\qquad \sum_{x,y:U(y|x)<V(y|x)<W(y|x)} P(x)(U(y|x) - V(y|x)) \log \frac{V(y|x)}{W(y|x)} + \\
&\qquad \sum_{x,y:U(y|x)<V(y|x),W(y|x)<V(y|x)} P(x)(U(y|x) - V(y|x)) \log \frac{V(y|x)}{W(y|x)} \\
&= \sum_{x,y:U(y|x)>V(y|x)} P(x)(U(y|x) - V(y|x))\kappa_W + \\
&\qquad \sum_{x,y:U(y|x)<V(y|x)<W(y|x)} P(x)(V(y|x) - U(y|x)) \log \frac{W(y|x)}{V(y|x)} + \\
&\qquad \sum_{x,y:U(y|x)<V(y|x),W(y|x)<V(y|x)} P(x)(V(y|x) - U(y|x)) \log \frac{W(y|x)}{V(y|x)} \\
&\leq \sum_{x,y:U(y|x)>V(y|x)} P(x)(U(y|x) - V(y|x))\kappa_W + \\
&\qquad \sum_{x,y:U(y|x)<V(y|x)<W(y|x)} P(x)(V(y|x) - U(y|x))\kappa_V + \\
&\qquad \sum_{x,y:U(y|x)<V(y|x),W(y|x)<V(y|x)} P(x)(V(y|x) - U(y|x))0 \\
&\leq \sum_{x,y} P(x)|U(y|x) - V(y|x)| \max\{\kappa_W, \kappa_V\} \\
&\leq \epsilon \max\{\kappa_W, \kappa_V\},
\end{aligned}$$

where (a) follows because $V(y|x) \leq 1$ and $\log 1/W(y|x) \leq \kappa_W$ for the x, y where $(U(y|x) - V(y|x)) \neq 0$. Similar reasoning yields the other inequalities above.

Proposition 14. *For a fixed blocklength code with feedback, for each message m and $P \in \mathcal{P}_n$, $V \in \mathcal{V}_n(P)$, recall the 'conditional shell with feedback'*

$$B(m,P,V) \triangleq \{y^n : P(m,y^n) = P, V(m,y^n) = V\}.$$

Some properties of $B(m,P,V)$ are

(a) $B(m,P,V) \subset T_{PV} \subset \mathcal{Y}^n$

(b) If $y^n \in B(m,P,V)$,

$$\mathbb{P}_W(Y^n = y^n|M = m) = \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right)$$

(c) $|B(m,P,V)| \leq \exp(nH(V|P))$

(d) $\mathbb{P}_W(Y^n \in B(m,P,V)|M = m) \leq \exp(-nD(V||W|P))$.

Proof: Item (a) is true for the same reason that if $y^n \in T_V(P)$, $y^n \in T_{PV} \subset \mathcal{Y}^n$. As for (b), if $y^n \in B(m,P,V)$,

$$\begin{aligned}
\mathbb{P}_W(Y^n = y^n|M = m) &= \prod_{i=1}^{n} W(y_i|x_i(m,y^{i-1})) \\
&= \prod_{x,y} W(y|x)^{\sum_{i=1}^{n} 1(x_i(m,y^{i-1})=x, y_i=y)} \\
&= \exp\left(-n\left[\sum_{x,y} \log\frac{1}{W(y|x)}\frac{1}{n}\sum_{i=1}^{n} 1(x_i(m,y^{i-1}) = x, y_i = y)\right]\right) \\
&= \exp\left(-n\left[\sum_{x,y} \log\frac{1}{W(y|x)}P(x)V(y|x)\right]\right) \\
&= \exp\left(-n\left[\sum_{x,y} P(x)V(y|x)\left(\log\frac{V(y|x)}{W(y|x)} + \log\frac{1}{V(y|x)}\right)\right]\right) \\
&= \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right).
\end{aligned}$$

To prove (c), we can apply item (b) with the 'true channel' being set to V to get a bound as follows.

$$\begin{aligned}
1 &\geq \mathbb{P}_V(Y^n \in B(m,P,V)|M = m) \\
&= \sum_{y^n \in B(m,P,V)} \mathbb{P}_V(Y^n = y^n|M = m) \\
&= |B(m,P,V)| \exp(-n[D(V||V|P) + H(V|P)]) \\
&= |B(m,P,V)| \exp(-nH(V|P)).
\end{aligned}$$

Therefore, $|B(m,P,V)| \leq \exp(nH(V|P))$. Finally, we can combine (b) and (c) to get (d):

$$\begin{aligned}
\mathbb{P}_W(Y^n \in B(m,P,V)|M=m) &= \sum_{y^n \in B(m,P,V)} \mathbb{P}_W(Y^n = y^n|M=m) \\
&= \sum_{y^n \in B(m,P,V)} \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \\
&= |B(m,P,V)| \exp\left(-n\left[D(V||W|P) + H(V|P)\right]\right) \\
&\leq \exp\left(-nD(V||W|P)\right).
\end{aligned}$$

Proposition 15. *Suppose $P \in \mathcal{P}$ and $V, W \in \mathcal{W}$ with $\sum_x P(x) \sum_y |V(y|x) - W(y|x)| \leq \epsilon \leq 1/2$. Then,*

$$\begin{aligned}
&(a) \qquad\qquad |H(V|P) - H(W|P)| \leq \epsilon \log \frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} \\
&(b) \qquad\qquad |I(P,V) - I(P,W)| \leq 2\epsilon \log \frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon}.
\end{aligned}$$

Proof: Consider the distribution (P,V) on $\mathcal{X} \times \mathcal{Y}$ with $(P,V)(x,y) = P(x)V(y|x)$. Then,

$$\begin{aligned}
|H(V|P) - H(W|P)| &= |H(P) + H(V|P) - H(P) - H(W|P)| \\
&= |H(P,V) - H(P,W)| \\
&\overset{(a)}{\leq} \epsilon \log \frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon}.
\end{aligned}$$

In the above, (a) follows from Lemma 20 applied to joint distributions (P,V) and (P,W), while noting that

$$\sum_{x,y} |(P,V)(x,y) - (P,W)(x,y)| = \sum_x P(x) \sum_y |V(y|x) - W(y|x)| \leq \epsilon.$$

As for part (b) of the proposition, note that

$$\begin{aligned}
|I(P,V) - I(P,W)| &= |H(PV) - H(V|P) - H(PW) + H(W|P)| \\
&\leq |H(PV) - H(PW)| + |H(V|P) - H(W|P)| \\
&\leq \epsilon \log \frac{|\mathcal{Y}|}{\epsilon} + \epsilon \log \frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon} \\
&\leq 2\epsilon \log \frac{|\mathcal{X}||\mathcal{Y}|}{\epsilon},
\end{aligned}$$

where the bound on $|H(PV) - H(PW)|$ follows from Lemma 20 and noting that

$$\begin{aligned}\sum_y |(PV)(y) - (PW)(y)| &= \sum_y \left| \sum_x P(x)(V(y|x) - W(y|x)) \right| \\ &\leq \sum_x P(x) \sum_y |V(y|x) - W(y|x)| \\ &\leq \epsilon.\end{aligned}$$

Lemma 20 (Uniform continuity of entropy, [21], Theorem 16.3.2)**.** *Suppose P and P' are distributions on a finite alphabet $\mathcal{X}$ and $||P - P'||_1 = \sum_{x \in \mathcal{X}} |P(x) - P'(x)| \leq 1/2$, then*

$$|H(P) - H(P')| \leq ||P - P'||_1 \log \frac{|\mathcal{X}|}{||P - P'||_1}.$$

Lemma 21 (Pinsker's Inequality, [21], Lemma 12.6.1)**.** *Suppose P and P' are distributions on a finite alphabet $\mathcal{X}$. Then, with all logarithms and exponentials to the base e,*

$$D(P||P') \geq \frac{1}{2} \left(\sum_x |P(x) - P'(x)| \right)^2 .$$

Appendix B

Fixed delay coding appendix

B.1 Proof of Theorem 8

We will now rigorously prove Theorem 8 following the outline given in Section 3.6. The goal is to show that for a given rate $R > 0$,

$$\sup_{\mathcal{C} \in \mathcal{C}_R} E(\mathcal{C}) \leq \widetilde{E}(R) = \min_{V: C(V) \leq R} \max_{P: I(P,W) \geq R} D(V||W|P).$$

We will show that $E(\mathcal{C}) \leq \widetilde{E}(R - \epsilon)$ for all $\epsilon > 0$, and therefore, $E(\mathcal{C}) \leq \lim_{\epsilon \downarrow 0} \widetilde{E}(R - \epsilon)$, provided the limit exists. In Lemma 25, it is shown that $\widetilde{E}(R) = \lim_{\epsilon \downarrow 0} \widetilde{E}(R - \epsilon)$ for all $R > 0$, which will complete the proof.

B.1.1 Inducing error by forming a block code

So first, fix an $\epsilon \in (0, R)$, a $V \in \mathcal{W}$ such that $C(V) \leq R - \epsilon$ and a $\mathcal{C} \in \mathcal{C}_R$. Fix a $d, \widetilde{d} \geq 1$. The required properties of $d, \widetilde{d}$ will be described later in the proof. The first step is to construct a random block code of length $n \geq d$ as shown in Fig. 3.8. The message bits are

$$M = \left(B_1, \ldots, B_{\lfloor (n-d+1)R \rfloor}\right).$$

The encoder randomizes the code using the bits that arrive after time $n - d + 1$, that is the common randomness is

$$U = \left(B_{\lfloor (n-d+1)R \rfloor + 1}, \ldots, B_{\lfloor nR \rfloor}\right).$$

The encoder for this block code is the encoder for $\mathcal{C}$. The decoder for this block code is the delay-d decoder from $\mathcal{D}$ applied to all the bits in the message,

$$\widehat{M} = \left(\widehat{B}_1(d), \ldots, \widehat{B}_{\lfloor (n-d+1)R \rfloor}(d)\right).$$

The rate of the code is

$$\begin{aligned}\widetilde{R} &= \frac{\lfloor (n-d+1)R \rfloor}{n} \\ &\geq \frac{(n-d+1)R-1}{n} \\ &= R - \frac{(d-1)R+1}{n}.\end{aligned}$$

It follows that $\widetilde{R} \geq R - \epsilon/2$ if

$$n \geq \left\lceil \frac{2[(d-1)R+1]}{\epsilon} \right\rceil . \tag{B.1}$$

So for n large enough depending on d, R and ϵ, the rate of the block code is at least $R - \epsilon/2$.

Definition 12. *A rate-R* **decoder with feedforward information** $\overline{\mathcal{D}}$ *is set of decoders $\overline{\mathcal{D}} = \left\{\overline{\mathcal{D}}_{i,d}\right\}_{i\geq 1, d\geq 1}$ that have access to both all received channel outputs as well as all the past message bit realizations. So,*

$$\overline{\mathcal{D}}_{i,d} : \{0,1\}^{i-1} \times \mathcal{Y}^{\lceil i/R \rceil + d - 1} \to \{0,1\}.$$

Clearly, any regular decoder $\mathcal{D}$ is also a feedforward decoder since the feedforward decoder can choose to ignore the past message bits.

For any feedforward decoder, [37] shows that there is a feedforward decoder that uses only the d recent channel outputs since the bit arrived at the encoder as well as the past message bits, which performs at least as well as the original feedforward decoder for channel W.

Lemma 22. *[Lemma 4.1 of [37]] For a memoryless channel, a rate-R encoder $\mathcal{E}$ and a rate-R feedforward decoder $\overline{\mathcal{D}}$, there is another feedforward decoder $\mathcal{D}^f = \left\{\mathcal{D}^f_{i,d}\right\}_{i\geq 1, d\geq 0}$ using only the past d channel outputs and the past message bits that performs at least as well as $\overline{\mathcal{D}}$. That is,*

$$\begin{aligned}\mathcal{D}^f_{i,d} &: \{0,1\}^{i-1} \times \mathcal{Y}^d \to \{0,1\} \\ \widehat{B}^f_i(d) &= \mathcal{D}^f_{i,d}\left(B^{i-1}, Y_{\lceil i/R \rceil}^{\lceil i/R \rceil + d - 1}\right) \\ \mathbb{P}_W\left(B_i \neq \mathcal{D}^f_{i,d}\left(B^{i-1}, Y_{\lceil i/R \rceil}^{\lceil i/R \rceil + d - 1}\right)\right) &\leq \mathbb{P}_W\left(B_i \neq \overline{\mathcal{D}}_{i,d}\left(B^{i-1}, Y^{\lceil i/R \rceil + d - 1}\right)\right) \\ &\leq \mathbb{P}_W\left(B_i \neq \mathcal{D}_{i,d}\left(Y^{\lceil i/R \rceil + d - 1}\right)\right).\end{aligned}$$

The intuition behind this lemma is that if the channel is memoryless, the past message bits provide us with more information about the inputs coming after bit i has entered the encoder than the past channel outputs.

Lemma 22 tells us that if we can prove a lower bound to the error probability for the class of feedforward decoders that use only recent channel outputs, then we can also lower bound the error probability to our actual decoder. The upshot is that a change-of-measure argument applied to these feedforward decoders using recent channel outputs only requires the 'true' channel W to act like a 'test' channel V for d time units, as opposed to the full n symbols of the block code.

Define $\widetilde{B}_i^f(d) = \widetilde{B}_i^f(d) \oplus B_i$ to be the error sequence for decoder $\mathcal{D}^f$. Now, Lemmas 4.2 and 4.3 of [37] show that, provided n is large enough as in Eqn. (B.1),

$$H\left(\widetilde{B}^{f,\lfloor(n-d+1)R\rfloor}(d)\right) \geq n(R-\epsilon/2) - I(X^n;Y^n),$$

where $I(X^n;Y^n)$ denotes the mutual information between the random variables X^n and Y^n (e.g. as in the textbook of Cover and Thomas [21]). It is straightforward to check (it is a part of the converse to the channel coding theorem in [21]) that under a channel V with $C(V) \leq R - \epsilon$,

$$\begin{aligned} I(X^n;Y^n) &\leq nC(V) \leq n(R-\epsilon) \\ H\left(\widetilde{B}^{f,\lfloor(n-d+1)R\rfloor}(d)\right) &\geq n\epsilon/2. \end{aligned}$$

The sum of the marginal entropies is at least the entropy of all the error bits, so

$$\begin{aligned} \sum_{i=1}^{\lfloor(n-d+1)R\rfloor} H\left(\widetilde{B}_i^f(d)\right) &\geq n\epsilon/2 \\ \frac{1}{nR}\sum_{i=1}^{\lfloor(n-d+1)R\rfloor} H\left(\widetilde{B}_i^f(d)\right) &\geq \epsilon/2R. \end{aligned}$$

Since entropy is non-negative and $\lfloor(n-d+1)R\rfloor \leq nR$, it follows that there is at least one i such that $H(\widetilde{B}_i^f(d)) \geq \epsilon/2$. Then, for this i, it follows from the monotonicity of the binary entropy function $h_b(t) = -t\log t - (1-t)\log(1-t), t \in [0,1]$, that

$$\mathbb{P}_V\left(B_i \neq \mathcal{D}_{i,d}^f\left(B^{i-1}, Y_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1}\right)\right) \geq h_b^{-1}\left(\frac{\epsilon}{2R}\right). \tag{B.2}$$

For each $b^i \in \{0,1\}^i$, denote the error set for $\mathcal{D}_{i,d}^f$ to be

$$A(b^i) \triangleq \left\{y_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1} \in \mathcal{Y}^d : \mathcal{D}_{i,d}^f\left(b^{i-1}, y_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1}\right) \neq b_i\right\}.$$

Combining this definition of the error set with (B.2) gives

$$\begin{aligned} h_b^{-1}\left(\frac{\epsilon}{2R}\right) \leq \sum_{b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}\in\{0,1\}^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}} &\frac{1}{2^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}}\times \\ &\mathbb{P}_V\left(A(b^i)\middle|B^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor} = b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}\right). \end{aligned} \tag{B.3}$$

Define the set G as

$$G \triangleq \left\{ b^{\lfloor (\lceil i/R \rceil + d - 1)R \rfloor} : \mathbb{P}_V \left(A(b^i) \middle| B^{\lfloor (\lceil i/R \rceil + d - 1)R \rfloor} = b^{\lfloor (\lceil i/R \rceil + d - 1)R \rfloor} \right) \geq h_b^{-1} \left(\frac{\epsilon}{2R} \right) \right\}.$$

It is easily verified that

$$\mathbb{P}(G) \geq h_b^{-1} \left(\frac{\epsilon}{2R} \right)$$

as assuming the contraposition leads to a contradiction of (B.3). At this point, we diverge from the proof in [37]. We will now use the assumption that $\mathcal{C}$ is a good anytime code to show that the type of the input over the d time steps of interest must be able to support a rate close to R over the channel W.

