Robust and adaptive communication under uncertain interference

Anand D. Sarwate

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2008-86
http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-86.html

July 17, 2008

Robust and adaptive communication under uncertain interference

by

Anand Dilip Sarwate

B.S. (Massachusetts Institute of Technology) 2002
B.S. (Massachusetts Institute of Technology) 2002
M.S. (University of California, Berkeley) 2005

A dissertation submitted in partial satisfaction

of the requirements for the degree of

Doctor of Philosophy

in

Engineering—Electrical Engineering and Computer Sciences

and the Designated Emphasis

in

Communication, Computation, and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael Gastpar, Chair
Professor Venkatachalam Anantharam
Professor David Blackwell

Fall 2008

The dissertation of Anand Dilip Sarwate is approved:

Chair Date

Date

Date

University of California, Berkeley

Fall 2008

Robust and adaptive communication under uncertain interference

by

Anand Dilip Sarwate

Abstract

Robust and adaptive communication under uncertain interference

by

Anand Dilip Sarwate

Doctor of Philosophy in Engineering—Electrical Engineering and Computer Sciences

and the Designated Emphasis

in

Communication, Computation, and Statistics

University of California, Berkeley

Professor Michael Gastpar, Chair

In the future, wireless communication systems will play an increasingly integral role in society. Cutting-edge application areas such as cognitive radio, ad-hoc networks, and sensor networks are changing the way we think about wireless services. The demand for ubiquitous communication and computing requires flexible communication protocols that can operate in a range of conditions. This thesis adopts and extends a mathematical model for these communication systems that accounts for uncertainty and time variation in link qualities. The arbitrarily varying channel (AVC) is an information theoretic channel model that has a time varying state with no statistical description. We assume the state is chosen by an adversarial jammer, reflecting the demand that our constructions work for all state sequences. In this thesis we show how resources such as secret keys, feedback, and side-information can help communication under this kind of uncertainty.

In order to put our results in context we provide a detailed taxonomy of the known results on AVCs in a unified setting. We then prove new results on list decoding with constrained states, a relaxation of the main problem in which the receiver may output a short list of possible messages. In particular, we show constant list sizes can achieve capacity under an average-error criterion and that a list size L can achieve within $O(1/L)$ from the capacity under a maximal-error criterion, complementing the known results for unconstrained state sequences.

If the encoder and decoder share a secret key, they can use a randomized code to make their communication more robust. An important practical consideration in using joint randomization for communication schemes is the tradeoff between key size and error probability. Inspired by ad-hoc networks, we propose a new AVC model called the AVC with "nosy noise," in which the jammer can observe the transmitted codeword non-causally. We show that a key size of $O(\log n)$ bits is sufficient to achieve capacity for codes of blocklength n in this model as well as in the case for the standard AVC. If a secure feedback channel is available, the key can be shared via feedback. Limited feedback can also be used to adapt the rate to the actual channel state sequence. We develop an AVC framework for rateless coding and show schemes that achieve rates arbitrarily close to the empirical mutual information.

Finally, we address the Gaussian version of the AVC, where we show that a key size of $O(\log n)$ bits is again sufficient to achieve capacity. This result allows us to find an achievable rate region for degraded broadcast channels. In the case where randomized coding is infeasible, we show how a known interference signal at the transmitter can enlarge the capacity region. This result has applications to watermarking and a model for spectrum-sharing communication systems.

Professor Michael Gastpar, Chair Date

Acknowledgements

Much of this work would have not been possible without discussions with my fellow students. Special thanks go to Bobak Nazer, Krish Eswaran, Hari Palaiyanur, and Pulkit Grover. Some of the work in Chapter 4 was done with Krish. Thanks are also due to research collaborators on projects that did not appear in this thesis : Venkat Anantharam, Alex Dimakis, Martin Wainwright, Anant Sahai, Tuncer Can Aysal, Anna Scaglione, and Mehmet Ercan Yildiz. Working on other problems was incredibly rewarding and helped me maintain some perspective. Surviving graduate school would have been impossible without the tireless Ruth Gjerde, who can make magic happen for EE graduate students.

I owe some of my residual sanity to all of the artistic groups with whom I worked during my graduate studies: the UC Berkeley Department of Dance and Performance Studies, Sudipto Chatterjee, the University Chorus, Perfect Fifth, the San Francisco Symphony Chorus, Pacific Collegium, and the Haydn Singers. Many thanks to all of my fellow actors, singers, and artists.

More thanks go out to my family and friends – without their support I doubt I could have made it.

Finally, I want to thank Michael Gastpar for taking the chance on me and helping me from start to finish.

For my parents, who gave me their love of learning.

ज्ञानं ज्ञेयं परिज्ञाता त्रिविधा कर्मचोदना ।
करणं कर्म कर्तेति त्रिविधः कर्मसंग्रहः ॥ १८॥

–भगवद् गीता १८

Contents

Auch klei- ne Din- ge kön- nen uns ent- zü- cken, auch klei- ne Din- ge kön- nen theu- er sein.

– *Auch kleine Dinge*, Hugo Wolf

Chapter 1

Introduction

1.1 Interference in new communication systems

As we move forward into the 21st century, communication technologies will play an ever increasing in commercial and social interactions. New applications from personal multimedia systems in “smart homes” to sensor networks monitoring oceanic conditions to municipal WiFi access require protocols for reliably transmitting data and control information. The demand for these new complex and interactive technologies requires theoretical models that can shed light on performance limits, illustrate tradeoffs between different design parameters, and provide guidelines for developing strategies to efficiently use the communication resources.

The dominant paradigm for organizing wireless communication services is *centralization.* This centralization comes in two forms. In cellular systems, the system designer can plan physical deployments to ensure good coverage and limit inter-cell interference. The FCC also uses a centralized planning system to allocate spectrum by requiring service providers to limit the amount of energy that leaks out of their allocated frequency band. Although centralized planning makes the design of wireless technologies easier, some claim that it has also led to an inefficient use of the radio spectrum resources.

Because of the consumer demand for data-rich ubiquitous connectivity, research on new communication systems has turned to *decentralized* architectures. One vision of the future includes wireless devices that make local peer-to-peer connections and seamlessly hand-off transmission to internet-enabled access points if the user walks out of range. Devices should also be able to adapt their protocols based on local spectrum availability in order to take advantage of unused resources. This emphasis on local spectrum reuse and adaptation represents a shift away from current centralized service planning.

The work in this thesis is inspired in part by engineering challenges that may arise in the design of communication technologies for three emerging applications for decentralized communication protocols: sensor networks, wireless ad-hoc networks, and "cognitive radio." These applications involve communication in environments that are difficult to model, which in turn requires the communication protocols to be robust to modeling errors. This difficulty may stem from the cost of measuring channel characteristics, the behavior of other users, or the interaction of heterogeneous systems using the same resources.

1.1.1 Sensor networks

The term "sensor network" is a convenient label for systems that consist of distributed components, each containing some sort of communication interface, deployed in an environment that is to be sensed and/or controlled. Applications for these networks range from environmental monitoring [105] to distributed target tracking [34, 118] to ubiquitous computing or pervasive networking [60] environments. A historical survey can be found in Chong and Kumar [37].

Sensor networks are often envisioned as consisting of a large number of cheap low-power "motes" equipped with sensors to detect a spatially varying quantity such as temperature or electromagnetic field strength and a wireless radio to permit communication between sensors or between sensors and a central processor. In these applications, a premium is put on energy efficiency, since the motes have limited battery power and cannot afford to be wasteful. Although this power-limited view is popular in theoretical studies, sensor networks have also been proposed for industrial monitoring, where the motes may be wired for data and power.

One advantage of producing cheap motes is that the same hardware should be usable for many different applications. The cornucopia of sensor network applications encompasses a wide range of channel conditions. A good communication protocol for such networks should be insensitive to these variations or be able to adapt to them. In particular, it may not be possible to characterize the interference from external sources prior to deployment.