B.1.2 The second block code construction

We wish to study the type of the input during the d time steps after bit i arrives at the encoder. In order to do so, we will construct a second randomized block code as shown in Fig. 3.9. The length of the block code is d and the message bits are

$$M = \left(B_i, \dots, B_{\left\lfloor (\lceil i/R \rceil + d - \widetilde{d})R \right\rfloor} \right).$$

The rate of the block code is therefore

$$\begin{aligned} \frac{\left\lfloor (\lceil i/R \rceil + d - \widetilde{d})R \right\rfloor - i + 1}{d} &\geq \frac{(\lceil i/R \rceil + d - \widetilde{d})R - i}{d} \\ &\geq R - \frac{\widetilde{d}}{d} R. \end{aligned}$$

Fix a $\gamma > 0$. The rate of this block code is at least $R - \gamma$ provided that $\widetilde{d} \leq \lfloor \gamma d / R \rfloor$. The randomness used at the encoder is

$$U = \left(B_1, \dots, B_{i-1}, B_{\left\lfloor (\lceil i/R \rceil + d - \widetilde{d})R \right\rfloor + 1}, \dots, B_{\lfloor (\lceil i/R \rceil + d - 1)R \rfloor} \right),$$

that is, both the bits that come before i and the ones that arrive at the encoder after time $\lceil i/R \rceil + d - \widetilde{d}$. The decoder used will be the feedforward delay-$\widetilde{d}$ decoder that uses past message bits as well as recent channel outputs $\mathcal{D}^f$ from Lemma 22. So, the decoded message is

$$\widehat{M} = \left(\widehat{B}_i^f(\widetilde{d}), \dots, \widetilde{B}^f_{\left\lfloor (\lceil i/R \rceil + d - \widetilde{d})R \right\rfloor}(\widetilde{d}) \right).$$

Since we have used the feedforward decoder, the performance of this block code, averaged over the randomness in U is bounded by

$$\begin{aligned}
\mathbb{P}_W(\widehat{M} \neq M) &\overset{(A)}{\leq} \sum_{i'=i}^{\lfloor(\lceil i/R\rceil + d - \widetilde{d})R\rfloor} \mathbb{P}_W\left(\widehat{B}^f_{i'}(\widetilde{d}) \neq B_{i'}\right) \\
&\leq dR \sup_{i'} \mathbb{P}_W\left(\widehat{B}^f_{i'}(\widetilde{d}) \neq B_{i'}\right) \\
&\leq dR \sup_{i'} \mathbb{P}_W\left(\widehat{B}_{i'}(\widetilde{d}) \neq B_{i'}\right),
\end{aligned}$$

where (A) follows by union bound. By assumption, $E(\mathcal{C}) \geq 2\widetilde{E}(R)/3$, otherwise there is nothing to prove for this $\mathcal{C}$. Therefore, there is some d_0 such that if $\widetilde{d} \geq d_0(\mathcal{C})$,

$$\sup_{i'} \mathbb{P}_W\left(\widehat{B}_{i'}(\widetilde{d}) \leq B_{i'}\right) \leq \exp(-\widetilde{d}\widetilde{E}(R)/2).$$

So for $\widetilde{d} \geq d_0$,

$$\mathbb{P}_W(M \neq \widehat{M}) \leq dR\exp(-\widetilde{d}\widetilde{E}(R)/2) = \exp\left(-d\left[\frac{\widetilde{d}}{2d}\widetilde{E}(R) - \frac{1}{d}\log dR\right]\right),$$

and for $d \geq d_1(\widetilde{d}/d)$, $\log(dR)/d \leq \widetilde{E}(R)\widetilde{d}/4d$, so

$$\mathbb{P}_W(M \neq \widehat{M}) \leq dR\exp(-\widetilde{d}\widetilde{E}(R)/2) = \exp\left(-d\left[\frac{\widetilde{d}}{4d}\widetilde{E}(R)\right]\right).$$

Letting $\mathbb{P}_W(\widehat{M} \neq M|U)$ denote the error probability conditioned on a value of the common randomness, we have by Markov's inequality that

$$\mathbb{P}\left(\mathbb{P}_W(\widehat{M} \neq M|U) > \exp\left(d\left[\frac{\widetilde{d}}{8d}\widetilde{E}(R)\right]\right) \times \exp\left(-d\left[\frac{\widetilde{d}}{4d}\widetilde{E}(R)\right]\right)\right) \leq \exp\left(-d\left[\frac{\widetilde{d}}{8d}\widetilde{E}(R)\right]\right).$$

So for $\left(1 - \exp\left(-d\left[\frac{\widetilde{d}}{8d}\widetilde{E}(R)\right]\right)\right)$ fraction of the realizations of U, the error probability of the deterministic block code for that value of U is at most $\exp\left(-d\left[\frac{\widetilde{d}}{8d}\widetilde{E}(R)\right]\right)$ provided that $d \geq \max(d_1(\widetilde{d}/d), d_0(\mathcal{C}))$. Let $\mathcal{U}$ be the set of u for which the bound

$$\mathbb{P}_W(\widehat{M} \neq M|U = u) \leq \exp\left(-d\left[\frac{\widetilde{d}}{8d}\widetilde{E}(R)\right]\right)$$

holds. For these realizations of the block code, we will apply Lemma 23 with the tuple (n,δ,α,R) in Lemma 23 being the parameters $(d,\gamma,\widetilde{d}\widetilde{E}(R)/8d,R-\gamma)$ here. The lemma then says that if d is large enough, depending on $\gamma,\widetilde{d}\widetilde{E}(R)/8d,|\mathcal{X}|$ and $|\mathcal{Y}|$, then

$$\mathbb{P}\left(\text{Type of } X_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1}=P,I(P,W)\leq R-2\gamma\middle|U\in\mathcal{U}\right)\leq\exp\left(-d\left[\frac{\gamma}{4}-\frac{|\mathcal{X}|}{d}\log(d+1)\right]\right).$$

Therefore, unconditioned on the common randomness,

$$\begin{aligned}\mathbb{P}&\left(\text{Type of } X_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1}=P,I(P,W)>R-2\gamma\right)\\&\geq\mathbb{P}\left(\text{Type of } X_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1}=P,I(P,W)>R-2\gamma,U\in\mathcal{U}\right)\\&\geq\mathbb{P}\left(\text{Type of } X_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1}=P,I(P,W)>R-2\gamma|U\in\mathcal{U}\right)\times\\&\qquad\left(1-\exp\left(-d\left[\frac{\widetilde{d}}{8d}\widetilde{E}(R)\right]\right)\right)\\&\geq\left(1-\exp\left(-d\left[\frac{\gamma}{4}-\frac{|\mathcal{X}|}{d}\log(d+1)\right]\right)\right)\times\\&\qquad\left(1-\exp\left(-d\left[\frac{\widetilde{d}}{8d}\widetilde{E}(R)\right]\right)\right),\end{aligned}$$

for d and $\widetilde{d}$ large enough. Define the set

$$G_\gamma\triangleq\left\{b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}:\text{Type of } x_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1}=P,I(P,W)>R-2\gamma\right\}.$$

So, we have that

$$\mathbb{P}(G_\gamma)\geq\left(1-\exp\left(-d\left[\frac{\gamma}{4}-\frac{|\mathcal{X}|}{d}\log(d+1)\right]\right)\right)\left(1-\exp\left(-d\left[\frac{\widetilde{d}}{8d}\widetilde{E}(R)\right]\right)\right)$$

for large enough d and $\widetilde{d}$. Recall the set

$$G\triangleq\left\{b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}:\mathbb{P}_V\left(A(b^i)\middle|B^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}=b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}\right)\geq h_b^{-1}\left(\frac{\epsilon}{2R}\right)\right\}.$$

and that

$$\mathbb{P}(G)\geq h_b^{-1}\left(\frac{\epsilon}{2R}\right).$$

Now,

$$
\begin{aligned}
\mathbb{P}(G \cap G_\gamma) &= 1 - \mathbb{P}(G^c \cup G_\gamma^c) \\
&\geq 1 - \mathbb{P}(G^c) - \mathbb{P}(G_\gamma^c) \\
&\geq 1 - \left(1 - h_b^{-1}\left(\frac{\epsilon}{2R}\right)\right) - \\
&\quad \left(1 - \left(1 - \exp\left(-d\left[\frac{\gamma}{4} - \frac{|\mathcal{X}|}{d}\log(d+1)\right]\right)\right)\left(1 - \exp\left(-d\left[\frac{\widetilde{d}}{8d}\widetilde{E}(R)\right]\right)\right)\right) \\
&\geq \frac{1}{2}h_b^{-1}\left(\frac{\epsilon}{2R}\right), \qquad \text{(B.4)}
\end{aligned}
$$

where the last line holds for all d larger than some finite $d_2(\epsilon/2R, \mathcal{C}, \gamma, d/\widetilde{d}, \widetilde{d}, \widetilde{E}(R), |\mathcal{X}|, |\mathcal{Y}|)$ since the exponents are all positive in the second to last line above. When we take the limit of d going to infinity, we will particularly need $\widetilde{d}/d$ to not decay too fast (i.e. keep it is a small constant, take the limit as $d \to \infty$, and repeat the entire argument for a smaller ratio $\widetilde{d}/d$). So, we have succeeded in showing that the type supports a rate near R for large enough d. At this point, we can finish the proof with a change of measure argument.

B.1.3 Finishing with a change of measure argument

Lemma 24 shows that if an event has a non-vanishing error probability under channel V, and the type of the input is known to have high mutual information, the probability of the event under W can be lower bounded with a divergence term that is constrained by the input distribution. In particular, if $b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor} \in G \cap G_\gamma$, this means that

$$
\begin{aligned}
\mathbb{P}_W\left(\widehat{B}_i^f(d) \neq B_i \middle| b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}\right) &= \mathbb{P}_W\left(Y_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1} \in A(b^i) \middle| b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}\right) \\
&\geq \frac{1}{2}h_b^{-1}\left(\frac{\epsilon}{2R}\right) \times \exp\left(-d\left[\max_{P:I(P,W)\geq R-2\gamma} D(V||W|P) + \right.\right. \\
&\qquad \left.\left.\sqrt{\frac{4|\mathcal{X}||\mathcal{Y}|}{d}\log(d+1)}\max\{\kappa_V, \kappa_W\}\right]\right),
\end{aligned}
$$

provided that d is larger than some finite number depending on $h_b^{-1}(\epsilon/2R), |\mathcal{X}|$ and $|\mathcal{Y}|$. Using the above, we have

$$
\begin{aligned}
\zeta &\triangleq \mathbb{P}_W(\widehat{B}_i(d) \neq B_i) \\
&\geq \mathbb{P}_W(\widehat{B}_i^f(d) \neq B_i) \\
&= \sum_{b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}} \mathbb{P}\left(b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}\right) \mathbb{P}_W\left(Y_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1} \in A(b^i) \middle| b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}\right) \\
&\geq \sum_{b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor} \in G\cap G_\gamma} \mathbb{P}\left(b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}\right) \mathbb{P}_W\left(Y_{\lceil i/R\rceil}^{\lceil i/R\rceil+d-1} \in A(b^i) \middle| b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}\right) \\
&\geq \frac{1}{2} h_b^{-1}\left(\frac{\epsilon}{2R}\right) \sum_{b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor} \in G\cap G_\gamma} \mathbb{P}\left(b^{\lfloor(\lceil i/R\rceil+d-1)R\rfloor}\right) \times \\
&\quad \exp\left(-d\left[\max_{P:I(P,W)\geq R-2\gamma} D(V||W|P) + \sqrt{\frac{4|\mathcal{X}||\mathcal{Y}|}{d}\log(d+1)\max\{\kappa_V,\kappa_W\}}\right]\right) \\
&\geq \frac{1}{4} h_b^{-1}\left(\frac{\epsilon}{2R}\right)^2 \exp\left(-d\left[\max_{P:I(P,W)\geq R-2\gamma} D(V||W|P) + \right.\right. \\
&\qquad\qquad \left.\left.\sqrt{\frac{4|\mathcal{X}||\mathcal{Y}|}{d}\log(d+1)\max\{\kappa_V,\kappa_W\}}\right]\right).
\end{aligned}
$$

Now, if we let $\gamma \leq \min\{\epsilon/2$, and let d and $\widetilde{d}$ tend to ∞ while the ratio $\widetilde{d}/d$ is at most γ (while not being much smaller, up to integer effects), we can take log and the limit as $d \to \infty$ to get

$$\limsup_{d\to\infty} -\frac{1}{d}\log P_e(d,\mathcal{C}) \leq \max_{P:I(P,W)\geq R-\epsilon} D(V||W|P).$$

If V is an optimizing channel for $\widetilde{E}(R-\epsilon)$, this shows that $E(\mathcal{C}) \leq \widetilde{E}(R-\epsilon)$. Lemma 25(c) shows that $\lim_{\epsilon\downarrow 0}\widetilde{E}(R-\epsilon) = \widetilde{E}(R)$, so it follows that $E(\mathcal{C}) \leq \widetilde{E}(R)$, so the theorem is proved.

B.1.4 Input types for good block codes

In this section, we show that if a sequence of block codes has an exponentially decaying probability of error, most of the codewords in the codes eventually have types that 'support enough rate' across the channel.

Lemma 23. *Suppose we have a sequence of deterministic block codes of length n going to ∞ and rate at least R for each n. Suppose that there is an $\alpha > 0$ so that for all n greater* *than some finite $\widetilde{n}(\alpha)$,*

$$P_e^{(n)}(W) \leq \exp(-n\alpha),$$

where $P_e^{(n)}(W)$ denotes the error probability of the length n code over channel W. Fix a $\delta \in (0, 2\alpha)$. There exists some finite $n'(\delta, \alpha, |\mathcal{X}|, |\mathcal{Y}|)$ such that for all $n \geq n'(\delta, \alpha, |\mathcal{X}|, |\mathcal{Y}|)$,

$$\frac{\left|\bigcup_{P \in \mathcal{P}_n : I(P,W) \leq R - \delta} \{m \in \mathcal{M}_n : \phi_n(m) \in T_P\}\right|}{|\mathcal{M}_n|} \leq \exp\left(-n\left[\frac{\delta}{4} - \frac{|\mathcal{X}|}{n}\log(n+1)\right]\right),$$

where $\phi_n(m)$ is the codeword for message m in the length n code, $\mathcal{M}_n$ is the message set for the length n code and $\mathcal{P}_n$ is the set of types of length n for $\mathcal{X}$. Hence, the probability that the type of the input supports rate less than $R - \delta$ is decaying exponentially (albeit with a small exponent).