In this thesis, we look at the benefits of using randomization on communication over channels with unknown but power-limited interference. Randomization requires the communicating parties to share a secure key, and we develop tradeoffs between the key size and the achievable probability of error. We can extend these results to a "streaming" construction that can allow rate-adaptation in the face of varying channel characteristics.

1.1.2 Ad-hoc networks

In an ad-hoc wireless network, several devices form a network to facilitate the transmission of data. Ad-hoc wireless networks are similar to sensor networks in many

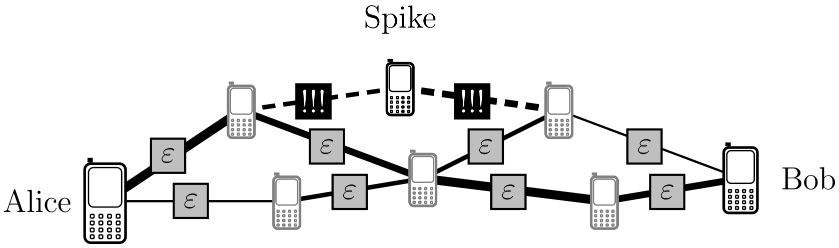


Figure 1.1: An ad-hoc network with a compromised user. Alice and Bob wish to exchange data. Although links between users following the protocol can be modeled by channels that randomly erase packets, a malicious user, Spike, may substitute fake traffic.

respects, and indeed the underlying physical modeling of the communication channels is often the same. Research on ad-hoc networks also addresses issues such as networking [121] and mobility [71].

Data transmitted between two users may be forwarded by third parties in the network, so one potential concern is data security and integrity. Although cryptographic protocols at the application layer may guarantee (under suitable complexity assumptions) that packets cannot be decrypted by intermediate users, the packets may still be corrupted by a third party that has, for example, downloaded a computer virus (see Figure 1.1). One class of problems that we study can apply to error control coding for this kind of adversarial tampering, which in general can be more harmful than the random-error model. Again, randomization will be a key component in our strategies. Because cryptographic keys may be used at the application layer, one way to interpret our results is a quantification of the tradeoff between extra overhead (in terms of the key rate) and data integrity (in terms of error probability).

1.1.3 Cognitive radio

Another hot application area in modern communication engineering is cognitive radio [114, 115], which refers to wireless systems that utilize sensing to adapt their behavior in order to coexist with other systems. Measurements of spectrum utilization indicate that there is a wide variation, geographically and temporally, in usage patterns [158,63]. Cognitive radios have been proposed as a solution for opportunistically communicating on unused spectrum. The spectrum is licensed to *primary* users who share the band with *secondary* or "cognitive" users. The secondary systems must follow certain etiquette rules, such as remaining silent when the primary system is transmitting. A cognitive system with a wideband spectrum sensor could theoretically perform dynamic spectrum management across several bands to scavenge sufficient resources for its communication needs.

With the Federal Communication Commission's recent decision to allow spectrum reuse in the 700 MHz band [64], these systems are moving ever closer to reality and have raised a number of interesting theoretical questions from limits of sensing [148] to allocation mechanisms [83] to information theoretic models [50, 90, 69] for interacting systems. In this thesis we will look at a channel model which does not make many assumptions on the time-dynamics of an interfering signal. Within this model there are several interesting problems highlighting the benefits of randomization, feedback, and knowledge of the primary signal on the capacity of cognitive radio systems.

1.1.4 Towards a mathematical model

Claude Shannon's landmark 1948 paper provided a mathematical foundation for the study of communication systems via probability theory [135]. Information theory studies communication by positing a statistical relationship between the received and transmitted signals. The simplest model for this relationship is a *discrete memoryless* *channel* $V(y|x)$, which is a conditional probability distribution for the output y conditioned on the input x. This model and its properties are discussed extensively in basic texts on information theory [41,44], and have been used over the last 60 years to derive important insights into real communication channels.

Strategies for communication over discrete memoryless channels (DMCs) involve communicating over blocks of many channel inputs. As the blocklength n becomes very large, large-deviations results in probability theory can be used to approximate the input-output relationship of the channel. Loosely speaking, at large blocklengths the channel's behavior is nearly deterministic, so coding schemes can use the large-deviations approximation.

The DMC model does not capture time variation in the channel – the transition probabilities $V(y|x)$ are fixed for all time. This model can be extended to include dynamics that can be described by finite state machines, Markov chains, linear time-invariant filters, or other transformations. These extended models have led to significant improvements for many widely used communication systems. However, in order to realize these gains, the dynamics of the channel must be modeled in advance.

For applications such as ad-hoc networks, sensor networks, and cognitive radio, the channel dynamics may not be known in advance and hence we cannot design communication protocols that are optimized for the particular communication channel. As discussed earlier, robustness and adaptivity are two important goals for communication in these applications. In order to look at these issues we must adopt a fundamentally nonstationary channel model. In the next section we will describe the *arbitrarily varying channel*, which is one approach to introducing uncertainty in the channel dynamics without making statistical assumptions.

1.2 Arbitrarily varying channels

The arbitrarily varying channel (AVC) is an information-theoretic model for communication channels with an unknown state. It is a simple extension to the discrete memoryless channel proposed by Shannon [135]. The analysis is worst-case, in that the state is assumed to be chosen by a malicious adversary, or *jammer*, whose objective is to minimize the maximum reliable rate of communication. In this section we will introduce the basic model and its variants.

First let us establish some notation conventions for the rest of this work. An appendix of notation with references to definitions is given in Appendix A. We will generally use calligraphic type for sets, and use the shorthand $[M] = \{1, 2, \ldots, M\}$ for integers M. For a set $\mathcal{X}$, the set $\mathcal{P}(\mathcal{X})$ is the set of all probability distributions on $\mathcal{X}$ and $\mathcal{P}_n(\mathcal{X})$ is the set of all probability distributions of composition n. We will write $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ for all conditional distributions on $\mathcal{Y}$ conditioned on $\mathcal{X}$. For random variables (X, Y) with joint distribution P_{XY} we will write P_X and P_Y for the marginal distributions and $P_{X|Y}$ for the conditional distribution of X given Y. For a distribution P and conditional distribution $V(y|x)$, the distribution PV or $P \times V$ is the joint distribution $P(x)V(y|x)$.

Given a sequence $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$, let $N(x|\mathbf{x}) = |\{i : x_i = x\}|$, the number of times x appears in $\mathbf{x}$. We denote the type of $\mathbf{x}$ by

$$T_{\mathbf{x}} = \frac{1}{n}(N(x_1|\mathbf{x}), N(x_2|\mathbf{x}), \ldots, N(x_{|\mathcal{X}|}|\mathbf{x})) \ . \tag{1.1}$$

The set of all length-n sequences of a fixed type P will be denoted by

$$\mathcal{T}_n(P) = \{\mathbf{x} \in \mathcal{X}^n : T_{\mathbf{x}} = P\} \ . \tag{1.2}$$

We will use $H(X)$ and $I(X \ \wedge \ Y)$ to denote the entropy of a random variable X

and the mutual information between two random variables X and Y. Suppose X and Y take values in $\mathcal{X}$ and $\mathcal{Y}$, respectively, and have joint distribution $P_{XY}(x, y)$ with marginal distributions P_X and P_Y. Then

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)} \tag{1.3}$$

$$I\left(X \ \wedge \ Y\right) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x) P_Y(y)} \ . \tag{1.4}$$

If $P_{Y|X}$ is the distribution in $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ of Y conditioned on X, the conditional entropy $H(Y|X)$ is

$$\begin{aligned} H(Y|X) &= H(Y) - I\left(X \ \wedge \ Y\right) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{XY}(x, y) \log \frac{1}{P_{Y|X}(y|x)} \ . \end{aligned} \tag{1.5}$$

We will also write entropies as functions of distributions, so $H(P_X) = H(X)$ and $H(P_{Y|X}|P_X) = H(Y|X)$. For two distributions P and Q in $\mathcal{P}(\mathcal{X})$ the Kullback-Leibler divergence $D\left(P \ \| \ Q\right)$ is defined by

$$D\left(P \ \| \ Q\right) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \ . \tag{1.6}$$

Properties of these functions can be found in standard information theory references [41, 44].