Proof: Fix an n in the sequence and the block code for length n and drop the subscript n from $\mathcal{M}_n$. For each $P \in \mathcal{P}_n$, let $\mathcal{M}(P) = \{m \in \mathcal{M} : \phi(m) \in T_P\}$, where T_P is the type class for type P [47]. Define

$$C_P = -\frac{1}{n}\log\frac{|\mathcal{M}(P)|}{|\mathcal{M}|}.$$

Consider the subcode one gets from only taking messages in $\mathcal{M}(P)$ and using the maximum likelihood (ML) decoder for the thinned out subcode. Let $P_e(W; P)$ denote the (average) error probability of this subcode. The rate of this subcode is at least $R_P = R - C_P$. Let $P_e(W; P)$ denote the error probability of this subcode with ML decoding over channel W, and let $P_c(W; P) = 1 - P_e(W; P)$. For any $\tau \in (0, 1/2)$, we can apply Lemma 12 to get

$$\begin{aligned} P_c(W;P) \leq \exp&\left(-n\left[\frac{\tau^2}{2} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) + \\ &\exp\left(-n\left[R_P - I(P,W) - 2\tau\log\frac{|\mathcal{X}||\mathcal{Y}|}{\tau} - \frac{|\mathcal{Y}|}{n}\log(n+1)\right]\right). \end{aligned} \tag{B.5}$$

Now, if we let $\mathcal{D}_m$ denote the decoding set for each message in the code, we see that

$$\begin{aligned} P_e(W) &= \frac{1}{|\mathcal{M}|}\sum_{P \in \mathcal{P}_n}\sum_{m \in \mathcal{M}(P)} \mathbb{P}_W\left(Y^n \notin \mathcal{D}_m | X^n = \phi(m)\right) \\ &\geq \frac{|\mathcal{M}(P)|}{|\mathcal{M}|}\frac{1}{|\mathcal{M}(P)|}\sum_{m \in \mathcal{M}} \mathbb{P}_W\left(Y^n \notin \mathcal{D}_m | X^n = \phi(m)\right) \\ &\geq \exp(-nC_P)P_e(W;P) \end{aligned} \tag{B.6}$$

for any $P \in \mathcal{P}_n$. Now, fix a $\delta > 0$. Suppose we focus on a $P \in \mathcal{P}_n$ such that $I(P,W) \leq R - \delta$. Then, from (B.5), we can say that

$$P_c(W;P) \leq \exp\left(-n\left[\frac{\tau^2}{2} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) + \exp\left(-n\left[\delta - C_P - 2\tau\log\frac{|\mathcal{X}||\mathcal{Y}|}{\tau} - \frac{|\mathcal{Y}|}{n}\log(n+1)\right]\right).$$

Now, there exists a finite $n_1(\delta, |\mathcal{Y}|)$ such that if $n \geq n_1$, we also have,

$$P_c(W;P) \leq \exp\left(-n\left[\frac{\tau^2}{2} - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) + \exp\left(-n\left[3\delta/4 - C_P - 2\tau\log\frac{|\mathcal{X}||\mathcal{Y}|}{\tau}\right]\right).$$

If we let

$$g(\delta, |\mathcal{X}|, |\mathcal{Y}|) \triangleq \sup_{\tau \in (0,1/2)} \left\{\frac{\tau^2}{2} : 2\tau\log\frac{|\mathcal{X}||\mathcal{Y}|}{\tau} \leq \frac{\delta}{4}\right\},$$

noting that $g(\delta, |\mathcal{X}|, |\mathcal{Y}|) > 0$ for all $\delta > 0$, we also have

$$P_c(W;P) \leq \exp\left(-n\left[g(\delta, |\mathcal{X}|, |\mathcal{Y}|) - \frac{|\mathcal{X}||\mathcal{Y}|}{n}\log(n+1)\right]\right) + \exp\left(-n\left[\delta/2 - C_P\right]\right)$$

by the appropriate choice of τ. Further, there exists a finite $n_2(\delta, |\mathcal{X}|, |\mathcal{Y}|)$ (WLOG, $n_2 \geq n_1$), such that if $n \geq n_2$, we also have

$$P_c(W;P) \leq \exp\left(-ng(\delta, |\mathcal{X}|, |\mathcal{Y}|)/2\right) + \exp\left(-n\left[\delta/2 - C_P\right]\right).$$

Therefore, for any $P \in \mathcal{P}_n$ with $I(P,W) \geq R - \delta$ and $n \geq n_2(\delta, |\mathcal{X}|, |\mathcal{Y}|)$,

$$P_e(W) \geq \exp(-nC_P)\left[1 - \exp(-ng(\delta, |\mathcal{X}|, |\mathcal{Y}|)/2) - \exp(-n(\delta/2 - C_P))\right],$$

where $g(\delta, |\mathcal{X}|, |\mathcal{Y}|) > 0$ for all $\delta > 0$. Now, assume that for all $n \geq n_0(\alpha)$, $P_e(W) \leq \exp(-n\alpha)$, where $\alpha > 0$. So for $n \geq \max(n_0(\alpha), n_2(\delta, |\mathcal{X}|, |\mathcal{Y}|))$, $P \in \mathcal{P}_n$ with $I(P,W) \leq R - \delta$,

$$\exp(-n\alpha) \geq P_e(W) \geq \exp(-nC_P)\left[1 - \exp(-ng(\delta, |\mathcal{X}|, |\mathcal{Y}|)/2) - \exp(-n(\delta/2 - C_P))\right].$$

Now, further suppose that $C_P \leq \delta/4$, so we would have

$$\begin{aligned}\exp(-n\alpha) &\geq \exp(-n\delta/4)\left[1 - \exp(-ng(\delta, |\mathcal{X}|, |\mathcal{Y}|)/2) - \exp(-n\delta/4)\right]\\ \exp(-n\alpha/2) &\geq \left[1 - \exp(-ng(\delta, |\mathcal{X}|, |\mathcal{Y}|)/2) - \exp(-n\delta/4)\right],\end{aligned}$$

where the last line holds since $\delta < 2\alpha$. Now the LHS above is decaying exponentially in n, while the RHS is converging to 1, this is a contradiction for n greater than some finite $n_3(\delta, \alpha, |\mathcal{X}|, |\mathcal{Y}|) \geq \max(n_0(\alpha), n_2(\delta, |\mathcal{X}|, |\mathcal{Y}|))$. Therefore, for $n \geq n_3$, there is no $P \in \mathcal{P}_n$ for which $I(P,W) \leq R - \delta$ and $C_P \leq \delta/4$. The number of types of length n for $\mathcal{X}$ is at most $(n+1)^{|\mathcal{X}|}$, so we have for $n \geq n_3$,

$$\sum_{P:I(P,W)\leq R-\delta} |\mathcal{M}(P)| \leq (n+1)^{|\mathcal{X}|} \exp(-n(R - \delta/4)),$$

which completes the proof of the lemma, with $n' = n_3$.

B.1.5 Another change of measure lemma

Lemma 24. *Let A be a set of vectors in $\mathcal{Y}^d$ and let $x^d \in T_P \subset \mathcal{X}^d$ for some type $P \in \mathcal{P}_d$ with $I(P,W) \geq R - 2\gamma$. Suppose, that for some $\beta > 0$,*

$$\mathbb{P}_V\left(Y^d \in A | X^d = x^d\right) \geq \beta.$$

Then,

$$\mathbb{P}_W\left(Y^d \in A | X^d = x^d\right) \geq \frac{\beta}{2} \exp\left(-d\left[\max_{P:I(P,W)\geq R-2\gamma} D(V||W|P) + \sqrt{\frac{4|\mathcal{X}||\mathcal{Y}|}{d}\log(d+1)\max\{\kappa_V, \kappa_W\}}\right]\right),$$

for all d at least some finite $d_4(\beta, |\mathcal{X}|, |\mathcal{Y}|)$, where $\kappa_V = \max_{x,y:V(y|x)>0} - \log V(y|x)$ and similarly for κ_W.

Proof: For $\theta > 0$, define a typical set under channel V,

$$T_\theta = \left\{ y^d : y^d \in T_U(x^d), U \in \mathcal{V}_d(P), \sum_{x,y} P(x)|V(y|x) - U(y|x)| \leq \theta \right\}.$$

Then, as shown in Lemma 19,

$$\mathbb{P}_V(Y^d \notin T_\theta | X^d = x^d) \leq (d+1)^{|\mathcal{X}||\mathcal{Y}|} \exp(-d\theta^2/(2)).$$

Now,

$$\begin{aligned}\mathbb{P}_V\left(Y^d \in A \cap T_\theta | X^d = x^d\right) &= 1 - \mathbb{P}_V(Y^d \in A^c \cup T_\theta^c | X^d = x^d) \\ &\geq 1 - (1-\beta) - (d+1)^{|\mathcal{X}||\mathcal{Y}|}\exp(-d\theta^2/(2)) \\ &= \beta - (d+1)^{|\mathcal{X}||\mathcal{Y}|}\exp(-d\theta^2/(2)). \qquad \text{(B.7)}\end{aligned}$$

Now, for $y^d \in T_\theta$,

$$\frac{\mathbb{P}_W(Y^d = y^d|X^d = x^d)}{\mathbb{P}_V(Y^d = y^d|X^d = x^d)} = \prod_{l=1}^{d} \frac{W(y_l|x_l)}{V(y_l|x_l)}$$
$$= \exp\left(-d\left[\sum_{x,y} P(x)U(y|x)\log\frac{V(y|x)}{W(y|x)}\right]\right),$$

where $U \in \mathcal{V}_d(P)$ is some conditional type such that $\sum_{x,y} P(x)|V(y|x) - U(y|x)| \leq \theta, y^d \in T_U(x^d)$. Hence,

$$\sum_{x,y} P(x)U(y|x)\log\frac{V(y|x)}{W(y|x)} \leq \sum_{x,y} P(x)V(y|x)\log\frac{V(y|x)}{W(y|x)} +$$
$$\sum_{x,y} P(x)|U(y|x) - V(y|x)| \max\{\kappa_V, \kappa_W\}$$
$$\leq D(V||W|P) + \theta \max\{\kappa_V, \kappa_W\}.$$

Therefore,

$$\begin{aligned}
\zeta &\triangleq \mathbb{P}_W(Y^d \in A|X^d = x^d)\\
&\geq \mathbb{P}_W(Y^d \in A \cap T_\theta|X^d = x^d)\\
&= \sum_{y^d \in A \cap T_\theta} \mathbb{P}_W(Y^d = y^d|X^d = x^d)\\
&= \sum_{y^d \in A \cap T_\theta} \mathbb{P}_V(Y^d = y^d|X^d = x^d)\frac{\mathbb{P}_W(Y^d = y^d|X^d = x^d)}{\mathbb{P}_V(Y^d = y^d|X^d = x^d)}\\
&\geq \sum_{y^d \in A \cap T_\theta} \mathbb{P}_V(Y^d = y^d|X^d = x^d)\exp\left(-d\left[D(V||W|P) + \theta \max\{\kappa_V, \kappa_W\}\right]\right)\\
&= \exp\left(-d\left[D(V||W|P) + \theta \max\{\kappa_V, \kappa_W\}\right]\right)\mathbb{P}_V(Y^d \in A \cap T_\theta|X^d = x^d)\\
&\geq \exp\left(-d\left[D(V||W|P) + \theta \max\{\kappa_V, \kappa_W\}\right]\right)\left(\beta - (d+1)^{|\mathcal{X}||\mathcal{Y}|}\exp(-d\theta^2/2)\right),
\end{aligned}$$

where the last line follows from (B.7). By setting

$$\theta = \sqrt{\frac{4|\mathcal{X}||\mathcal{Y}|}{d}\log(d+1)},$$

we get

$$\mathbb{P}_W(Y^d \in A|X^d = x^d) \geq \left(-d\left[D(V||W|P) + \sqrt{\frac{4|\mathcal{X}||\mathcal{Y}|}{d}\log(d+1)}\max\{\kappa_V, \kappa_W\}\right]\right) \times$$
$$\left(\beta - (d+1)^{-|\mathcal{X}||\mathcal{Y}|}\right)$$

and hence the lemma holds for d large enough depending on $\beta, |\mathcal{X}|$ and $|\mathcal{Y}|$, since we can trivially take a max over all P such that $I(P, W) \geq R - 2\gamma$.

B.1.6 Left continuity of $\widetilde{E}(R)$

Lemma 25. *Fix a W and $R \geq 0$. Recall that*

$$\widetilde{E}(R) \triangleq \min_{V:C(V)\leq R} \max_{P:I(P,W)\geq R} D(V||W|P).$$

For $R \geq C(W)$, let $\widetilde{E}(R) = 0$ by convention[1]*. Then,*

(a) $\widetilde{E}(R)$ is a monotone nonincreasing function of R.

(b) For all $R \geq 0$,$E_{sp}(R) \leq \widetilde{E}(R) \leq E_h(R)$.

(c) For all $R > 0$, $\widetilde{E}(R)$ is left continuous, that is $\widetilde{E}(R) = \lim_{\epsilon\downarrow 0} \widetilde{E}(R-\epsilon)$.

Proof: Let $\mathcal{V}_R = \{V \in \mathcal{W} : C(V) \leq R\}$ and $\mathcal{P}_R = \{P \in \mathcal{P} : I(P,W) \geq R\}$. Then, $\mathcal{V}_R$ is increasing and $\mathcal{P}_R$ is decreasing as a set with increasing R. Therefore, $\widetilde{E}(R)$ is a monotone nonincreasing function of R, proving (a). For (b), note that $\widetilde{E}(R) \leq E_h(R)$ obviously because the inner maximization's feasible set in $\widetilde{E}(R)$ is expanded to give $E_h(R)$. Recalling the definition of the sphere-packing exponent,

$$\begin{aligned}
E_{sp}(R) &\triangleq \max_{P} \min_{V:I(P,V)\leq R} D(V||W|P) \\
&\overset{(A)}{=} \max_{P:I(P,W)\geq R} \min_{V:I(P,V)\leq R} D(V||W|P) \\
&\overset{(B)}{\leq} \max_{P:I(P,W)\geq R} \min_{V:C(V)\leq R} D(V||W|P) \\
&\overset{(C)}{\leq} \min_{V:C(V)\leq R} \max_{P:I(P,W)\geq R} D(V||W|P) \\
&= \widetilde{E}(R),
\end{aligned}$$

where (A) follows because if $I(P,W) \leq R$, W is included in the minimization, (B) follows because we are making the feasible set in the minimization smaller and (C) follows because max-min is smaller than min-max. As for the proof of (c), we need to be careful about discontinuities in $\widetilde{E}(R)$, which really can only happen when the divergence jumps to ∞.

For real vector spaces, define a set 'distance' $d(\mathcal{A},\mathcal{B})$ between two sets $\mathcal{A},\mathcal{B}$ to be

$$d(\mathcal{A},\mathcal{B}) = \max\left\{\sup_{a\in\mathcal{A}} \inf_{b\in\mathcal{B}} ||a-b||_1, \sup_{b\in\mathcal{B}} \inf_{a\in\mathcal{A}} ||a-b||_1\right\}$$

[1] There is no feasible P in the inner maximization if $R > C(W)$ and $V = W$ is feasible for the outer minimization for $R = C(W)$.

where $||\cdot||_1$ denotes the $\mathcal{L}_1$ norm. We say that $\mathcal{V}_R$ is continuously increasing with R because for all $R \in (0, C(W)), \epsilon > 0$, there is a $\delta > 0$ such that if $|R' - R| \leq \delta$, $d(\mathcal{V}_R, \mathcal{V}_{R'}) \leq \epsilon$. This is due to the fact that capacity is continuous (and convex-∪) with the channel V with respect to $\mathcal{L}_1$ norm. Similarly, $\mathcal{P}_R$ is a continuously decreasing set with R. So, we have

$$\widetilde{E}(R) = \min_{V \in \mathcal{V}_R} \max_{P \in \mathcal{P}_R} D(V||W|P).$$

Now, for $R \in (0, C(W))$, let

$$\mathcal{X}_R = \left\{ x \in \mathcal{X} : \max_{P \in \mathcal{P}_R} P(x) > 0 \right\}.$$

We claim that $\mathcal{X}_R$ is independent of R, i.e. there is a $\mathcal{X}' \subset \mathcal{X}$ such that $\mathcal{X}' = \mathcal{X}_R$ for all $R \in (0, C(W))$. To see this, fix an $R, R' \in (0, C(W))$ with $R' < R$. Since $\mathcal{P}_R$ is shrinking with increasing R, $\mathcal{X}_R$ is shrinking with increasing R, so $\mathcal{X}_R \subseteq \mathcal{X}_{R'}$. Now, suppose $x \in \mathcal{X}_{R'}$. There is a $P' \in \mathcal{P}_{R'}$ with $P'(x) > 0$. Let $P^* \in \mathcal{P}_{C(W)}$ be a capacity achieving distribution and for $\alpha \in (0,1)$, let $P_\alpha = \alpha P^* + (1-\alpha)P'$. The mutual information $I(P, W)$ is concave-∩ in the input distribution P, so

$$\begin{aligned} I(P_\alpha, W) &\geq \alpha I(P^*, W) + (1-\alpha) I(P', W) \\ &\geq \alpha C(W) + (1-\alpha) R'. \end{aligned}$$

For α close enough to 1, but strictly smaller, $P_\alpha \in \mathcal{P}_R$, and $P_\alpha(x) > 0$. Therefore, $x \in \mathcal{X}_R$. Hence, $\mathcal{X}_R \subset \mathcal{X}_{R'}$ also. So now, let $\mathcal{X}'$ be $\mathcal{X}_R$ for $R \in (0, C(W))$. Let

$$\begin{aligned} \mathcal{V}' &\triangleq \{V \in \mathcal{W} : \forall\, x \in \mathcal{X}', \forall\, y \in \mathcal{Y}, W(y|x) = 0 \Rightarrow V(y|x) = 0\} \\ \mathcal{V}'_R &\triangleq \mathcal{V}_R \cap \mathcal{V}'. \end{aligned}$$

Then, in the definition of $\widetilde{E}(R)$ for $R \in (0, C(W))$, we can restrict attention to $V \in \mathcal{V}'_R$ because if $V \notin \mathcal{V}'_R$, there is a $P \in \mathcal{P}_R$ that causes the inner maximization to be infinite.

$$\begin{aligned} \widetilde{E}(R) &= \min_{V \in \mathcal{V}'_R} \max_{P \in \mathcal{P}_R} D(V||W|P) \\ &= \min_{V \in \mathcal{V}'_R} g(V, R) \\ g(V, R) &\triangleq \max_{P \in \mathcal{P}_R} D(V||W|P). \end{aligned}$$

Now, restricting to the domain of $\mathcal{V}'$, $g(V, R)$ is continuous in V for a fixed R because $D(V||W|P)$ is continuous in V. Similarly, $g(V, R)$ is continuous in R for a fixed V provided that V in $\mathcal{V}'$ because $D(V||W|P)$ is continuous in P and $\mathcal{P}_R$ is continuously decreasing in R. Now, if $\widetilde{E}(R) = \infty$, monotonicity guarantees left continuity at R. If $\widetilde{E}(R) < \infty$, this means that $\mathcal{V}'_R$ is not empty. $E_{sp}(R)$ and $E_h(R)$ become infinite at the same R, both are left continuous, and $\widetilde{E}(R)$ is sandwiched between the two, so for small enough $\delta > 0$, this means that $\mathcal{V}'_{R-\delta}$ is also not empty. Since $\mathcal{V}'_R$ is continuously varying with R, and $g(V, R)$ is continuous in R and V, it follows that $\widetilde{E}(R)$ is left continuous.

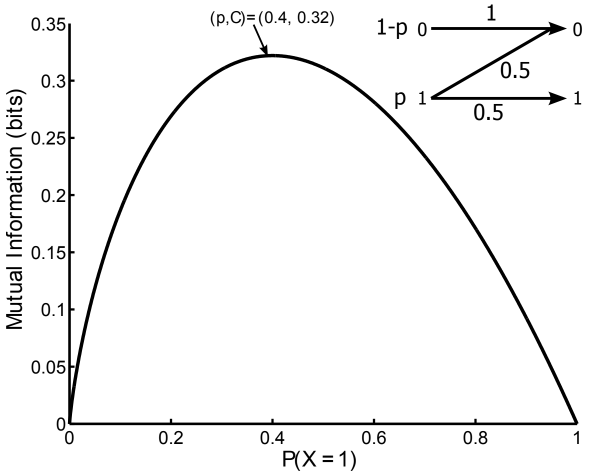

Figure B.1: Mutual information $I_Z(p, 1/2)$ for a Z-channel with crossover probability 1/2. The mutual information is strictly concave-$\cap$ on $(0, 1)$ and 0 at the endpoints, so there is a unique maximizing p and the capacity is I_Z evaluated at that p.

B.2 The Z-Channel

B.2.1 Capacity and capacity achieving distribution

In this section, we will derive simple expressions for $\widetilde{E}(R)$, $E_h(R)$ and $E_{sp}(R)$ when W is in the family of Z-channels. A Z-channel (shown in Fig. 2.3) is an asymmetric binary input, binary output ($\mathcal{X} = \mathcal{Y} = \{0, 1\}$) channel for which

$$W = \begin{bmatrix} 1 & 0 \\ \delta & 1 - \delta \end{bmatrix},$$

where δ is called the crossover probability. The Z-channel is particularly amenable to computation and visualization of error exponents because it is a one-parameter family of channels and the input distributions for the channel can also be described with one parameter. For the remainder of this section, we will let p denote the probability that an input distribution P places on 1, so $P(1) = p$.

Define the mutual information for a Z-channel of crossover probability δ and input distribution p to be $I_Z(p, \delta)$, and note that

$$I_Z(p, \delta) = h_b\left(p(1 - \delta)\right) - p h_b(\delta).$$

Similarly, let $C_Z(\delta)$ be the capacity of a Z channel with crossover probability δ,

$$C_Z(\delta) = \max_{p \in [0,1]} I_Z(p, \delta).$$

Since $I_Z(p,\delta)$ is strictly concave-$\cap$ (see Fig. B.1) for $\delta \in [0,1)$, it follows that there is a unique capacity achieving p for each δ.

Proposition 16. *Let $p^*(\delta)$ denote the capacity achieving p for a Z-channel with crossover probability δ. Then, for all $\delta \in [0,1)$,*

$$p^*(\delta) = \left[(1-\delta)\left(1+\exp\left(\frac{h_b(\delta)}{1-\delta}\right)\right)\right]^{-1}, \tag{B.8}$$

and

$$\begin{aligned} C_Z(\delta) &= h_b\left(p^*(\delta)(1-\delta)\right) - p^*(\delta)h_b(\delta) \\ &= h_b\left(\left[1+\exp\left(\frac{h_b(\delta)}{1-\delta}\right)\right]^{-1}\right) - \frac{h_b(\delta)}{(1-\delta)\left(1+\exp\left(\frac{h_b(\delta)}{1-\delta}\right)\right)}. \end{aligned} \tag{B.9}$$

A plot of $p^*(\delta)$ is shown in Fig. 2.5 and a plot of $C_Z(\delta)$ is given in Fig. 2.4.

Proof: Assume that all log's and exponentials are base e. Then, for $t \in (0,1)$,

$$\frac{dh_b(t)}{dt} = \log\frac{1-t}{t},$$

and hence,

$$\begin{aligned} \frac{\partial I_Z(p,\delta)}{\partial \delta} &= \frac{\partial}{\partial p}\left[h_b\left(p(1-\delta)\right) - ph_b(\delta)\right] \\ &= (1-\delta)\log\frac{1-p(1-\delta)}{p(1-\delta)} - h_b(\delta). \end{aligned}$$

Now, $I_Z(p,\delta)$ is strictly concave-$\cap$ in p, nonnegative and $I_Z(0,\delta) = I_Z(1,\delta) = 0$, so we can set the derivative to 0 and solve for p to yield the capacity achieving $p^*(\delta)$ in (B.8). Then $C_Z(\delta) = I_Z(p^*(\delta),\delta)$, yielding (B.9).

B.2.2 Evaluating error exponents

The mutual information $I_Z(p,\delta)$ is convex-$\cup$ in δ for a fixed p and the capacity, $C_Z(\delta)$ is convex-$\cup$ in δ as it is the maximum of a set of convex-$\cup$ functions. It is easy to see as well that both are strictly decreasing in δ. The range of $C_Z(\delta)$ is $[0,1]$, so let it's inverse function be, for $R \in [0,1]$,

$$f(R) \triangleq C_Z^{-1}(R) = \min\{\beta \in [0,1] : C_Z(\beta) \leq R\}. \tag{B.10}$$

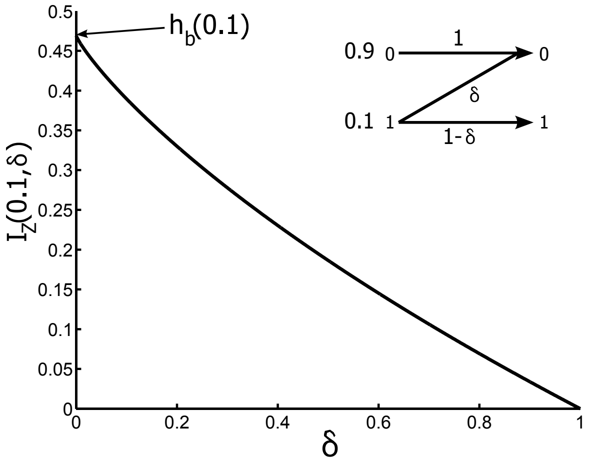

Figure B.2: $I_Z(p,\delta)$ is a strictly decreasing, convex-$\cup$ function of δ for all $p \in (0,1)$. It is plotted here for $p = 0.1$. The left endpoint is $h_b(p)$.

Similarly, for a fixed $p \in [0,1]$, the range of $I_Z(p,\delta)$ is $[0, h_b(p)]$ (see Fig. B.2), so for $R \in [0, h_b(p)]$, let the inverse function be denoted

$$g(R,p) \triangleq I_Z^{-1}(p,R) = \min\{\beta \in [0,1] : I_Z(p,\beta) \leq R\}. \tag{B.11}$$

Both $f(R)$ and $g(R,p)$ are well defined, convex-$\cup$ and strictly monotonically decreasing with R (for all $p \in (0,1)$ for $g(R,p)$).

We will evaluate the three exponents $E_h(R), \widetilde{E}(R)$ and $E_{sp}(R)$ when W is a Z-channel with crossover probability δ. For all three exponents, it will be sufficient to consider test channels V that are also Z-channels. The reason is that if a binary input, binary output channel V has $V(0|1) > 0$, then $D(V||W|P) = \infty$ if $P(1) > 0$. Therefore, no such channel would be included in the minimizations over channels V for the exponents. Now, if V is a Z-channel with crossover probability β, then

$$\begin{aligned} D(V||W|P) &= \sum_{x,y\in\{0,1\}} P(x)V(y|x)\log\frac{V(y|x)}{W(y|x)} \\ &= P(0)\times 0 + P(1)\left[\beta\log\frac{\beta}{\delta} + (1-\beta)\log\frac{1-\beta}{1-\delta}\right] \\ &= P(1)D_b(\beta||\delta), \end{aligned}$$

where $D_b(\beta||\delta)$ denotes the binary divergence

$$D_b(\beta||\delta) \triangleq \beta\log\frac{\beta}{\delta} + (1-\beta)\log\frac{1-\beta}{1-\delta}.$$

Proposition 17. *If W is a Z-channel with crossover probability δ, then, for $R \in [0, C_Z(\delta)]$,*

$$\begin{aligned} E_h(R) &= D_b\left(f(R)||\delta\right) \\ \widetilde{E}(R) &= p_r(R,\delta) D_b\left(f(R)||\delta\right) \\ E_{sp}(R) &= \max_{p\in[0,1]:I_Z(p,\delta)\geq R} pD_b\left(g(R,p)||\delta\right), \end{aligned}$$

where g and f are defined in (B.11) and (B.10) respectively and

$$p_r(R,\delta) \triangleq \max\{p \in [0,1] : I_Z(p,\delta) \geq R\}.$$

Proof: First,

$$E_h(R) = \min_{\beta\in[0,1]:C_Z(\beta)\leq R} \max_{p\in[0,1]} pD_b(\beta||\delta).$$

The inner maximization is independent of β, and D_b is nonnegative so $p = 1$ is the maximizer. Also, since $C_Z(\beta)$ is monotonically decreasing and $R \leq C_Z(\delta)$, and $D_b(\beta||\delta)$ is increasing in β for $\beta \geq \delta$, it follows that the minimizing β is $f(R)$.

Second,

$$\widetilde{E}(R) = \min_{\beta\in[0,1]:C_Z(\beta)\leq R} \max_{p\in[0,1]:I_Z(p,\delta)\geq R} pD_b(\beta||\delta).$$

Similarly to the argument for $E_h(R)$, the maximizing p in the inner max is the largest p such that $I(p,\delta) \geq R$, which we denote $p_r(R,\delta)$, and the minimizing β is $f(R)$.

Finally,

$$\begin{aligned} E_{sp}(R) &= \max_{p\in[0,1]} \min_{\beta:I_Z(p,\beta)\leq R} pD_b(\beta||\delta) \\ &= \max_{p\in[0,1]:I_Z(p,\delta)\geq R} \min_{\beta:I_Z(p,\beta)\leq R} pD_b(\beta||\delta) \\ &= \max_{p\in[0,1]:I_Z(p,\delta)\geq R} E_{sp}(R,p) \\ E_{sp}(R,p) &\triangleq \min_{\beta:I_Z(p,\beta)\leq R} pD_b(\beta||\delta) \\ &= pD_b\left(g(R,p)||\delta\right), \end{aligned}$$

where we may as well remove p such that $I(p,\delta) < R$ because $E_{sp}(R,p)$ evaluates to 0 for those p (as $D_b(\delta||\delta) = 0$). Similarly to $C_Z(\beta)$, for a fixed p, $I_Z(p,\beta)$ is monotonically decreasing in β. For those p that have $I_Z(p,\delta) \geq R$, the set of β that have $I_Z(p,\beta) < R$ all lie to the right of δ, hence the minimizing β is $g(R,p)$.

Obtaining an analytical expression for the maximizing p in $E_{sp}(R)$ seems to be difficult. Fig. B.3 shows a plot of $E_{sp}(R,p)$, and it can be seen to be non-concave-$\cap$, but quasi-concave ($\cap$) in p. Define the sphere-packing optimizing p to be

$$p^*_{sp}(R,\delta) \triangleq \arg \max_{p\in[0,1]:I_Z(p,\delta)\geq R} E_{sp}(R,p).$$

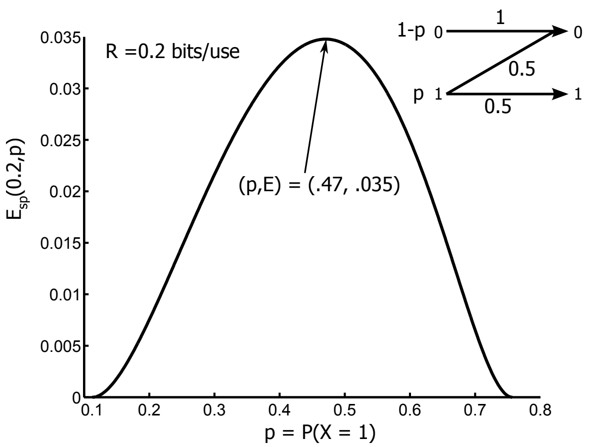

Figure B.3: A plot of $E_{sp}(R,p)$ in base 2 for $R = 0.2$ (bits/channel use), when $\delta = 0.5$, $C_Z(\delta) = 0.322$ bits/channel use. As a function of p, $E_{sp}(R,p)$ is not concave-$\cap$, but is quasi-concave.

Figure 2.6 shows a plot of $p^*_{sp}(R,\delta)$ for $\delta = 0.5$ and R near capacity. As can be seen from the figure, $p^*_{sp}(R,\delta)$ approaches $p^*(\delta)$ from above as $R \to C_Z(\delta)$ with a non-zero slope.

The expressions in Proposition 17 can be used to quickly plot the three exponents, as shown in Fig. 3.1, because evaluating the inverse functions of C_Z and I_Z can be done in logarithmic time as the accuracy desired tends to 0. Evaluating the max over p for E_{sp} is done by brute force.