Let $d_{\max}(P, Q)$ be the maximum deviation (ℓ_∞ distance) between two probability distributions P and Q:

$$d_{\max}(P, Q) = \max_{x \in \mathcal{X}} |P(x) - Q(x)| \ . \tag{1.7}$$

The set T_P^ϵ is the ϵ-typical set around P:

$$T_P^\epsilon = \{\mathbf{x} : d_{\max}(P, T_{\mathbf{x}}) < \epsilon\} \ . \tag{1.8}$$

The set $T_V^\epsilon(\mathbf{x})$ is the (V, ϵ)-shell around $\mathbf{x}$:

$$T_V^\epsilon(\mathbf{x}) = \left\{ \mathbf{y} : d_{\max}\left(T_{\mathbf{y}}, \sum_x T_{\mathbf{x}}(x) V(y|x) \right) < \epsilon \right\} \ . \tag{1.9}$$

1.2.1 Basic definitions

We will model our time-varying channel by an *arbitrarily varying channel* (shown in Figure 1.2), which is a set $\mathcal{W} = \{W(\cdot|\cdot, s) : s \in \mathcal{S}\}$ of channels from an input alphabet $\mathcal{X}$ to an output alphabet $\mathcal{Y}$ parameterized by a state $s \in \mathcal{S}$. Unless otherwise specified, we will assume the sets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{S}$ are finite. If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ and $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ are length n vectors, the probability of observing the output $\mathbf{y}$ given the input $\mathbf{x}$ and state $\mathbf{s}$ over the AVC $\mathcal{W}$ is given by:

$$W(\mathbf{y}|\mathbf{x}, \mathbf{s}) = \prod_{i=1}^{n} W(y_i|x_i, s_i) \ . \tag{1.10}$$

The interpretation of (1.10) is that the channel state can change arbitrarily from time to time. We will think of this as an adversarial model in which the state is controlled by a *jammer* who wishes to stymie the communication between the encoder and decoder. As we will see, the capabilities of this adversary can be captured in the error criterion.

One extension of this model is to introduce constraints on the input and state sequences [45]. Let $g : \mathcal{X} \to \mathbb{R}^+$ and $l : \mathcal{S} \to \mathbb{R}^+$ be cost functions on the input and state sets . For discrete channels we will assume $\max_{x \in \mathcal{X}} g(x) = \gamma^* < \infty$ and $\max_{s \in \mathcal{S}} l(s) = \lambda^* < \infty$. The cost of vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ is the sum

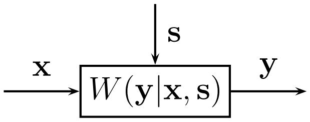


Figure 1.2: An arbitrarily varying channel.

of the cost on the elements:

$$g(\mathbf{x}) = \sum_{i=1}^{n} g(x_i) \tag{1.11}$$

$$l(\mathbf{s}) = \sum_{i=1}^{n} l(s_i) \ . \tag{1.12}$$

Without loss of generality we will assume $\min_x g(x) = 0$ and $\min_s l(s) = 0$. In point-to-point fixed blocklength channel coding problems, we will put constraint Γ and Λ on the average costs, so that

$$g(\mathbf{x}) \le n\Gamma \quad \text{a.s.} \tag{1.13}$$

$$l(\mathbf{s}) \le n\Lambda \quad \text{a.s.} \ . \tag{1.14}$$

We will also assume $\Gamma > 0$ and $\Lambda > 0$. If $\Gamma \ge \gamma^*$ we say the input is unconstrained, and if $\Lambda \ge \lambda^*$ we say the state is unconstrained. We will define the set

$$\mathcal{S}^n(\Lambda) = \{\mathbf{s} : l(\mathbf{s}) \le n\Lambda\} \tag{1.15}$$

to be the set of sequences with average cost less than or equal to Λ.

Example 1.1 – Bag of BSCs

Consider an AVC with $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{0,1\}$, and $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$ with

$s_j \in [0,1]$. For $s \in \mathcal{S}$, let the corresponding channel from $\mathcal{X}$ to $\mathcal{Y}$ be given by

$$W(y|x,s) = \begin{pmatrix} 1-s & s \\ s & 1-s \end{pmatrix} . \tag{1.16}$$

That is, the channel model is that of a binary symmetric channel (BSC) with time-varying crossover probability. A cost function $l(\cdot)$ on the state and a state constraint Λ can restrict the allowable mixtures of these channels.

Example 1.2 – Modulo additive

Consider an AVC with $\mathcal{X} = \{0,1,\ldots,p\}$, $\mathcal{Y} = \{0,1,\ldots,p\}$, and $\mathcal{S} = \{0,1,\ldots,p\}$, where

$$W(y|x,s) = \left(\begin{array}{c|c} 0 & I_{p-s} \\ \hline I_s & 0 \end{array} \right) , \tag{1.17}$$

where I_k is the $k \times k$ identity matrix. This corresponds to a channel of the form $Y = X \oplus S$, where the addition is performed modulo p. The simplest case is when $p = 2$. In this case, we will let $l(s) = s$, so that a cost constraint Λ becomes a bound on the empirical number of 1's in the state sequence.

Example 1.3 – Real additive channel

Consider an AVC with $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{0,1,2\}$, and $\mathcal{S} = \{0,1\}$ and the following channel matrices

$$W(y|x,0) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} , \qquad W(y|x,1) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} . \tag{1.18}$$

This corresponds to a channel of the form $Y = X + S$, where the addition is performed over the real numbers. To introduce constraints into this model, we can again let $l(s) = s$ to bound the empirical number of 1's in the state sequence. We can also use

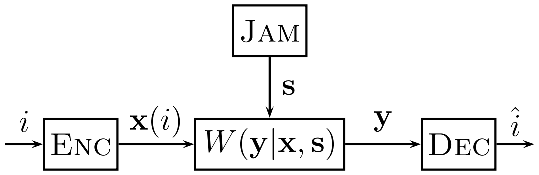


Figure 1.3: An arbitrarily varying channel with deterministic encoding.

an input cost function $g(x) = x$ with constraint Γ to bound the weight of the input sequence.

In the AVC literature, the form of the capacity formula is generally determined by two factors: the allowable coding strategies and the error criterion. In a randomized code, the encoder and decoder share a source of common randomness with which they may randomize their coding strategy, whereas a deterministic code uses a fixed mapping from messages to codewords. The state sequence may depend on different quantities – the message, the transmitted codeword, or both. Furthermore, we may relax the definition of correct decoding to allow the decoder to output a list of candidate codewords. All of these changes affect the way in which we define the error criterion.

A (n, N) **deterministic code** $\mathcal{C}$ for the AVC $\mathcal{W}$ with input constraint Γ is a pair of maps (ϕ, ψ) with

$$\phi : [N] \to \mathcal{X}^n \tag{1.19}$$

$$\psi : \mathcal{Y}^n \to [N] \ , \tag{1.20}$$

and for all $i \in [N]$ we have $g(\phi(i)) \leq n\Gamma$. The **rate** of the code is $n^{-1} \log N$. The deterministic encoder for the AVC is shown in Figure 1.3. The **decoding region** for message i is $D_i = \{\mathbf{y} : \psi(\mathbf{y}) = i\}$. We can also write a deterministic code $\mathcal{C}$ as a set of pairs $\{(\mathbf{x}(i), D_i) : i \in [N]\}$ with the encoder ϕ and decoder ψ defined implicitly.

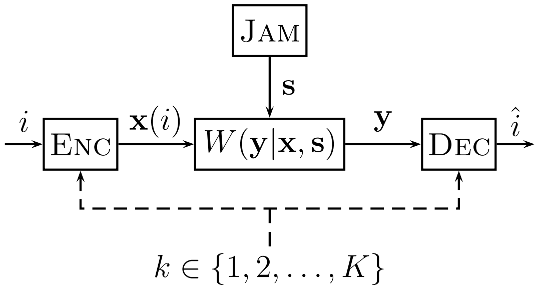


Figure 1.4: An arbitrarily varying channel with randomized encoding. The encoder and decoder share a secret key in $[K]$ that is unknown to the jammer.