B.2.3 Exponents near capacity

Proposition 18. *If W is a Z-channel of crossover probability $\delta \in (0,1)$,*

$$\lim_{R\to C_Z(\delta)} \frac{E_h(R)}{E_{sp}(R)} \geq \frac{1}{p^*(\delta)} \geq 2.$$

Proof: We know that $E_h(R) = D_b(f(R)||\delta)$. Now, since $C_Z(f(R)) = R$, this means that for all $p \in [0,1], I_Z(p, f(R)) \leq R$, hence we can upper bound $E_{sp}(R)$ as

$$\begin{aligned} E_{sp}(R) &\leq \max_{p\in[0,1]:I_Z(p,\delta)\geq R} pD_b(f(R)||\delta) \\ &= p_r(R,\delta)D_b(f(R)||\delta). \end{aligned}$$

Using the fact that $\lim_{R\to C_Z(\delta)} p_r(R,\delta) = p^*(\delta)$ yields

$$\begin{aligned}\lim_{R\to C_Z(\delta)} \frac{E_h(R)}{E_{sp}(R)} &\geq \lim_{R\to C_Z(\delta)} \frac{D_b(f(R)||\delta)}{p_r(R,\delta)D_b(f(R)||\delta)}\\ &= \frac{1}{\lim_{R\to C_Z(\delta)} p_r(R,\delta)}\\ &= \frac{1}{p^*(\delta)}.\end{aligned}$$

For all values of $\delta \in [0,1)$, $p^*(\delta) \leq 1/2$.

With regards to $\widetilde{E}(R)$, it can be seen from a plot (Fig. 3.3) that

$$\lim_{R\to C_Z(\delta)} \frac{\widetilde{E}(R)}{E_{sp}(R)} = 1.$$

This fact can also be proved by looking at Taylor expansions of the two exponents around capacity.

Proposition 19. *If W is a Z-channel with crossover probability $\delta \in (0,1)$,*

$$\lim_{R\to C_Z(\delta)} \frac{\widetilde{E}(R)}{E_{sp}(R)} = 1.$$

Proof: First, fix a δ and let $C = C_Z(\delta)$. It is straightforward to check that the second-order expansion of $D_b(\delta+\epsilon||\delta)$ around $\epsilon = 0$ is

$$D_b(\delta+\epsilon||\delta) = \frac{\epsilon^2}{2\delta(1-\delta)} + O(\epsilon^3),$$

where the notation $O(\epsilon^3)$ means that there are constants $K > 0$ and $\epsilon_K > 0$ such that if $|\epsilon| \leq \epsilon_K$,

$$|O(\epsilon^3)| \leq K|\epsilon|^3.$$

Now, recall that

$$\widetilde{E}(R) = p_r(R,\delta)D_b(f(R)||\delta).$$

where $p_r(R,\delta)$ is the largest p such that $I_Z(p,\delta) \geq R$ and $f(R)$ is the inverse function of $C_Z(\beta)$ defined in (B.10). By continuity[2] of $I_Z(p,\delta)$ there is a real valued function $k(\cdot)$ such that

$$p_r(R,\delta) = p^*(\delta) + k(R-C),$$

[2] $I_Z(p,\delta)$ has a derivative of 0 at $p = p^*(\delta)$, hence the first derivative of $p_r(R,\delta)$ at C is undefined, i.e. $-\infty$. That does not prevent us from using continuity however.

and $\lim_{t\to 0} k(t) = 0$. Note also that $f(R)$ has a Taylor expansion about $R = C$, which is

$$f(R) = \delta + f'(C)(R - C) + O(|R - C|^2).$$

Using these expansions, we have

$$\begin{aligned}
\widetilde{E}(R) &= (p^*(\delta) + k(R - C)) \times D_b \left(\delta + f'(C)(R - C) + O(|R - C|^2)||\delta\right) \\
&= (p^*(\delta) + k(R - C)) \times \left[\frac{\left(f'(C)(R - C) + O(|R - C|^2)\right)^2}{2\delta(1 - \delta)} + O(|R - C|^3)\right] \\
&= (p^*(\delta) + k(R - C)) \times \left[\frac{f'(C)^2(R - C)^2}{2\delta(1 - \delta)} + O(|R - C|^3)\right] \\
&= \frac{p^*(\delta)f'(C)^2}{2\delta(1 - \delta)}(R - C)^2 + \frac{f'(C)^2}{2\delta(1 - \delta)}(R - C)^2 k(R - C) + O(|R - C|^3). \qquad \text{(B.12)}
\end{aligned}$$

Now, we already know that $E_{sp}(R) \leq \widetilde{E}(R)$ for all R, so we want a lower bound on $E_{sp}(R)$, because a simple expression for the sphere-packing optimizing $p^*_{sp}(R, \delta)$ is lacking. Hence, we take the lower bound

$$\begin{aligned}
E_{sp}(R) &= \max_{p: I_Z(p,\delta) \geq R} p D_b(g(R, p)||\delta) \\
&\geq p^*(\delta) D_b(g(R, p^*(\delta))||\delta).
\end{aligned}$$

Having fixed $p = p^*(\delta)$ for all R, we can define the function of one variable,

$$\widetilde{g}(R) \triangleq g(R, p^*(\delta)).$$

We can take a Taylor expansion of $\widetilde{g}$ to yield

$$\widetilde{g}(R) = \delta + \widetilde{g}'(C)(R - C) + O((R - C)^2).$$

Plugging this into the lower bound for $E_{sp}(R)$ gives

$$\begin{aligned}
E_{sp}(R) &\geq p^*(\delta) D_b \left(\delta + \widetilde{g}'(C)(R - C) + O((R - C)^2)||\delta\right) \\
&= p^*(\delta) \left[\frac{\left(\widetilde{g}'(C)(R - C) + O(|R - C|^2)\right)^2}{2\delta(1 - \delta)} + O(|R - C|^3)\right] \\
&= p^*(\delta) \left[\frac{\widetilde{g}'(C)^2(R - C)^2}{2\delta(1 - \delta)} + O(|R - C|^3)\right] \\
&= \frac{p^*(\delta)\widetilde{g}'(C)^2}{2\delta(1 - \delta)}(R - C)^2 + O(|R - C|^3). \qquad \text{(B.13)}
\end{aligned}$$

At this point, we would like to show that $f'(C) = \widetilde{g}'(C)$. Since $f(R)$ is the inverse function of $C_Z(\beta)$, it follows that

$$f'(C) = \left[\frac{dC_Z(\beta)}{d\beta}\bigg|_{\beta=\delta} \right]^{-1}.$$

Now, $C_Z(\beta) = I_Z(p^*(\beta), \beta)$, so by chain rule,

$$\begin{aligned} C_Z'(\beta) &= \nabla I_Z(p,\tau)\bigg|_{(p,\tau)=(p^*(\beta),\beta)} \cdot \left[\frac{dp^*(\tau)}{d\tau}\bigg|_{\tau=\beta} \quad 1 \right] \\ &= \left[\frac{\partial I_z(p,\beta)}{\partial p}\bigg|_{p=p^*(\beta)} \quad \frac{\partial I_Z(p^*(\beta),\tau)}{\partial \tau}\bigg|_{\tau=\beta} \right] \cdot \left[\frac{dp^*(\tau)}{d\tau}\bigg|_{\tau=\beta} \quad 1 \right], \end{aligned}$$

where $\cdot$ denotes dot product in the above. Since $p^*(\beta)$ is the capacity achieving p for β, it follows by concavity of $I_Z(p,\beta)$ in p that

$$\begin{aligned} \frac{\partial I_z(p,\beta)}{\partial p}\bigg|_{p=p^*(\beta)} &= 0 \\ C_Z'(\delta) = \frac{\partial I_Z(p^*(\delta),\tau)}{\partial \tau}\bigg|_{\tau=\delta}&. \end{aligned}$$

Now, $\widetilde{g}(R) = g(R, p^*(\delta))$ is the inverse function of $I_Z(p^*(\delta), \beta)$ when viewed as a function of β. Therefore,

$$\begin{aligned} \widetilde{g}'(R) &= \left[\frac{\partial I_Z(p^*(\delta),\beta)}{\partial \beta}\bigg|_{\beta : I_Z(p^*(\delta),\beta)=R} \right]^{-1} \\ \widetilde{g}'(C) &= \left[\frac{\partial I_Z(p^*(\delta),\beta)}{\partial \beta}\bigg|_{\beta=\delta} \right]^{-1} \\ &= \left[\frac{dC_z(\beta)}{d\beta}\bigg|_{\beta=\delta} \right]^{-1} \\ &= f'(C). \end{aligned}$$

Therefore, using the fact that $f'(C) = \widetilde{g}'(C)$ and the expansions of (B.12) and (B.13),

$$\begin{aligned} 1 \le \lim_{R\to C} \frac{\widetilde{E}(R)}{E_{sp}(R)} &\le \lim_{R\to C} \frac{\frac{p^*(\delta) f'(C)^2}{2\delta(1-\delta)}(R-C)^2 + \frac{f'(C)^2}{2\delta(1-\delta)}(R-C)^2 k(R-C) + O(|R-C|^3)}{\frac{p^*(\delta)\widetilde{g}'(C)^2}{2\delta(1-\delta)}(R-C)^2 + O(|R-C|^3)} \\ &= \lim_{R\to C} \frac{1 + k(R-C)/p^*(\delta) + O(|R-C|)}{1 + O(|R-C|)} \\ &= 1, \end{aligned}$$

where the last line follows because $k(R-C) \to 0$ as $R \to C$.

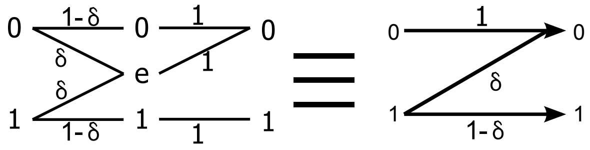

Figure B.4: Simulating a Z-channel by using a BEC and collapsing the erasure symbol into 0.

B.2.4 The back-story bound for the Z-channel

Let $E_{fb}(R;\mathcal{Z}(\delta))$ denote the error exponent with feedback for the Z-channel with crossover probability δ. and let $E_{sp}(R;BEC(\delta))$ denote the sphere-packing exponent for the BEC with erasure probability δ. It is readily seen that for $R \in [0, 1-\delta]$,

$$E_{sp}(R;BEC(\delta)) = D_b(1-R||\delta)$$

and we have already seen that

$$E_{fb}(R;\mathcal{Z}(\delta)) \leq E_h(R;\mathcal{Z}_(\delta)) = D_b(C_Z^{-1}(R)||\delta).$$

Now, one can simulate $\mathcal{Z}(\delta)$ with $BEC(\delta)$ as shown in Figure B.4 by collapsing the erasure symbol and 0 into 0 at the output of the BEC. We know from the 'back-story' bound that

$$E_{fb}(R;\mathcal{Z}(\delta)) \leq E_{sp}(R;BEC(\delta)).$$

The question then becomes whether $E_{sp}(R;BEC(\delta)) < E_h(R;\mathcal{Z}_(\delta)$ or not. From the plot of $C_Z(\delta)$ in Figure 2.4, we know that $C_Z(\beta) \leq 1-\beta$ for all $\beta \in [0,1]$ (with strict inequality for $\beta \in (0,1)$). Therefore, $1-R > C_Z^{-1}(R) > \delta$ if $\delta \in (0,1)$. Therefore

$$\begin{aligned} E_{sp}(R;BEC(\delta)) &= D_b(1-R||\delta) \\ &> D_b(C_Z^{-1}(R)||\delta) \\ &= E_h(R;\mathcal{Z}(\delta)). \end{aligned}$$

So the back-story bound is not tighter than Haroutunian if we use a BEC as the back-story for a Z-channel. It seems unlikely to get a different result by using a larger channel than the BEC as the back-story to the Z-channel. Numerical evaluations of the BSC followed by a Z-channel as the back-story to a binary asymmetric channel have also not yielded anything tighter than the Haroutunian bound.

Appendix C

Arbitrarily varying sources appendix

C.1 The cheating switcher with clean observations

Theorem 9 of Section 4.3: Define the set of distributions

$$\mathcal{C} = \left\{ p \ : \ \begin{array}{c} \sum_{x \in \mathcal{V}} p(x) \geq P\big(\forall\ l, x_l \in \mathcal{V}\big) \\ \forall\ \mathcal{V} \text{ such that} \\ \mathcal{V} \subseteq \mathcal{X} \end{array} \right\},$$

where the event $\{\forall\ l, x_l \in \mathcal{V}\}$ is shorthand for $\{(x_1, \ldots, x_m) : x_l \in \mathcal{V}, l = 1, \ldots, m\}$. Also, define

$$\widetilde{R}(D) \triangleq \max_{p \in \mathcal{C}} R(p, D).$$

For a general set of distributions $\mathcal{Q} \subset \mathcal{P}(\mathcal{X})$, let $D_{\min}(\mathcal{Q}) \triangleq \sup_{p \in \mathcal{Q}} D_{\min}(p)$. Suppose the switcher has either 1-step lookahead or full lookahead. In both cases, for $D > D_{\min}(\mathcal{C})$,

$$R(D) = \widetilde{R}(D)$$

For $D < D_{\min}(\mathcal{C})$, $R(D) = \infty$ by convention because the switcher can simulate a distribution for which the distortion D is infeasible for the coder.

C.1.1 Achievability for the coder

The main tool of the proof is:

Lemma 26 (Type Covering)**.** *Let $S_D(\widehat{\mathbf{x}}^n) \triangleq \{\mathbf{x}^n \in \mathcal{X}^n : d_n(\mathbf{x}^n, \widehat{\mathbf{x}}^n) \leq D\}$ be the set of $\mathcal{X}^n$ strings that are within distortion D of a given $\widehat{\mathcal{X}}^n$ string $\widehat{\mathbf{x}}^n$. Fix an $\epsilon > 0$. Then for all $n \geq n_0(d, \epsilon)$, for any $p \in \mathcal{P}_n(\mathcal{X})$, there exists a codebook $\mathcal{B} = \{\widehat{\mathbf{x}}^n(1), \widehat{\mathbf{x}}^n(2), \ldots, \widehat{\mathbf{x}}^n(M)\}$ where $M \leq \exp(n(R(p, D) + \epsilon))$ and*

$$T_p^n \subseteq \bigcup_{\widehat{\mathbf{x}}^n \in \mathcal{B}} S_D(\widehat{\mathbf{x}}^n),$$

where T_p^n is the set of $\mathcal{X}^n$ strings with type p.

Proof: See [47], Lemma 2.4.1. Note that $n_0(d, \epsilon)$ is independent of both p and D.

We now show how the coder can get arbitrarily close to $\widetilde{R}(D)$ for large enough n. For a $\delta > 0$,

$$\mathcal{C}_\delta \triangleq \left\{ \begin{array}{ccc} & & \sum_{x \in \mathcal{V}} p(x) \geq P(\forall l, x_l \in \mathcal{V}) - \delta \\ p \in \mathcal{P} & : & \forall\ \mathcal{V} \text{ such that} \\ & & \mathcal{V} \subseteq \mathcal{X} \end{array} \right\}.$$

Lemma 27 (Converse for switcher)**.** *Let $\epsilon > 0$. For all n sufficiently large*

$$\frac{1}{n} \ln M(n, D) \leq \widetilde{R}(D) + \epsilon.$$

Proof: Fix a $\lambda > 0$ and $\lambda \leq \lambda(\epsilon) < D - D_{\min}(\mathcal{C})$ to be defined later. We know $R(p, D - \lambda)$ is a continuous function of p ([47]). It follows then that because $\mathcal{C}_\delta$ is monotonically decreasing (as a set) with δ that for all $\epsilon > 0$, there is a $\delta > 0$ so that

$$\max_{p \in \mathcal{C}_\delta} R(p, D - \lambda) \leq \max_{p \in \mathcal{C}} R(p, D - \lambda) + \epsilon/3.$$

We will have the coder use a codebook such that all $\mathcal{X}^n$ strings with types in $\mathcal{C}_\delta$ are covered within distortion $D - \lambda$. The coder can do this for large n with at most M codewords in the codebook $\mathcal{B}$, where

$$\begin{aligned} M &\leq (n+1)^{|\mathcal{X}|} \exp\left(n \left(\max_{p \in \mathcal{C}_\delta} R(p, D - \lambda) + \epsilon/3 \right) \right) \\ &\leq \exp(n(\max_{p \in \mathcal{C}} R(p, D - \lambda) + \epsilon)). \end{aligned}$$

Explicitly, this is done by taking a union of the codebooks provided by the type-covering lemma and noting that the number of types in $\mathcal{P}_n(\mathcal{X})$ is less than $(n+1)^{|\mathcal{X}|}$. Next, we will show that the probability of the switcher being able to produce a string with a type not in $\mathcal{C}_\delta$ goes to 0 exponentially with n.