In stating later results we will also write $\{\mathbf{x}(i) : i \in [N]\}$ or $\{\mathbf{x}_i : i \in [N]\}$ for the set of codewords in a code.

The **error for message i and state sequence $\mathbf{s} \in \mathcal{S}^n(\Lambda)$** is given by

$$\varepsilon(i, \mathbf{s}) = 1 - W\left(D_i | \mathbf{x}(i), \mathbf{s}\right) \ . \tag{1.21}$$

The **maximal** and **average error for a (n, N) deterministic code over an AVC $\mathcal{W}$ with cost constraint** Λ are given by

$$\varepsilon = \max_{i} \max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \varepsilon(i, \mathbf{s}) \tag{1.22}$$

$$\overline{\varepsilon} = \max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \frac{1}{N} \sum_{i=1}^{n} \varepsilon(i, \mathbf{s}) \ . \tag{1.23}$$

A (n, N) **randomized code C** for the AVC $\mathcal{W}$ with input constraint Γ is random variable taking on values in the set of deterministic codes. It is written as a pair of random maps (Φ, Ψ) where each realization is an (n, N) deterministic code satisfying the constraint Γ. If (Φ, Ψ) almost surely takes values in a set of K codes, then we call this an (n, N, K) randomized code. The **key size** of a randomized code (Φ, Ψ) is the entropy of the code $H(\mathbf{C})$. Note that the realization of the code is shared by

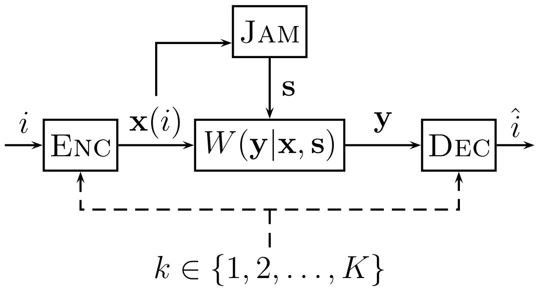


Figure 1.5: The nosy noise error model – the jammer knows both the message i and the codeword $\phi_k(i)$.

the encoder and decoder, so the key is known by both parties.

In the case where $\mathbf{C}$ is uniformly distributed on a set of K codes, the key size is simply $\log K$. For $K = \exp(nR_K)$ we can define the **key rate** to be R_K. We can also think of an (n, N, K) randomized code as a family of codes $\{(\phi_k, \psi_k) : k \in [K]\}$ indexed by a set of K *keys*, as shown in Figure 1.4. The *rate* of the code is $R = n^{-1} \log N$. The **decoding region** for message i under key k is $D_{i,k} = \{\mathbf{y} : \psi_k(\mathbf{y}) = i\}$. In the case where the bound on K is not explicit or unspecified, we write the random decoding region for message i as $\mathbf{D}_i = \{\mathbf{y} : \Psi(\mathbf{y}) = i\}$.

The power of randomized codes comes from modifying the definition of the error probability. Rather than demanding that the decoder error be small for every message and every key value, we instead require it to be small for every message *averaged over key values*. Randomization allows several different codewords to represent the same message. For maximal error, there are two cases to consider, depending on whether or not the state can depend on the actual *codeword*. For deterministic codes this distinction does not come up, since taking the maximum of $\varepsilon(i, \mathbf{s})$ over all messages i is the same as taking the maximum over codewords $\mathbf{x}(i)$.

The **standard maximal error** for a (n, N) randomized code over an AVC $\mathcal{W}$

with cost constraint Λ is given by

$$\varepsilon = \max_i \max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \mathbb{E}\left[1 - W\left(\mathbf{D}_i | \Phi(i), \mathbf{s}\right)\right] \ , \tag{1.24}$$

where the expectation is over the randomized code (Φ, Ψ). Here the variables $\mathbf{D}_i$ and $\Phi(i)$ correspond to the same realization of the key. The **nosy maximal error** for a (n, N) randomized code over an AVC $\mathcal{W}$ with cost constraint Λ is given by

$$\hat{\varepsilon} = \max_i \max_{J : \mathcal{X}^n \to \mathcal{S}^n(\Lambda)} \mathbb{E}\left[1 - W\left(\mathbf{D}_i | \Phi(i), J(\Phi(i))\right)\right] \ , \tag{1.25}$$

where the expectation is over the randomized code (Φ, Ψ). Again, the variables $\mathbf{D}_i$, $\Phi(i)$, and $J(\Phi(i))$ correspond to the same realization of the key. We call an AVC under the nosy maximal error criterion an **AVC with nosy noise**. We will not consider an average error criterion for randomized codes. Figure 1.5 shows the channel model under the nosy noise assumption. In the AVC with nosy noise, the jammer's strategies take the form of mappings $J : \mathcal{X}^n \to \mathcal{S}^n(\Lambda)$ from the codeword vectors to state sequences. This is a more pessimistic assumption on the jammer's capabilities, since it assumes that it has noncausal access to the transmitted codeword. Under randomized coding we will show that from a capacity standpoint all that matters is whether the jammer has access to the *current* input symbol.

A rate R is called achievable if for every $\epsilon > 0$ there exists a sequence of (n, N) codes of rate $R_n \geq R - \delta$ satisfying the input constraints whose probability of error is at most ϵ. Whether R is achievable will depend on the error criterion (maximal, average, nosy) and allowable codes (deterministic, randomized). For a fixed error criterion and type of coding, the supremum of achievable rates is the capacity of the arbitrarily varying channel.

We will write capacities using the letter C. The subscript will indicate the type of coding allowed : d for deterministic, r for randomized. A bar will indicate the average error criterion, a hat the nosy noise criterion, and no accent will indicate maximal error. Constraints will be given in parentheses, with the absence of a constraint indicating the unconstrained case. Thus $\overline{C}_d(\Gamma, \Lambda)$ is the deterministic coding capacity under average error with input constraint Γ and cost constraint Λ, C_r is the randomized coding capacity under maximal error with no constraints, and $\hat{C}_r(\Lambda)$ is the randomized coding capacity with nosy noise and state constraints.

1.2.2 Information quantities

For a fixed input distribution $P(x)$ on $\mathcal{X}$ and channel $V(y|x)$, we will also use the notation $I\left(P, V\right)$ to denote the mutual information between the input and output of the channel:

$$I\left(P, V\right) = \sum_{x,y} V(y|x)P(x) \log \frac{V(y|x)P(x)}{P(x) \sum_{x'} V(y|x')P(x')} \ . \tag{1.26}$$

We define the following sets:

$$\mathcal{I}(\Gamma) = \left\{ P \in \mathcal{P}(\mathcal{X}) : \sum_{s} P(x)g(x) \le \Gamma \right\} \tag{1.27}$$

$$\mathcal{Q}(\Lambda) = \left\{ Q \in \mathcal{P}(\mathcal{S}) : \sum_{s} Q(s)l(s) \le \Lambda \right\} \tag{1.28}$$

$$\mathcal{U}(P, \Lambda) = \left\{ U \in \mathcal{P}(\mathcal{S}|\mathcal{X}) : \sum_{s,x} U(s|x)P(x)l(s) \le \Lambda \right\} \ . \tag{1.29}$$

For an AVC $\mathcal{W} = \{W(y|x, s) : s \in \mathcal{S}\}$ with state constraint Λ we define two sets of

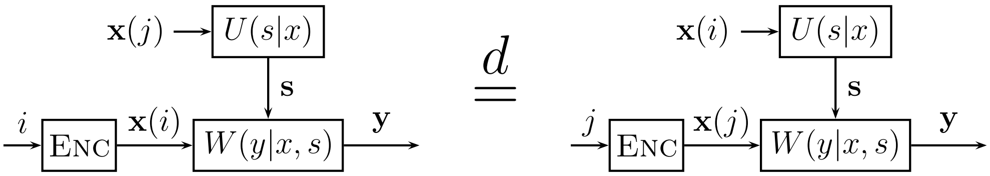


Figure 1.6: A symmetrizable channel. The jammer can simulate a codeword of the user to yield the same output distribution

channels:

$$\mathcal{W}_{std}(\Lambda) = \left\{ V(y|x) : V(y|x) = \sum_s W(y|x,s)Q(s), \quad Q(s) \in \mathcal{Q}(\Lambda) \right\} \tag{1.30}$$

$$\mathcal{W}_{dep}(P,\Lambda) = \left\{ V(y|x) : V(y|x) = \sum_s W(y|x,s)U(s|x), \quad U(s|x) \in \mathcal{U}(P,\Lambda) \right\} . \tag{1.31}$$

We will suppress the explicit dependence on Λ. The set in (1.30) is called the *convex closure* of $\mathcal{W}$, and the set in (1.31) is the *row-convex closure* of $\mathcal{W}$. In earlier works $\mathcal{W}_{dep}(P,\Lambda)$ is sometimes written as $\overline{\overline{\mathcal{W}}}$.