Consider a type $p \in \mathcal{P}_n(\mathcal{X}) \cap (\mathcal{P}(\mathcal{X}) - \mathcal{C}_\delta)$. By definition, there is some $\mathcal{V} \subseteq \mathcal{X}$ such that $\sum_{x \in \mathcal{V}} p(x) < P(x_l \in \mathcal{V}, 1 \leq l \leq m) - \delta$. Let $\zeta_k(\mathcal{V})$ be the indicator function

$$\zeta_k(\mathcal{V}) = \prod_{l=1}^{m} \mathbf{1}(x_{l,k} \in \mathcal{V}).$$

ζ_k indicates the event that the switcher cannot output a symbol outside of $\mathcal{V}$ at time k. Then $\zeta_k(\mathcal{V})$ is a Bernoulli random variable with a probability of being 1 equal to $\kappa(\mathcal{V}) \triangleq P(x_l \in \mathcal{V}, 1 \leq l \leq m)$. Since the subsources are IID over time, $\zeta_k(\mathcal{V})$ is a sequence of IID binary random variables with distribution $q' \triangleq (1 - \kappa(\mathcal{V}), \kappa(\mathcal{V}))$.

Now for the type $p \in \mathcal{P}_n(\mathcal{X}) \cap (\mathcal{P}(\mathcal{X}) - \mathcal{C}_\delta)$, we have that for all strings $\mathbf{x}^n$ in the type class T_p, $\frac{1}{n}\sum_{i=1}^n \mathbf{1}(x_i \in \mathcal{V}) < \kappa(\mathcal{V}) - \delta$. Let p' be the binary distribution $(1 - \kappa(\mathcal{V}) + \delta, \kappa(\mathcal{V}) - \delta)$. Therefore $||p' - q'||_1 = 2\delta$, and hence we can bound the binary divergence $D(p'||q') \geq 2\delta^2$ by Pinsker's inequality. Using standard types properties [21] gives

$$\begin{aligned} P\left(\frac{1}{n}\sum_{k=1}^n \zeta_k(\mathcal{V}) < \kappa(\mathcal{V}) - \delta\right) &\leq (n+1)\exp(-nD(p'||q')) \\ &\leq (n+1)\exp(-2n\delta^2). \end{aligned}$$

This bound holds for all $\mathcal{V} \subset \mathcal{X}, \mathcal{V} \neq \emptyset$, so we sum over types not in $\mathcal{C}_\delta$ to get

$$\begin{aligned} P(p_{\mathbf{x}^n} \notin \mathcal{C}_\delta) &\leq \sum_{p \in \mathcal{P}_n(\mathcal{X}) \cap (\mathcal{P}(\mathcal{X}) - \mathcal{C}_\delta)} (n+1)\exp(-2n\delta^2) \\ &\leq (n+1)^{|\mathcal{X}|}\exp(-2n\delta^2) \\ &= \exp\left(-n\left(2\delta^2 - |\mathcal{X}|\frac{\ln(n+1)}{n}\right)\right). \end{aligned}$$

Then, regardless of the switcher strategy,

$$\mathbb{E}[d(\mathbf{x}^n; \mathcal{B})] \leq D - \lambda + d^* \times \exp\left(-n\left(2\delta^2 - |\mathcal{X}|\frac{\ln(n+1)}{n}\right)\right).$$

So for large n we can get arbitrarily close to distortion $D - \lambda$ while the rate is at most $\max_{p \in \mathcal{C}} R(p, D - \lambda) + \epsilon$. Using the fact that the IID rate-distortion function is continuous in D (uniformly over p such that $D_{\min}(p) < D$, see (C.8)) gives us that the coder can achieve at most distortion D on average while the asymptotic rate is at most $\widetilde{R}(D) + 2\epsilon$ (provided $\lambda \leq \lambda(\epsilon)$ is small enough). Since ϵ is arbitrary, $R(D) \leq \widetilde{R}(D)$.

C.1.2 Achievability for the switcher

This section shows that $R(D) \geq \widetilde{R}(D)$ when the switcher has 1-step lookahead. We will show that the switcher can target any distribution $p \in \mathcal{C}$ and produce a sequence of IID symbols with distribution p. In particular, the switcher can target the distribution that yields $\max_{p \in \mathcal{C}} R(p, D)$, so $R(D) \geq \widetilde{R}(D)$.

The switcher will use a memoryless randomized strategy. Let $\mathcal{V} \subseteq \mathcal{X}$ and suppose that at some time k the set of symbols available to choose from for the switcher is exactly $\mathcal{V}$, i.e. $\{x_{1,k}, \ldots, x_{m,k}\} = \mathcal{V}$. Recall $\beta(\mathcal{V}) \triangleq P(\{x_{1,1}, \ldots, x_{m,1}\} = \mathcal{V})$ is the probability that at any time the switcher must choose among elements of $\mathcal{V}$ and no other symbols. Then let $f(x|\mathcal{V})$ be a probability distribution on $\mathcal{X}$ with support $\mathcal{V}$, i.e. $f(x|\mathcal{V}) \geq 0$, $\forall\, x \in \mathcal{X}$, $f(x|\mathcal{V}) = 0$ if $x \notin \mathcal{V}$, and $\sum_{x \in \mathcal{V}} f(x|\mathcal{V}) = 1$. The switcher will have such a randomized rule for every nonempty subset $\mathcal{V}$ of $\mathcal{X}$ such that $|\mathcal{V}| \leq m$. Let $\mathcal{D}$ be the set of distributions on $\mathcal{X}$ that can be achieved with these kinds of rules,

$$\mathcal{D} = \left\{ p \quad : \quad \begin{array}{c} p(\cdot) = \sum_{\mathcal{V} \subseteq \mathcal{X}, |\mathcal{V}| \leq m} \beta(\mathcal{V}) f(\cdot|\mathcal{V}), \\ \forall\, \mathcal{V} \text{ s.t. } \mathcal{V} \subseteq \mathcal{X},\ |\mathcal{V}| \leq m, \\ f(\cdot|\mathcal{V}) \text{ is a PMF on } \mathcal{V} \end{array} \right\}.$$

It is clear by construction that $\mathcal{D} \subseteq \mathcal{C}$ because the conditions in $\mathcal{C}$ are those that only prevent the switcher from producing symbols that do not occur enough on average, but put no further restrictions on the switcher. So we need only show that $\mathcal{C} \subseteq \mathcal{D}$. The following gives such a proof by contradiction.

Lemma 28 (Achievability for switcher)**.** *The set relation $\mathcal{C} \subseteq \mathcal{D}$ is true.*

Proof: Without loss of generality, let $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$. Suppose $p \in \mathcal{C}$ but $p \notin \mathcal{D}$. It is clear that $\mathcal{D}$ is a convex set. Let us view the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$. Since $\mathcal{D}$ is a convex set, there is a hyperplane through p that does not intersect $\mathcal{D}$. Hence, there is a vector $(a_1, \ldots, a_{|\mathcal{X}|})$ such that $\sum_{i=1}^{|\mathcal{X}|} a_i p(i) = t$ for some real t but $t < \min_{q \in \mathcal{D}} \sum_{i=1}^{|\mathcal{X}|} a_i q(i)$. Without loss of generality, assume $a_1 \geq a_2 \geq \ldots \geq a_{|\mathcal{X}|}$ (otherwise permute symbols). Now, we will construct $f(\cdot|\mathcal{V})$ so that the resulting q has $\sum_{i=1}^{|\mathcal{X}|} a_i p(i) \geq \sum_{i=1}^{|\mathcal{X}|} a_i q(i)$, which contradicts the initial assumption. Let

$$f(i|\mathcal{V}) \triangleq \begin{cases} 1 & \text{if } i = \max(\mathcal{V}) \\ 0 & \text{else} \end{cases},$$

so for example, if $\mathcal{V} = \{1, 5, 6, 9\}$, then $f(9|\mathcal{V}) = 1$ and $f(i|\mathcal{V}) = 0$ if $i \neq 9$. Call q the distribution on $\mathcal{X}$ induced by this choice of $f(\cdot|\mathcal{V})$. Recall that $\kappa(\mathcal{V}) = P(x_l \in \mathcal{V}, 1 \leq l \leq m)$. Then, we have

$$\begin{aligned} \sum_{i=1}^{|\mathcal{X}|} a_i q(i) &= a_1 \kappa(\{1\}) + a_2[\kappa(\{1,2\}) - \kappa(\{1\})] + \\ &\quad \cdots + a_{|\mathcal{X}|} \Big[\kappa(\{1, \ldots, |\mathcal{X}|\}) - \kappa(\{1, \ldots, |\mathcal{X}| - 1\}) \Big] \end{aligned}$$

By the constraints in the definition (4.3) of $\mathcal{C}$, we have the following inequalities for p:

$$\begin{aligned} p(1) &\geq \kappa(\{1\}) = q(1) \\ p(1) + p(2) &\geq \kappa(\{1,2\}) = q(1) + q(2) \\ &\vdots \\ \sum_{i=1}^{|\mathcal{X}|-1} p(i) &\geq \kappa(\{1, \ldots, |\mathcal{X}| - 1\}) = \sum_{i=1}^{|\mathcal{X}|-1} q(i). \end{aligned}$$

Therefore, the difference of the objective is

$$\begin{aligned}
\sum_{i=1}^{|\mathcal{X}|} a_i(p(i) - q(i)) &= a_{|\mathcal{X}|}\left[\sum_{i=1}^{|\mathcal{X}|} p(i) - q(i)\right] + \\
&\quad (a_{|\mathcal{X}|-1} - a_{|\mathcal{X}|})\left[\sum_{i=1}^{|\mathcal{X}|-1} p(i) - q(i)\right] + \\
&\quad \cdots + (a_1 - a_2)\left[p(1) - q(1)\right] \\
&= \sum_{i=1}^{|\mathcal{X}|-1} (a_i - a_{i+1})\left[\sum_{j=1}^{i} p(j) - \sum_{j=1}^{i} q(j)\right] \\
&\geq 0.
\end{aligned}$$

The last step is true because of the monotonicity in the a_i and the inequalities we derived earlier. Therefore, we see that $\sum_{i=1}^{|\mathcal{X}|} a_i p(i) \geq \sum_{i=1}^{|\mathcal{X}|} a_i q(i)$ for the p we had chosen at the beginning of the proof. This contradicts the assumption that $\sum_{i=1}^{|\mathcal{X}|} a_i p(i) < \min_{q\in\mathcal{D}} \sum_{i=1}^{|\mathcal{X}|} a_i q(i)$, therefore it must be that $\mathcal{C} \subseteq \mathcal{D}$.

C.2 Adversarial switching with noisy observations

Theorem 10 of Section 4.4: For the AVS problem of Fig. 4.2, where the adversary has access to the states either with 1-step lookahead or full lookahead,

$$R(D) = \max_{p\in\mathcal{D}_{states}} R(p, D),$$

where

$$\mathcal{D}_{states} = \left\{ p : \begin{array}{c} p(\cdot) = \sum_{t\in\mathcal{T}} \alpha(t) f(\cdot|t) \\ f(\cdot|t) \in \overline{\mathcal{G}}(t), \forall\, t \in \mathcal{T} \end{array} \right\}.$$

Proof: It is clear that $R(D) \geq \max_{p\in\mathcal{D}_{states}} R(p, D)$ because the switcher can select distributions $f(\cdot|t) \in \overline{\mathcal{G}}(t)$ for all $t \in \mathcal{T}$ and upon observing a state t, the switcher can randomly select the switch position according to the convex combination that yields $f(\cdot|t)$. With this strategy, the AVS is simply an IID source with distribution $p(\cdot) = \sum_t \alpha(t) f(\cdot|t)$. Hence, $R(D) \geq \max_{p\in\mathcal{D}_{states}} R(p, D)$.

We will now show that $R(D) \leq \max_{p\in\mathcal{D}_{states}} R(p, D)$. This can be done in the same way as in Appendix C.1. We can use the type covering lemma to cover sequences with types in or very near $\mathcal{D}_{states}$ and then we need only show that the probability of $\mathbf{x}^n$ having a type ϵ-far from $\mathcal{D}_{states}$ goes to 0 with block length n.

Lemma 29. *Let $p_{\mathbf{x}^n}$ be the type of $\mathbf{x}^n$ and for $\epsilon > 0$ let $\mathcal{D}_{states,\epsilon}$ be the set of $p \in \mathcal{P}(\mathcal{X})$ with $\mathcal{L}_1$ distance at most ϵ from a distribution in $\mathcal{D}_{states}$. Then, for $\epsilon > 0$,*

$$P(p_{\mathbf{x}^n} \notin \mathcal{D}_{states,\epsilon}) \leq 4|\mathcal{T}||\mathcal{X}| \exp(-n\xi(\epsilon)),$$

where $\xi(\epsilon) > 0$ for all $\epsilon > 0$. So for large n, $p_{\mathbf{x}^n}$ is in $\mathcal{D}_{states,\epsilon}$ with high probability.

Proof: Let $\mathbf{t}^n$ be the n-length vector of the observed states. We assume that the switcher has advance knowledge of all these states before choosing the switch positions. First, we show that with high probability, the states that are observed are strongly typical. Let $N(t|\mathbf{t}^n)$ be the count of occurrence of $t \in \mathcal{T}$ in the vector $\mathbf{t}^n$. Fix a $\delta > 0$ and for $t \in \mathcal{T}$, define the event

$$A_\delta^t = \left\{ \left| \frac{N(t|\mathbf{t}^n)}{n} - \alpha(t) \right| > \delta \right\}. \tag{C.1}$$

Since $N(t|\mathbf{t}^n) = \sum_{i=1}^n \mathbf{1}(t_i = t)$ and each term in the sum is an IID Bernoulli variable with probability of 1 equal to $\alpha(t)$, we have by Hoeffding's tail inequality [57],

$$P(A_\delta^t) \leq 2\exp(-2n\delta^2).$$

Next, we need to show that the substrings output by the AVS at the times when the state is t have a type in or very near $\overline{\mathcal{G}}(t)$. This will be done by a martingale argument similar to that given in Lemma 3 of [4]. Let $\mathbf{t}^\infty$ denote the infinite state sequence $(t_1, t_2, \ldots)$ and let $\mathcal{F}_0 = \sigma(\mathbf{t}^\infty)$ be the sigma field generated by the states $\mathbf{t}^\infty$. For $i = 1, 2, \ldots$, let $\mathcal{F}_i = \sigma(\mathbf{t}^\infty, \mathbf{s}^i, \mathbf{x}_1^i, \ldots, \mathbf{x}_m^i)$. Note that $\{\mathcal{F}_i\}_{i=0}^\infty$ is a filtration and for each i, x_i is included in $\mathcal{F}_i$ trivially because $x_i = x_{s_i,i}$.

Let C_i be the $|\mathcal{X}|$-dimensional unit vector with a 1 in the position of x_i. That is, $C_i(x) = \mathbf{1}(x_i = x)$ for each $x \in \mathcal{X}$. Define T_i to be

$$T_i = C_i - \mathbb{E}[C_i|\mathcal{F}_{i-1}]$$

and let $S_0 = 0$. For $k \geq 1$,

$$S_k = \sum_{i=1}^k T_i.$$

We claim that $S_k, k \geq 1$ is a martingale[1] with respect to the filtration $\{\mathcal{F}_i\}$ defined previously. To see this, note that $\mathbb{E}[|S_k|] < \infty$ for all k since S_k is bounded (not uniformly). Also, $S_k \in \mathcal{F}_k$ because $T_i \in \mathcal{F}_i$ for each i. Finally,

$$\begin{aligned}
\mathbb{E}[S_{k+1}|\mathcal{F}_k] &= \mathbb{E}[T_{k+1} + S_k|\mathcal{F}_k] \\
&= \mathbb{E}[T_{k+1}|\mathcal{F}_k] + S_k \\
&= \mathbb{E}[C_{k+1} - \mathbb{E}[C_{k+1}|\mathcal{F}_k]|\mathcal{F}_k] + S_k \\
&= \mathbb{E}[C_{k+1}|\mathcal{F}_k] - \mathbb{E}[C_{k+1}|\mathcal{F}_k] + S_k \\
&= S_k.
\end{aligned}$$

[1] S_k is a vector, so we show that each component of the vector is a martingale. For ease of notation, we drop the dependence on the component of the vector until it is explicitly needed.