A central idea in the study of AVCs under average error is that of symmetrizability. We call a channel $V(y|x_1,x_2,\ldots,x_m)$ from $\mathcal{X}^m$ to $\mathcal{Y}$ **symmetric** if for any permutation π on $[m]$,

$$V(y|x_1,x_2,\ldots,x_m) = V(y|x_{\pi(1)},x_{\pi(2)},\ldots,x_{\pi(m)}) \quad \forall (x_1,x_2,\ldots,x_m,y) . \tag{1.32}$$

An AVC $\mathcal{W}$ is **symmetrizable** under the input distribution P and cost constraint Λ if there exists a distribution $U(s|x) \in \mathcal{U}(P,\Lambda)$ such that

$$V(y|x,x') = \sum_{s \in \mathcal{S}} W(y|x,s)U(s|x') \tag{1.33}$$

is symmetric. That is,

$$\sum_{s\in\mathcal{S}} W(y|x,s)U(s|x') = \sum_{s\in\mathcal{S}} W(y|x',s)U(s|x) \qquad \forall (x,x',y) \in \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \ . \quad (1.34)$$

The intuitive meaning of (1.34), as shown in Figure 1.6, is that the jammer can simulate the transmitter by choosing a codeword $\mathbf{x}'$ from the codebook $\mathcal{C}$ and passing it through the channel U to get a state sequence $\mathbf{s}$. The decoder will be unable to tell if the transmitted codeword was $\mathbf{x}$ or $\mathbf{x}'$ because the average channel in (1.34) symmetric between its two inputs.

1.2.3 List decoding and feedback

The no-frills version of the AVC model is that of deterministic coding under maximal error. The capacity for this model is still an open question, in some cases is equivalent to finding a zero-error capacity of a DMC [2]. The difficulty of finding the deterministic capacity under maximal error has spawned several modifications of the problem which we can think of as relaxations from the original coding problem. These relaxations either enhance the allowable coding strategies or weaken the error criterion.

Two relaxations that we have already seen are using randomized coding and the average probability of error criterion. Chapter 2 addresses another relaxation in the form of *list decoding*. In list decoding, we allow the decoder to output a list of candidate messages and declare an error only if the true message is not on the list. In this thesis we will only discuss deterministic list codes.

An (n, N, L) **deterministic list code** C for the AVC is a pair of maps (ϕ, ψ) where the encoding function is $\phi : [N] \to \mathcal{X}^n$ and the decoding function is $\psi : \mathcal{Y}^n \to [N]^L$. The *rate* of the code is $R = \log(N/L)$. The **codebook** is the set of vectors

$\{\mathbf{x}_i : 1 \leq i \leq N\}$, where $\mathbf{x}_i = \phi(i)$. The decoding region for message i is $D_i = \{\mathbf{y} : i \in \psi(\mathbf{y})\}$. We can again specify a code by the pairs $\{(\mathbf{x}_i, D_i) : i = 1, 2, \ldots, N\}$, with the encoder and decoder defined implicitly.

We have two notions of error probability, *maximal* and *average*. The maximal error is given by

$$\varepsilon_L = \max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \max_i \left(1 - W(D_i | X^n = \mathbf{x}_i, \mathbf{s})\right) \ . \tag{1.35}$$

The average probability of error is given by

$$\overline{\varepsilon}_L = \max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \frac{1}{N} \sum_{i=1}^{N} \left(1 - W(D_i | \mathbf{x}_i, \mathbf{s})\right) \ . \tag{1.36}$$

The definition of achievability and capacity for list codes are the same as for the deterministic and randomized coding strategies above. We will denote the list-L capacity for maximal error by C_L and for average error by $\overline{C}_L$.

The concept of symmetrizability extends to lists. For an integer m let $\mathcal{U}_{\mathrm{sym}}(m)$ be the set of channels $U : \mathcal{X}^m \to \mathcal{S}$ that symmetrize the AVC $\mathcal{W}$:

$$\begin{aligned} \mathcal{U}_{\mathrm{sym}}(m) = \Bigg\{ U(s|x^m) : V(y|x, x_1, \ldots, x_m) = \sum_s W(y|x, s) U(s|x_1, x_2, \ldots, x_m) \\ \text{is symmetric} \Bigg\} \ . \end{aligned} \tag{1.37}$$

We call an AVC m-**symmetrizable** if $\mathcal{U}_{\mathrm{sym}}(m) \neq \emptyset$. The **symmetrizability** L_{sym} of an unconstrained AVC is the largest integer such that $\mathcal{W}$ is L_{sym}-symmetrizable.

Another relaxation of the coding problem is to allow the encoder access to the outputs of the channel via a noiseless feedback link. In this thesis we will not study the noiseless feedback case, but the problem has been investigated by Ahlswede [5]. His feedback construction uses a concatenation of list codes. The decoder decodes the first code into a list of candidate codewords, and because of the feedback the encoder can generate the same list. The encoder then uses another list code to disambiguate this list, and the decoder list-decodes the disambiguation information. By iterating this process the decoder can eventually decode the correct codeword.

1.2.4 Side information and correlated sources

We can also consider channel models in which the encoder or decoder is given side information about the state sequence $\mathbf{s}$ that governs the channel. This model has been used for memories [18] and localized error models [12]. The main difference is to allow the encoder or decoder mapping to depend on $\mathbf{s}$, so the encoder maps $[N] \times \mathcal{S}^n \to \mathcal{X}^n$ in the case where the encoder has full non-causal side information. The remaining coding definitions are the same. Another interesting case is where the channel has two state sequences, one known to the transmitter and the other generated by the jammer. In the Gaussian AVC, the side information at the encoder can help enlarge the capacity region, as we will show in Chapter 5.

1.3 Previous results on discrete AVCs

We will now summarize some of the previous results on AVCs in order to provide some context for the results in later chapters. Lapidoth and Narayan wrote a survey on models of channel uncertainty [104] which covers some of these results but in the interest of completeness we will briefly describe the results and the relevant arguments. Many of the capacity results are given by one of the following two quantities:

$$C_{\mathrm{std}}(\Gamma, \Lambda) = \max_{P \in \mathcal{I}(\Gamma)} \min_{V \in \mathcal{W}_{std}(\Lambda)} I\left(P, V\right) \tag{1.38}$$

$$C_{\mathrm{dep}}(\Gamma, \Lambda) = \max_{P \in \mathcal{I}(\Gamma)} \min_{V \in \mathcal{W}_{dep}(P,\Lambda)} I\left(P, V\right) \ . \tag{1.39}$$

If $\Gamma = \gamma^*$ then the input constraint is vacuous and we will write $C_{\text{std}}(\Lambda)$ for $C_{\text{std}}(\Gamma, \Lambda)$ and $C_{\text{std}}(\Lambda)$ for $C_{\text{dep}}(\Gamma, \Lambda)$.

In information theory there is a distinction between *information quantities* such as $C_{\text{std}}(\Gamma, \Lambda)$ and $C_{\text{dep}}(\Gamma, \Lambda)$ that are defined in terms of mutual information expressions and *operational quantities* such as the randomized coding capacity under maximal error $C_r(\Gamma, \Lambda)$. Previous researchers (see Theorems 1 and 2) have shown that

$$C_r(\Gamma, \Lambda) = C_{\text{std}}(\Gamma, \Lambda) \ . \tag{1.40}$$

In Theorem 14 in Chapter 3 we will show that the randomized coding capacity under nosy noise $\hat{C}_r(\Gamma, \Lambda)$ is given by

$$\hat{C}_r(\Gamma, \Lambda) = C_{\text{dep}}(\Gamma, \Lambda) \ . \tag{1.41}$$

Because $\mathcal{W}_{std} \subseteq \mathcal{W}_{dep}$, in general we have $C_{\text{dep}} \leq C_{\text{std}}$. In some cases equality can hold, as in the following example.

Example 1.4 – Bit-flipping (mod-two adder)

Consider an AVC with input alphabet $\mathcal{X} = \{0, 1\}$, state alphabet $\mathcal{S} = \{0, 1\}$ and output alphabet $\mathcal{Y} = \{0, 1\}$, with

$$y = x \oplus s \ , \tag{1.42}$$

where $\oplus$ denotes addition modulo two. This is a “bit-flipping AVC” in which the jammer can flip the input ($s = 1$). We choose $l(s) = s$ so that the state constraint Λ bounds the fraction of bits which can be flipped by the jammer. In this example we will not impose a constraint on the input. For this AVC it can be shown [45, 101] that

$$C_{\text{std}}(\Lambda) = 1 - h_b(\Lambda) \tag{1.43}$$

$$C_{\text{dep}}(\Lambda) = 1 - h_b(\Lambda) \ , \tag{1.44}$$

where $h_b(t) = -t \log t - (1-t) \log(1-t)$ is the binary entropy function. In this case, we have $C_{\text{std}}(\Lambda) = C_{\text{dep}}(\Lambda)$. Furthermore, the capacity under randomized coding and maximal error $C_r(\Lambda) = C_{\text{std}}(\Lambda)$ and the capacity under randomized coding and nosy noise is $\hat{C}_r(\Lambda) = C_{\text{dep}}(\Lambda)$.

Although for this bit-flipping example the two max-min expressions have the same value, this is not the case for general AVCs. In the previous example the addition was taken over the finite field $\mathbb{F}_2$. If we instead take the addition over the integers the two quantities are different.

Example 1.5 – Real additive channel
Consider an AVC with input alphabet $\mathcal{X} = \{0, 1\}$, state alphabet $\mathcal{S} = \{0, 1\}$ and output alphabet $\mathcal{Y} = \{0, 1, 2\}$, with

$$y = x + s \ . \tag{1.45}$$

We choose $l(s) = s$ and $g(x) = x$ so that the constraints Λ and Γ on the jammer and encoder bound the weight of their inputs.

Let $Q = (1-q, q)$ be a distribution on $\mathcal{S}$ and consider the average channel V under Q:

$$V = \begin{pmatrix} 1-q & q & 0 \\ 0 & 1-q & q \end{pmatrix} \ , \tag{1.46}$$

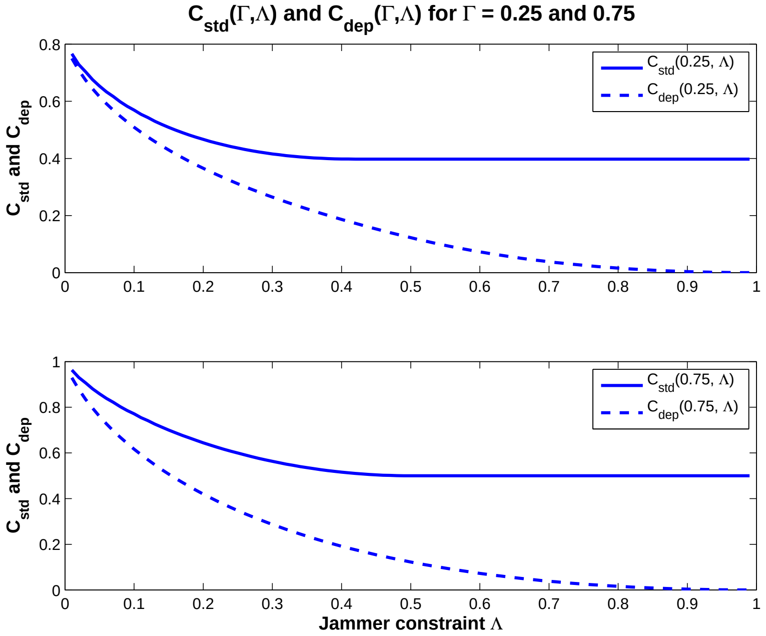


Figure 1.7: The max-min values $C_{\mathrm{std}}(\Gamma, \Lambda)$ and $C_{\mathrm{dep}}(\Gamma, \Lambda)$ for the real adder channel in Example 1.5 as a function of Λ for fixed values of $\Gamma = 0.25$ (upper plot) and $\Gamma = 0.75$ (lower plot). Within each plot the higher curve is $C_{\mathrm{std}}(\Gamma, \Lambda)$ and the lower curve is $C_{\mathrm{dep}}(\Gamma, \Lambda)$.

If the input distribution is $P = (1-p, p)$, the mutual information can be written as

$$\begin{aligned} I\left(X \ \wedge \ Y\right) &= H(X) - H(X|Y) \\ &= h_b(p) - ((1-p)q + p(1-q))h_b\left(\frac{(1-p)q}{(1-p)q + p(1-q)}\right) . \qquad (1.47) \end{aligned}$$

For the unconstrained channel, Csiszár and Narayan [45] showed that the optimal p and q are both equal to $1/2$, which yields $C_{\mathrm{std}}(\Gamma, \Lambda) = 1/2$ for $\Gamma \geq 1/2$ and $\Lambda \geq 1/2$. Unfortunately, for other values of the constraints it is more difficult to find a closed-form expression for the capacity.

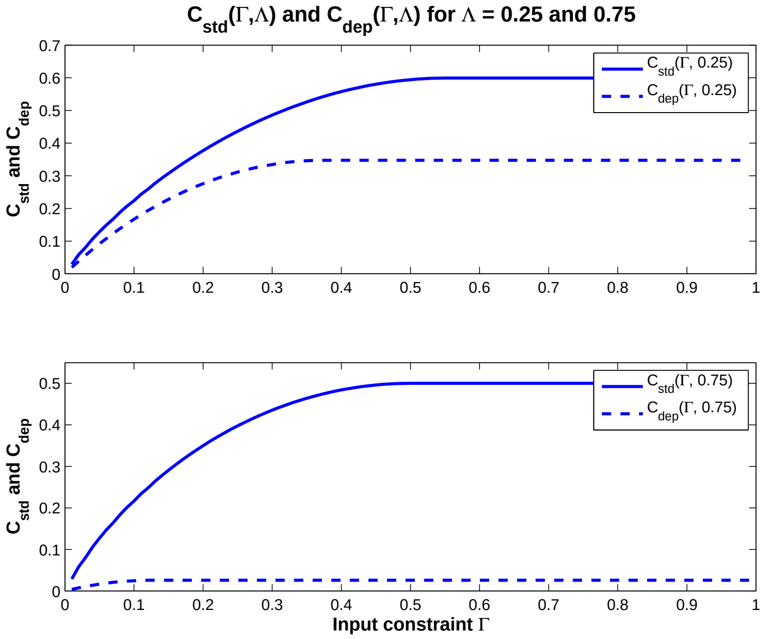


Figure 1.8: The max-min values $C_{\mathrm{std}}(\Gamma, \Lambda)$ and $C_{\mathrm{dep}}(\Gamma, \Lambda)$ for the real adder channel in Example 1.5 as a function of Γ for fixed values of $\Lambda = 0.25$ (upper plot) and $\Lambda = 0.75$ (lower plot). Within each plot the higher curve is $C_{\mathrm{std}}(\Gamma, \Lambda)$ and the lower curve is $C_{\mathrm{dep}}(\Gamma, \Lambda)$.