Now, define for each $t \in \mathcal{T}$,

$$T_i^t = T_i \cdot \mathbf{1}(t_i = t)$$

and analogously,

$$S_k^t = \sum_{i=1}^{k} T_i^t.$$

It can be easily verified that S_k^t is a martingale with respect to $\mathcal{F}_i$ for each $t \in \mathcal{T}$. Expanding, we also see that

$$\begin{aligned} \frac{1}{N(t|\mathbf{t}^n)} S_n^t &= \frac{1}{N(t|\mathbf{t}^n)} \sum_{i=1}^{n} T_i \mathbf{1}(t_i = t) \\ &= \frac{1}{N(t|\mathbf{t}^n)} \sum_{i:\ t_i = t} C_i - \\ &\quad \frac{1}{N(t|\mathbf{t}^n)} \sum_{i:\ t_i = t} \mathbb{E}[C_i|\mathcal{F}_{i-1}]. \end{aligned} \tag{C.2}$$

The first term in the difference above is the type of the output of the AVS during times when the state is t. For any i such that $t_i = t$,

$$\mathbb{E}[C_i|\mathcal{F}_{i-1}] = \sum_{l=1}^{m} P(l|\mathcal{F}_{i-1}) p_l(\cdot|t) \in \overline{\mathcal{G}}(t).$$

In the above, $P(l|\mathcal{F}_{i-1})$ represents the switcher's possibly random strategy because the switcher chooses the switch position at time i with knowledge of events in $\mathcal{F}_{i-1}$. The symbol generator's outputs, conditioned on the state at the time are independent of all other random variables, so $\sum_{l=1}^{m} P(l|\mathcal{F}_{i-1}) p_l(\cdot|t)$ is the probability distribution of the output at time i conditioned on $\mathcal{F}_{i-1}$.

Thus, the second term in the difference of (C.2) is in $\overline{\mathcal{G}}(t)$ because it is the average of $N(t|\mathbf{t}^n)$ terms in $\overline{\mathcal{G}}(t)$ and $\overline{\mathcal{G}}(t)$ is a convex set. Therefore, $S_n^t/N(t|\mathbf{t}^n)$ measures the difference between the type of symbols output at times when the state is t and some distribution guaranteed to be in $\overline{\mathcal{G}}(t)$.

Let $p_{\mathbf{x}^n}$ be the empirical type of the string $\mathbf{x}^n$, and let $p_{\mathbf{x}^n}^t$ be the empirical type of the sub-string of $\mathbf{x}^n$ corresponding to the times i when $t_i = t$. Then,

$$p_{\mathbf{x}^n} = \sum_{t \in \mathcal{T}} \frac{N(t|\mathbf{t}^n)}{n} p_{\mathbf{x}^n}^t.$$

Let $\overline{\mathcal{G}}(t)_\epsilon$ be the set of distributions at most ϵ in $\mathcal{L}_1$ distance from a distribution in $\overline{\mathcal{G}}(t)$. Recall that for $|\mathcal{X}|$ dimensional vectors, $\|p - q\|_\infty < \epsilon/|\mathcal{X}|$ implies $\|p - q\|_1 < \epsilon$. Hence, we have

$$\begin{aligned} P\left(\bigcup_{t\in\mathcal{T}}\{p^t_{\mathbf{x}^n}\notin\overline{\mathcal{G}}(t)_\epsilon\}\right) &\leq \sum_{t\in\mathcal{T}} P\left(\bigcup_{x\in\mathcal{X}}\left\{\left|\frac{1}{N(t|\mathbf{t}^n)}S^t_n(x)\right| > \frac{\epsilon}{|\mathcal{X}|}\right\}\right) \\ &\leq \sum_t\sum_x P\left(\left|\frac{1}{N(t|\mathbf{t}^n)}S^t_n(x)\right| > \frac{\epsilon}{|\mathcal{X}|}\right). \qquad \text{(C.3)}\end{aligned}$$

Let $(A^t_\delta)^c$ denote the complement of the event A^t_δ. So, for every (t,x) we have

$$\begin{aligned} P\left(\left|\frac{1}{N(t|\mathbf{t}^n)}S^t_n(x)\right| > \frac{\epsilon}{|\mathcal{X}|}\right) &\leq P(A^t_\delta) + P\left(\left|\frac{1}{N(t|\mathbf{t}^n)}S^t_n(x)\right| > \frac{\epsilon}{|\mathcal{X}|}, (A^t_\delta)^c\right) \\ &\leq 2\exp(-2n\delta^2) + \\ &\qquad P\left(\left|\frac{1}{N(t|\mathbf{t}^n)}S^t_n(x)\right| > \frac{\epsilon}{|\mathcal{X}|}, (A^t_\delta)^c\right).\end{aligned}$$

In the event of $(A^t_\delta)^c$, we have $N(t|\mathbf{t}^n) \geq n(\alpha(t)-\delta)$, so

$$\begin{aligned} P\left(\left|\frac{1}{N(t|\mathbf{t}^n)}S^t_n(x)\right| > \frac{\epsilon}{|\mathcal{X}|}, (A^t_\delta)^c\right) &\leq P\left(|S^t_n(x)| > n(\alpha(t)-\delta)\frac{\epsilon}{|\mathcal{X}|}, (A^t_\delta)^c\right) \\ &\leq P\left(|S^t_n(x)| > n(\alpha(t)-\delta)\frac{\epsilon}{|\mathcal{X}|}\right).\end{aligned}$$

$S^t_k(x)$ is a martingale with bounded differences since $|S^t_{k+1}(x) - S^t_k(x)| = |T^t_{k+1}(x)| \leq 1$. Hence, we can apply Azuma's inequality [58] to get

$$P\left(\left|\frac{1}{N(t|\mathbf{t}^n)}S^t_n(x)\right| > \frac{\epsilon}{|\mathcal{X}|}, (A^t_\delta)^c\right) \leq 2\exp\left(-n\frac{(\alpha(t)-\delta)^2\epsilon^2}{2|\mathcal{X}|^2}\right). \qquad \text{(C.4)}$$

Plugging this back into (C.3),

$$\begin{aligned} P\left(\bigcup_{t\in\mathcal{T}}\{p^t_{\mathbf{x}^n}\notin\overline{\mathcal{G}}(t)_\epsilon\}\right) &\leq 2|\mathcal{T}||\mathcal{X}|\Bigg(\exp(-2n\delta^2) + \\ &\qquad \exp\left(-n\frac{(\alpha_*-\delta)^2\epsilon^2}{2|\mathcal{X}|^2}\right)\Bigg) \\ &\leq 4|\mathcal{X}||\mathcal{T}|\exp(-n\xi(\epsilon,\delta))\end{aligned}$$

where

$$\begin{aligned} \xi(\epsilon,\delta) &= \min\left\{2\delta^2, \frac{(\alpha_*-\delta)^2\epsilon^2}{2|\mathcal{X}|^2}\right\} \\ \alpha_* &\triangleq \min_{t\in\mathcal{T}}\alpha(t).\end{aligned}$$

We assume without loss of generality that $\alpha_* > 0$ since $\mathcal{T}$ is finite. We will soon need that $\delta \le \epsilon/|\mathcal{T}|$, so let

$$\widetilde{\xi}(\epsilon) = \max_{0<\delta<\min\{\epsilon/|\mathcal{T}|,\alpha_*\}} \xi(\epsilon, \delta)$$

and note that it is always positive provided $\epsilon > 0$, since $\xi(\epsilon, \delta) > 0$ whenever $\delta \in (0, \alpha_*)$. Hence,

$$P\left(\bigcup_{t\in\mathcal{T}} \{p^t_{\mathbf{x}^n} \notin \overline{\mathcal{G}}(t)_\epsilon\}\right) \le 4|\mathcal{X}||\mathcal{T}| \exp(-n\widetilde{\xi}(\epsilon)).$$

We have shown that with probability at least $1 - 4|\mathcal{X}||\mathcal{T}| \exp(-n\widetilde{\xi}(\epsilon))$, for each $t \in \mathcal{T}$ there is some $p^t \in \overline{\mathcal{G}}(t)$ such that $\|p^t_{\mathbf{x}^n} - p^t\|_1 \le \epsilon$ and $(A^t_{\epsilon/|\mathcal{T}|})^c$ occurs. Let

$$p = \sum_{t\in\mathcal{T}} \alpha(t)p^t.$$

By construction, $p \in \mathcal{D}_{states}$. To finish, we show that $\|p_{\mathbf{x}^n} - p\|_1 \le 2\epsilon$.

$$\begin{aligned}
\|p_{\mathbf{x}^n} - p\|_1 &= \sum_{x\in\mathcal{X}} |p_{\mathbf{x}^n}(x) - p(x)| \\
&= \sum_x \left| \sum_{t\in\mathcal{T}} \frac{N(t|\mathbf{t}^n)}{n} p^t_{\mathbf{x}^n}(x) - \alpha(t)p^t(x)\right| \\
&\le \sum_t \sum_x \left|\frac{N(t|\mathbf{t}^n)}{n} p^t_{\mathbf{x}^n}(x) - \alpha(t)p^t(x)\right| \\
&= \sum_t \alpha(t) \sum_x \left|\frac{N(t|\mathbf{t}^n)}{n\alpha(t)} p^t_{\mathbf{x}^n}(x) - p^t(x)\right| \\
&\le \sum_t \alpha(t) \sum_x |p^t_{\mathbf{x}^n}(x) - p^t(x)| + \left|\frac{N(t|\mathbf{t}^n)}{n\alpha(t)} - 1\right| p^t_{\mathbf{x}^n}(x).
\end{aligned}$$

From (C.1), we are assumed to be in the event that

$$\left|\frac{N(t|\mathbf{t}^n)}{n\alpha(t)} - 1\right| \le \frac{\delta}{\alpha(t)}$$

Hence,

$$\begin{aligned}
\|p_{\mathbf{x}^n} - p\|_1 &\le \sum_t \alpha(t)\left(\epsilon + \frac{\delta}{\alpha(t)}\right) \\
&= \epsilon + |\mathcal{T}|\delta \le 2\epsilon.
\end{aligned}$$

We have proved $P(p_{\mathbf{x}^n} \notin \mathcal{D}_{states,2\epsilon}) \le 4|\mathcal{X}||\mathcal{T}| \exp(-n\widetilde{\xi}(\epsilon))$, so we arrive at the conclusion of the lemma by letting $\xi(\epsilon) = \widetilde{\xi}(\epsilon/2)$.

C.3 Uniform continuity of $R(p, D)$

Lemma 9 of Section 4.7.1:

Let $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$ be a distortion function. $\widetilde{d}$ is the minimum nonzero distortion from (4.1). Also, assume that for each $x \in \mathcal{X}$, there is an $\hat{x}_0(x) \in \widehat{\mathcal{X}}$ such that $d(x, \hat{x}_0(x)) = 0$. Then, for $p, q \in \mathcal{P}(\mathcal{X})$ with $\|p - q\|_1 \leq \frac{\widetilde{d}}{4d^*}$, for any $D \geq 0$,

$$|R(p, D) - R(q, D)| \leq \frac{7d^*}{\widetilde{d}} \|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1}.$$

Proof: Let $W^*_{p,D} \in \text{argmin}_{W \in \mathcal{W}(p,D)} I(p, W)$. Then

$$|R(p, D) - R(q, D)| = |I(p, W^*_{p,D}) - I(q, W^*_{q,D})|.$$

Consider $d(p, W^*_{q,D})$, the distortion of source p across q's distortion D achieving channel.

$$\begin{aligned} d(p, W^*_{q,D}) &\leq d(q, W^*_{q,D}) + |d(p, W^*_{q,D}) - d(q, W^*_{q,D})| \\ &\leq D + \|p - q\|_1 d^*. \end{aligned}$$

By definition, $W^*_{q,D}$ is in $\mathcal{W}(p, d(p, W^*_{q,D}))$, so $R(p, d(p, W^*_{q,D})) \leq I(p, W^*_{q,D})$.

$$\begin{aligned} R(p, d(p, W^*_{q,D})) &\leq I(p, W^*_{q,D}) \\ &\leq |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + I(q, W^*_{q,D}) \\ &= |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + R(q, D) \qquad \text{(C.5)} \end{aligned}$$

Expanding mutual informations yields

$$\begin{aligned} |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| \leq\; & |H(p) - H(q)| + |H(pW^*_{q,D}) - H(qW^*_{q,D})| + \\ & |H(p, W^*_{q,D}) - H(q, W^*_{q,D})|. \end{aligned}$$

Above, for a distribution p on $\mathcal{X}$ and channel W from $\mathcal{X}$ to $\widehat{\mathcal{X}}$, $H(pW)$ denotes the entropy of a distribution on $\widehat{\mathcal{X}}$ with probabilities $(pW)(\widehat{x}) = \sum_x p(x)W(\widehat{x}|x)$. $H(p, W)$ denotes the entropy of the joint source on $\mathcal{X} \times \widehat{\mathcal{X}}$ with probabilities $(p, W)(x, \widehat{x}) = p(x)W(\widehat{x}|x)$. It is straightforward to verify that $\|pW - qW\|_1 \leq \|p - q\|_1$ and $\|(p, W) - (q, W)\|_1 \leq \|p - q\|_1$. So using Lemma 8 three times, we have

$$\begin{aligned} |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| &= |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| \\ &\leq 3\|p - q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p - q\|_1}. \end{aligned}$$

Now, we have seen $d(p, W^*_{q,D}) \leq D + d^*\|p - q\|_1$. We will use the uniform continuity of $R(p, D)$ in D to bound $|R(p, D) - R(p, D + d^*\|p - q\|_1)|$. This will give an upper bound on $R(p, D) - R(q, D)$ as seen through (C.5), namely,

$$\begin{aligned} R(p, D) - R(q, D) &\leq |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + \\ &\qquad R(p, D) - R(p, d(p, W^*_{q,D})) \\ &\leq |I(p, W^*_{q,D}) - I(q, W^*_{q,D})| + \\ &\qquad R(p, D) - R(p, D + d^*\|p - q\|_1), \end{aligned} \tag{C.6}$$

where the last step follows because $R(p, D)$ is monotonically decreasing in D. For a fixed p, the rate-distortion function in D is convex-$\cup$ and decreasing and so has steepest descent at $D = 0$. Therefore, for any $0 \leq D_1, D_2 \leq d^*$,

$$|R(p, D_1) - R(p, D_2)| \leq |R(p, 0) - R(p, |D_2 - D_1|)|.$$

Hence, we can restrict our attention to continuity of $R(p, D)$ around $D = 0$. By assumption, $\mathcal{W}(p, 0) \neq \emptyset \ \forall p \in \mathcal{P}(\mathcal{X})$. Now consider an arbitrary $D > 0$, and let $W \in \mathcal{W}(p, D)$. We will show that there is some $W_0 \in \mathcal{W}(p, 0)$ that is close to W in an $\mathcal{L}_1$-like sense (relative to the distribution p). Since $W \in \mathcal{W}(p, D)$, we have by definition

$$\begin{aligned} D &\geq \sum_x p(x) \sum_{\widehat{x}} W(\widehat{x}|x) d(x, \widehat{x}) \\ &= \sum_x p(x) \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x) d(x, \widehat{x}) \\ &\geq \widetilde{d} \sum_x p(x) \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x). \end{aligned} \tag{C.7}$$

Now, we will construct a channel in $\mathcal{W}(p, 0)$, denoted W_0. First, for each $x, \widehat{x}$ such that $d(x, \widehat{x}) = 0$, let $V(\widehat{x}|x) = W(\widehat{x}|x)$. For all other $(x, \widehat{x})$, set $V(\widehat{x}|x) = 0$. Note that V is not a channel matrix if $W \notin \mathcal{W}(p, 0)$ since it is missing some probability mass. To create W_0, for each x, we redistribute the missing mass from $V(\cdot|x)$ to the pairs $(x, \widehat{x})$ with $d(x, \widehat{x}) = 0$. Namely, for $(x, \widehat{x})$ with $d(x, \widehat{x}) = 0$, we define

$$W_0(\widehat{x}|x) \ = \ V(\widehat{x}|x) + \frac{\sum_{\hat{x}':\ d(x,\hat{x}')>0} W(\hat{x}'|x)}{|\{\hat{x}' :\ d(x, \hat{x}') = 0\}|}.$$

For all $(x, \widehat{x})$ with $d(x, \widehat{x}) > 0$, define $W_0(\widehat{x}|x) = 0$. So, W_0 is a valid channel in $\mathcal{W}(p, 0)$.