	Deterministic	Randomized	List	Feedback
Average	C_{std} or 0	C_{std}	C_{std} or 0	C_{std}
Maximal	$\geq \min\left(C_{\text{dep}}, D(P)\right)$	C_{std}	C_{dep}	C_{dep}
Nosy	$\geq \min\left(C_{\text{dep}}, D(P)\right)$		C_{dep}	C_{dep}

Table 1.1: Previously known capacity results for unconstrained arbitrarily varying channels. The function $D(P)$ is defined in Section 1.3.2. For list decoding under average error the capacity is for a constant list size. For list decoding under maximal error and nosy noise, the capacity is achievable with growing list sizes. The randomized capacity under nosy noise is proved in Theorem 14 of Chapter 3. The feedback capacities hold under some additional assumptions.

We can compute the quantity $C_{\text{dep}}(\Gamma, \Lambda)$ explicitly :

$$C_{\text{dep}}(\Gamma, \Lambda) = \begin{cases} h_b\left(\frac{1-\Lambda}{2}\right) - \frac{1+\Lambda}{2} h_b\left(\frac{2\Lambda}{1+\Lambda}\right) & \Gamma \geq \frac{1-\Lambda}{2} \\ h_b(\Gamma) - (\Lambda + \Gamma) h_b\left(\frac{\Lambda}{\Lambda+\Gamma}\right) & \Gamma < \frac{1-\Lambda}{2} \end{cases} . \tag{1.48}$$

The details are given in Section C.1.

Figures 1.7 shows the two capacities as a function of the cost constraint for different values of the input constraint. Figure 1.8 shows the two quantities as a function of the input constraint for different cost constraints. It is clear that for this example $C_{\text{dep}}(\Gamma, \Lambda)$ is smaller than $C_{\text{std}}(\Gamma, \Lambda)$.

From 1960 until 1988, nearly all of the results on the arbitrarily varying channel were for the case of unconstrained states and inputs. In this thesis we will focus on AVCs with constraints. The results for unconstrained AVCs are summarized in Table 1.1. In the remainder of this section we will describe these results and the corresponding results for constrained AVCs if they exist.

1.3.1 Randomized coding and maximal error

The first result on arbitrarily varying channels was in the seminal paper by Blackwell, Breiman, and Thomasian [28]. In their model, the jammer is allowed to choose its state S_t at time t with knowledge of the inputs X_i and outputs Y_i for times $i = 1, 2, \ldots, t-1$. Communication over the AVC was modeled as a two person zero-sum game between the jammer and the encoder/decoder. The first player chooses a jamming strategy for selecting S_t, and the second player chooses a deterministic code. The payoff is 1 to the jammer if a decoding error is made. Mixed strategies for the second player correspond to randomized codes. Using mixed strategies the maximal probability of error under randomized coding is the value of the game.

Theorem 1 (Randomized coding under maximal error for unconstrained AVCs [28])**.** *For an AVC $\mathcal{W}$ without constraints, the randomized coding capacity is given by*

$$C_r = C_{\mathsf{std}} \ . \tag{1.49}$$

Furthermore, C_r is the capacity when the state S_t at time t can depend on all inputs X_i and outputs Y_i for $i = 1, 2, \ldots, t-1$.

Although it may be possible to use a similar game-theoretic argument to show an analogous result for constrained AVCs, the work of Csiszár and Narayan on discrete constrained AVCs [45, 46, 47] uses more "traditional" combinatorial arguments [44].

Theorem 2 (Randomized coding under maximal error for constrained AVCs [45])**.** *For an AVC $\mathcal{W}$ with input and state cost functions $\Gamma(\cdot)$ and $\Lambda(\cdot)$ with constraints Γ and Λ, the randomized coding capacity is given by*

$$C_r(\Gamma, \Lambda) = C_{\mathsf{std}}(\Gamma, \Lambda) \ . \tag{1.50}$$

1.3.2 Deterministic coding and maximal error

If the encoder and decoder use a deterministic code, the error under nosy noise is the same as the maximal error. Because maximizing over messages is the same as maximizing over codewords, we can see:

$$\begin{aligned}
\hat{\varepsilon} &= \max_{i} \max_{J:\mathcal{X}^n \to \mathcal{S}^n(\Lambda)} \left(1 - W\left(D_i | \phi(i), J(\phi(i))\right)\right) && (1.51)\\
&= \max_{i} \max_{J:[N] \to \mathcal{S}^n(\Lambda)} \left(1 - W\left(D_i | \phi(i), J(\phi(i))\right)\right) && (1.52)\\
&= \max_{i, \mathbf{s} \in \mathcal{S}^n(\Lambda)} \left(1 - W\left(D_i | \phi(i), \mathbf{s}\right)\right) && (1.53)\\
&= \varepsilon \ . && (1.54)
\end{aligned}$$

Thus for deterministic codes under the maximal error criterion we may assume that the jammer knows the codeword being transmitted. Therefore the deterministic coding capacity under nosy noise is the same as the deterministic coding capacity under maximal error:

$$\hat{C}_d = C_d \ . \tag{1.55}$$

One strategy for the jammer against message i is to choose a channel $U(s|x) \in \mathcal{U}$ and generate its input $\mathbf{s}$ by taking a codeword $\mathbf{x}(i)$ corresponding to message i and passing it through the channel U. If the encoder transmits message i, then for this choice of $\mathbf{s}$ the channel has the distribution of a DMC $V(y|x)$ given by

$$V(y|x) = \sum_{s} W(y|x, s) U(s|x) \ . \tag{1.56}$$

The quantity C_{dep} is a natural upper bound on the capacity under nosy noise $\hat{C}_d$, and by (1.55) it is also a bound on the capacity under maximal error C_d. In Chapter 3 we will show that the randomized coding capacity under nosy noise is given by $\hat{C}_r = C_{\text{dep}}$.

For AVCs with binary output alphabets Ahlswede and Wolfowitz [20] showed that the capacity under maximal error and deterministic coding C_d is equal to C_{dep}. Extensions of this result to other classes of AVCs were found by Ahlswede [2,3] and Kambo and Singh [91]. The best results to date are due to Csiszár and Körner [43]. They define a relation $x \overset{W}{\sim} x'$ between x and $x' \in \mathcal{X}$ if there are distributions $Q_1, Q_2 \in \mathcal{P}(\mathcal{S})$ such that

$$\sum_{s \in \mathcal{S}} W(y|x,s) Q_1(s) = \sum_{s \in \mathcal{S}} W(y|x',s) Q_2(s) \qquad \forall y \ . \tag{1.57}$$

Their result is an achievable rate using deterministic coding.

Theorem 3 (Deterministic coding under maximal error for unconstrained AVCs [43])**.** *For an unconstrained AVC $\mathcal{W}$, the following bound holds on the deterministic coding capacity under maximal error:*

$$C_d \geq \min\left(C_{\text{dep}}, D(P)\right) \ , \tag{1.58}$$

where

$$\mathcal{D} = \left\{ F(x,x') \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : F\left(\{X \overset{W}{\sim} X'\}\right) = 1,\ X \sim P,\ X' \sim P \right\} \tag{1.59}$$

$$D(P) = \min_{F \in \mathcal{D}} I\left(X \ \wedge \ X'\right) \ . \tag{1.60}$$

One explanation for the difficulty in establishing the deterministic coding capacity for general AVCs is that this problem has connections to a difficult open problems in information theory. Ahlswede showed that finding C_d for certain AVCs is equivalent to finding the zero-error capacity [137] of a corresponding DMC [2]. Although Lovász solved a special case of the zero-error capacity problem [106], finding the zero-error capacity in general is still a major open problem in information theory [97].

The AVC given in Example 1.4 on page 21 was a channel with binary input, output, and state, whose output is the modulo-two sum of the input and state. Designing a deterministic code for maximal error in this channel with state constraint Λ is the same as designing a binary code to correct all error patterns of Hamming weight less than or equal to Λn. Some error-correcting codes in the coding theory literature are also designed to correct error patterns of bounded Hamming weight.