Now for a fixed $x \in \mathcal{X}$,

$$\begin{aligned}
\sum_{\widehat{x}} |W(\widehat{x}|x) - W_0(\widehat{x}|x)| &= \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x) + \sum_{\widehat{x}:\ d(x,\widehat{x})=0} |W(\widehat{x}|x) - W_0(\widehat{x}|x)| \\
&= \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x) + \sum_{\widehat{x}:\ d(x,\widehat{x})=0} \left| \frac{\sum_{\hat{x}':\ d(x,\hat{x}')>0} W(\hat{x}'|x)}{|\{\hat{x}':\ d(x,\hat{x}')=0\}|} \right| \\
&= 2 \sum_{\widehat{x}:\ d(x,\widehat{x})>0} W(\widehat{x}|x).
\end{aligned}$$

Therefore, using (C.7)

$$\sum_x p(x) \sum_{\widehat{x}} |W(\widehat{x}|x) - W_0(\widehat{x}|x)| \ \leq \ \frac{2D}{\widetilde{d}}.$$

So, for $W = W^*_{p,D}$, there is a $W_0 \in \mathcal{W}(p,0)$ with the above 'modified $\mathcal{L}_1$ distance' with respect to p between W and W_0 being less than $2D/\widetilde{d}$. Going back to the bound on $|R(p,0)-R(p,D)|$,

$$\begin{aligned}
|R(p,0) - R(p,D)| &= \min_{W \in \mathcal{W}(p,0)} I(p,W) - I(p,W^*_{p,D}) \\
&\leq I(p,W_0) - I(p,W^*_{p,D}) \\
&\leq |H(pW_0) - H(pW^*_{p,D})| + |H(p,W_0) - H(p,W^*_{p,D})|.
\end{aligned}$$

It can be easily verified that $\|pW_0 - pW^*_{p,D}\|_1$ is at most $2D/\widetilde{d}$. Similarly, $\|(p,W_0) - (p,W^*_{p,D})\|_1 \leq 2D/\widetilde{d}$.

Now, assuming $D \leq \widetilde{d}/4$, we can again invoke Lemma 8 to get

$$\begin{aligned}
|R(p,0) - R(p,D)| &\leq \frac{2D}{\widetilde{d}} \ln \frac{\widetilde{d}|\mathcal{X}|}{2D} + \frac{2D}{\widetilde{d}} \ln \frac{\widetilde{d}|\mathcal{X}||\widehat{\mathcal{X}}|}{2D} \\
&\leq \frac{4D}{\widetilde{d}} \ln \frac{\widetilde{d}|\mathcal{X}||\widehat{\mathcal{X}}|}{2D}. \qquad \text{(C.8)}
\end{aligned}$$

Going back to (C.6), we see that if $\|p-q\|_1 \leq \frac{\widetilde{d}}{4d^*}$,

$$\begin{aligned}
|R(p, D + d^*\|p-q\|_1) - R(p,D)| &\leq \frac{4d^*\|p-q\|_1}{\widetilde{d}} \ln \frac{\widetilde{d}|\mathcal{X}||\widehat{\mathcal{X}}|}{2d^*\|p-q\|_1} \\
&\leq \frac{4d^*\|p-q\|_1}{\widetilde{d}} \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}.
\end{aligned}$$

The last step follows because $\widetilde{d}/d^* \leq 1$. Substituting into (C.6) gives

$$\begin{aligned} R(p,D) - R(q,D) &\leq 3\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1} + 4\frac{d^*}{\widetilde{d}}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1} \\ &\leq \frac{7d^*}{\widetilde{d}}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}. \end{aligned}$$

Finally, this bound holds uniformly on p and q as long as the condition on $\|p-q\|_1$ is satisfied. Therefore, we can interchange p and q to get the other side of the inequality

$$R(q,D) - R(p,D) \leq \frac{7d^*}{\widetilde{d}}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}.$$

C.4 Proof of Lemma 10

We now assume $d : \mathcal{X} \times \widehat{\mathcal{X}} \to [0, d^*]$ to be arbitrary. However, we let

$$d_0(x,\widehat{x}) = d(x,\widehat{x}) - \min_{\widetilde{x}\in\widehat{\mathcal{X}}} d(x,\widetilde{x})$$

so that Lemma 9 applies to d_0. Let $R_0(p,D)$ be the IID rate-distortion function for $p \in \mathcal{P}(\mathcal{X})$ at distortion D with respect to distortion measure $d_0(x,\widehat{x})$. By definition, $R(p,D)$ is the IID rate-distortion function for p with respect to distortion measure $d(x,\widehat{x})$. From Problem 13.4 of [21], for any $D \geq D_{\min}(p)$,

$$R(p,D) = R_0(p, D - D_{\min}(p)).$$

Hence, for $p,q \in \mathcal{P}(\mathcal{X})$, $D \geq \max(D_{\min}(p), D_{\min}(q))$,

$$\begin{aligned} |R(p,D) - R(q,D)| &= |R_0(p, D - D_{\min}(p)) - R_0(q, D - D_{\min}(q)| \\ &\leq |R_0(p, D - D_{\min}(p)) - R_0(p, D - D_{\min}(q))| + \\ &\quad |R_0(p, D - D_{\min}(q)) - R_0(q, D - D_{\min}(q))|. \end{aligned} \tag{C.9}$$

Now, we note that $|D_{\min}(p) - D_{\min}(q)| \leq d^*\|p-q\|_1$. The first term of (C.9) can be bounded using (C.8) and the second term of (C.9) can be bounded using Lemma 9. The first term can be bounded if $\|p-q\|_1 \leq \widetilde{d}_0/4d^*$ and the second can be bounded if $\|p-q\|_1 \leq \widetilde{d}_0/4d_0^*$. Since $d_0^* \leq d^*$, we only require $\|p-q\|_1 \leq \widetilde{d}_0/4d^*$.

$$\begin{aligned} |R(p,D) - R(q,D)| &\leq \frac{4d^*}{\widetilde{d}_0}\|p-q\|_1 \ln \frac{\widetilde{d}_0|\mathcal{X}||\widehat{\mathcal{X}}|}{2d^*\|p-q\|_1} + \\ &\qquad \frac{7d_0^*}{\widetilde{d}_0}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1} \\ &\leq \frac{11d^*}{\widetilde{d}_0}\|p-q\|_1 \ln \frac{|\mathcal{X}||\widehat{\mathcal{X}}|}{\|p-q\|_1}. \end{aligned}$$

Bibliography

[1] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[2] ——, “The zero-error capacity of a noisy channel,” *IRE Trans. Inform. Theory*, vol. 2, pp. 8–19, 1956.

[3] E. Haroutunian, “Lower bound for error probability in channels with feedback,” *Problemy Peredachi Informatsii*, pp. 36–44, 1977.

[4] T. Berger, “The source coding game,” *IEEE Trans. Inf. Theory*, vol. 17, pp. 71–76, Jan. 1971.

[5] C. Shannon, “Channels with side information at the transmitter,” *IBM J. Res. Devel.*, vol. 2, pp. 289–293, Oct. 1958.

[6] S. Gelfand and M. Pinsker, “Coding for channel with random parameters,” *Probl. Pered. Inform. (Probl. Inf. Transm.)*, vol. 9, pp. 19–31, 1980.

[7] C. Heegard and A. Gamal, “On the capacity of computer memory with defects,” *Information Theory, IEEE Transactions on*, vol. 29, no. 5, pp. 731 – 739, sep 1983.

[8] S. Jafar, “Capacity with causal and noncausal side information: A unified view,” *Information Theory, IEEE Transactions on*, vol. 52, no. 12, pp. 5468 –5474, dec. 2006.

[9] M. V. Burnashev, “Data transmission over a discrete channel with feedback,” *Problemy Peredachi Informatsii*, vol. 12, p. 1030, 1976.

[10] B. Nakiboglu, “Exponential bounds on error probability with feedback,” Ph.D. dissertation, Massachussetts Institute of Technology, Cambridge, MA, 2011.

[11] Y. Polyanskiy, “Channel coding: non-asymptotic fundamental limits,” Ph.D. dissertation, Princeton University, Princeton, NJ, 2010.

[12] A. Feinstein, “A new basic theorem of information theory,” *IRE Trans. Inform. Theory*, vol. 4, pp. 2–22, 1954.

[13] ——, “Error bounds in noisy channels without memory,” *IRE Trans. Inform. Theory*, vol. 1, pp. 13–14, 1955.

[14] P. Elias, “Coding for nosiy channels,” in *IRE Natl. Conv. Rec.*, 1955, pp. 37–46.

[15] R. M. Fano, *Transmission of Information.* Cambridge, MA and New York, NY: MIT Press and Wiley, 1961.

[16] R. G. Gallager, “A simple derivation of the coding theorem and some applications,” vol. 11, pp. 3–18, Jan. 1965.

[17] R. Gallager, *Information Theory and Reliable Communication.* New York, NY: John Wiley and Sons, 1971.

[18] C. Shannon, R. Gallager, and E. Berlekamp, “Lower bounds to error probability for coding on discrete memoryless channels. I,” *Information and Control*, vol. 10, pp. 65–103, 1967.

[19] E. A. Haroutunian, “Estimates of exponents in error probability for a semicontinuous memoryless channel,” *Problemy Peredachi Informatsii*, pp. 37–48, 1968.

[20] R. Blahut, “Hypothesis testing and information theory,” *Information Theory, IEEE Transactions on*, vol. 20, no. 4, pp. 405 – 417, Jul. 1974.

[21] T. Cover and J. Thomas, *Elements of Information Theory.* New York, NY: John Wiley and Sons, 1991.

[22] C. Shannon, R. Gallager, and E. Berlekamp, “Lower bounds to error probability for coding on discrete memoryless channels. II,” *Information and Control*, vol. 10, pp. 522–552, 1967.

[23] K. S. Zigangirov, “Upper bounds for the error probability for channels with feedback,” *Problems of Information Transmission*, vol. 6, pp. 87–92, 1970.

[24] A. G. D’yachkov, “Upper bounds on the error probability for discrete memoryless channels with feedback,” *Problems of Information Transmission*, vol. 11, pp. 13–28, 1975.

[25] M. V. Burnashev, “On the reliability function of a binary symmetrical channel with feedback,” *Problems of Information Transmission*, vol. 24, pp. 3–10, 1988.

[26] R. Dobrushin, “Asymptotic estimate of error probability in message transmission through memoryless channels using feedback,” *Problemy Kibernetiki [Russian]*, pp. 161–168, 1962.

[27] E. Haroutunian, "Estimation of optimal error probability in information transmission through channels with feedback," in *Proceedings of the Fourth All-Union Symposium on Redundancy in Infomation Systems [In Russian]*, Leningrad, USSR, 1970.

[28] ——, "Lower bound of error probability in data transmission through channels with feedback," in *Proceedings of the Second International Symposium on Information Theory [In Russian]*, Tsakhkadzor, Armenia, 1971, pp. 12–13.

[29] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, pp. 2307–2359, May 2010.

[30] A. Y. Sheverdyaev, "Lower bound for error probability in a discrete memoryless channel with feedback," *Problemy Peredachi Informatsii*, pp. 5–15, 1982.

[31] H. Palaiyanur and A. Sahai, "An upper bound for the block coding error exponent with delayed feedback," in *Proc. Int. Symp. Inform. Theory*, Austin, TX, Jun. 2010.

[32] I. Csiszar, "The method of types [information theory]," *Information Theory, IEEE Transactions on*, vol. 44, no. 6, pp. 2505 –2523, Oct. 1998.

[33] B. Nakiboglu, Personal Communication, 2008-2009.

[34] U. Augustin, "Noisy channels," Habilitationschrift at Universität-Nürenberg. Submitted to Springer Lecture Notes, 1978.

[35] T. S. Han, *Information-Spectrum Methods in Information Theory*, 1st ed. Heidelberg, Germany: Springer-Verlag, 2002.

[36] G. Como and B. Nakiboglu, "A lower bound on the error probability for block-codes with finite memory feedback," in *Proc. Int. Symp. Inform. Theory*, Austin, TX, Jun. 2010.

[37] A. Sahai, "Why do block length and delay behave differently if feedback is present," *IEEE Trans. Inf. Theory*, vol. 54, pp. 1860–1886, May 2008.

[38] A. Sahai and P. Grover, "The price of certainty: waterslide curves and the gap to capacity," 2008. [Online]. Available: http://arxiv.org/abs/0801.0352

[39] A. Sahai and S. Draper, "The hallucination bound for the bsc," in *Proc. Int. Symp. Inform. Theory*, Toronto, Canada, Jul. 2008.

[40] 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," vol. 52, no. 8, pp. 3369–3395, Aug. 2006.

[41] R. Bajcsy, "Active perception," *Proceedings of the IEEE*, vol. 76, no. 8, pp. 966–1005, Aug. 1988.

[42] T. Berger and J. Gibson, "Lossy source coding," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2693–2723, Oct 1998.

[43] J. Ziv, "Distortion-rate theory for individual sequences," *IEEE Trans. Inf. Theory*, vol. 26, no. 2, pp. 137–143, Mar 1980.

[44] E.-H. Yang and J. Kieffer, "Simple universal lossy data compression schemes derived from the lempel-ziv algorithm," *IEEE Trans. Inf. Theory*, vol. 42, pp. 239–245, 1996.

[45] R. Ahlswede, "Extremal properties of rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. 36, pp. 166–171, Jan. 1990.

[46] H. Palaiyanur, C. Chang, and A. Sahai, "The source coding game with a cheating switcher," To appear in IEEE Transactions on Information Theory, 2011.

[47] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. New York, NY: Academic Press, 1997.

[48] C. Shannon, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Natl. Conv. Rec.*, 1959, pp. 142–163.

[49] J. Wolfowitz, "Approximation with a fidelity criterion," in *5th Berkeley Symp. on Math. Stat. and Prob.*, vol. 1. Berkeley, California: University of California, Press, 1967, pp. 565–573.

[50] D. Sakrison, "The rate-distortion function for a class of sources," *Information and Control*, vol. 15, pp. 165–195, Mar. 1969.

[51] M. Harrison and I. Kontoyiannis, "Estimation of the rate-distortion function," 2007. [Online]. Available: http://arxiv.org/abs/cs/0702018v1

[52] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. L. Weinberger, "Inequalities for the l_1 deviation of the empirical distribution," Hewlett-Packard Labs, Tech. Rep., 2003. [Online]. Available: http://www.hpl.hp.com/techreports/2003/HPL-2003-97R1.html

[53] D. Neuhoff and R. K. Gilbert, "Causal source codes," *IEEE Trans. Inf. Theory*, vol. 28, pp. 701–713, Sep. 1982.

[54] R. Dobrushin, "Unified methods for the transmission of information: The general case," *Sov. Math.*, vol. 4, pp. 284–292, 1963.

[55] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. Inf. Theory*, vol. 44, Oct. 1998.

[56] G. Hardy, J. Littlewood, and G. Pólya, *Inequalities*, 2nd ed. Cambridge, UK: Cambridge University Press, 1952.

[57] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, Mar 1963.

[58] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Math. Journal*, vol. 19, pp. 357 – 367, 1967.