1.3.3 Deterministic coding and average error

Randomization relaxes the stringent requirements of deterministic coding for maximal error by making the encoder and decoder more powerful. Another relaxation is to change the error criterion from the maximum over all messages to the average. That is, instead of demanding the error under a state sequence $\mathbf{s}$ be small for every message m, we can require instead that the error be small for a vanishingly small fraction of messages. Under this coding model, Ahlswede proved that the capacity for unconstrained AVCs exhibits a *dichotomy* – the capacity is either 0 or equal to the randomized coding capacity [6]. The symmetrizability condition from (1.34) was shown to be sufficient to render the capacity 0 by Ericson [59] and necessary by Csiszár and Narayan [46].

Symmetrizability is also a necessary and sufficient condition for the capacity to be 0 in the constrained setting. For a given input distribution P, we first calculate the function

$$\lambda_1(P) = \min_{U \in \mathcal{U}_{\mathrm{sym}}(1)} \sum_{s,x} l(s) U(s|x) P(x) \ . \tag{1.61}$$

We call the channel symmetrizable under input distribution P with cost constraint Λ if $\lambda_1(P) < \Lambda$.

Theorem 4 (Deterministic coding dichotomy [6, 46]). *For an AVC $\mathcal{W}$ without constraints:*

$$\overline{C}_d = \begin{cases} 0 & \textit{if } \mathcal{W} \textit{ is symmetrizable} \\ C_{\mathrm{std}} & \textit{otherwise} \end{cases} \tag{1.62}$$

For an AVC with input constraint Γ and cost constraint Λ,

$$\overline{C}_d(\Gamma, \Lambda) = \begin{cases} 0 & \textit{if } \max_{P \in \mathcal{I}(\Gamma)} \lambda_1(P) < \Lambda \\ \max_{P \in \mathcal{I}(\Gamma) : \lambda_1(P) \geq \Lambda} \ \min_{V \in \mathcal{W}_{std}(\Lambda)} I\left(P, V\right) & \textit{if } \max_{P \in \mathcal{I}(\Gamma)} \lambda_1(P) > \Lambda \end{cases} \tag{1.63}$$

For constrained AVCs the second expression in (1.63) may in general be smaller than the $C_{\mathrm{std}}(\Gamma, \Lambda)$.

In order to prove this result in the unconstrained setting, Ahlswede used a subsampling argument known as the "elimination technique" [6]. Beginning with a randomized code $\mathbf{C}$ that achieves a rate below C_r, he showed that a new randomized code consisting of n^2 iid codebooks sampled from $\mathbf{C}$ has small average probability of error. If $\overline{C}_d > 0$ then the encoder sends a codeword consisting of two parts. It first chooses one of the n^2 codebooks uniformly at random. Because the deterministic coding capacity $\overline{C}_d$ is positive, there exists a deterministic code which can transmit the choice of codebook with small average probability of error. This requires only

$2 \log n$ bits and so the blocklength required is negligible. The second part of the encoder's codeword encodes the message using the selected codebook. Thus the overall codeword consists of a short prefix containing the choice of codebook followed by the encoded message.

The dichotomy between positive and zero capacity still holds in constrained AVCs. However, the prefixing argument will not work in general for constrained AVCs, since the jammer could use a state sequence with higher cost during the prefix. The subsampling argument can still be used to construct randomized codebooks with smaller key size. This line of argument is taken in Chapter 3. Csiszár and Narayan [46] proved their result using properties of constant composition codebooks and techniques due to Dobrushin and Stambler [52]. An important difference between the unconstrained and constrained setting is that the capacity for a constrained AVC may be positive but strictly lower than the randomized coding capacity, as shown by the following example.

Example 1.6 – Real additive channel [46]

Consider the additive channel $Y = X + S$ with $\mathcal{X} = \mathcal{S} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1, 2\}$. Let the cost functions be $g(x) = x$ and $l(s) = s$. If $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(S = 1) = q$ then

$$C_r(\Gamma, \Lambda) = \max_{p \le \Gamma} \min_{q \le \Lambda} H(pq, (1-p)(1-q), p + q - 2pq) - h(q) \ , \tag{1.64}$$

which is 1/2 for $\Gamma \ge 1/2$ and $\Lambda \ge 1/2$. The optimal p and q are both equal to 1/2.

The deterministic coding capacity under average error $\overline{C}_d(\Lambda, \Gamma)$ has different behavior in the region where $\Gamma \ge 1/2$ and $\Lambda \ge 1/2$. First, if $\Lambda > 1/2$ the encoder cannot choose $p = 1/2$ because the jammer can symmetrize that distribution. Therefore the encoder cannot choose the distribution that maximizes $C_r(\Gamma, \Lambda)$. Note that if $\Gamma \le \Lambda$ then the capacity is 0, since any input distribution satisfying the input cost constraint can be symmetrized by the jammer. However, if $\Gamma > \Lambda$ then the jammer cannot symmetrize all input distributions. In particular, $p = \Lambda$ maximizes the mutual information given in Theorem 4, so

$$\overline{C}_d(\Gamma, \Lambda) = \min_{q \leq \Lambda} H(\Lambda q, (1-\Lambda)(1-q), \Lambda + q - 2\Lambda q) - h(q) \qquad \Gamma > \Lambda > 1/2 \ . \tag{1.65}$$

Therefore we can see that for some constraint values, $0 < \overline{C}_d(\Gamma, \Lambda) < C_r(\Gamma, \Lambda)$, which shows that the deterministic coding capacity can be positive and strictly smaller than the randomized coding capacity.

1.3.4 List decoding

In Chapter 2 we present some new results on list decoding for AVCs with cost constraints. The proofs of these results are based on the proofs of earlier results on list decoding for unconstrained AVCs. For maximal error, Ahlswede found list decoding capacities for DMCs and AVCs [4] with an eye towards proving coding theorems for AVCs with feedback [5]. He later strengthened these results into the following Theorem.

Theorem 5 (List decoding under maximal error for unconstrained AVCs [4, 11])**.** *For an unconstrained AVC $\mathcal{W}$, any rate $R < C_{\text{dep}}$ is achievable under maximal error using list decoding with lists of size*

$$L(R) = \left\lfloor \frac{\log |\mathcal{Y}|}{C_{\text{dep}} - R} \right\rfloor + 1 \ . \tag{1.66}$$

For average error, Pinsker conjectured that all rates below C_{std} should be achievable using list decoding with constant list size. Ahlswede and Cai showed that this was indeed the case [13], but did not prove tight bounds on the list size. The list size was improved by Blinovsky and Pinsker [30], who also observed that either $\overline{C}_L = 0$ or $\overline{C}_L = C_r = C_{\text{std}}$ for all $L \geq 1$, which is the list decoding analogue of the dichotomy in Theorem 4. Independently, Hughes [85] and Blinovsky, Narayan, and Pinsker [29] showed that the concept of symmetrizability extends to lists as well and that lists greater than the list-symmetrizability L_{sym} (defined on page 19) are sufficient to achieve the randomized coding capacity $C_r = C_{\text{std}}$.

Theorem 6 (List decoding under average error for unconstrained AVCs [85,29])**.** *For an unconstrained AVC $\mathcal{W}$, the capacity under average error with list decoding using lists of size L is given by*

$$\overline{C}_L = \begin{cases} 0 & L \leq L_{\text{sym}} \\ C_{\text{std}} & L > L_{\text{sym}} \end{cases} \tag{1.67}$$

where L_{sym} is the symmetrizability of the channel defined on page page 19.

Blinovsky et al.'s proof uses the dichotomy to establish the capacity result and does not appear to extend naturally to constrained AVCs. In Chapter 2 we use the method of Hughes to prove some analogous results for the constrained case.

List decoding can be used in conjunction with noiseless feedback to yield a capacity expression for the arbitrarily varying channel. Because the encoder can track the decoder's actions, it can use a multi-stage coding strategy based on list codes in which each stage disambiguates the list for the previous stage. Ahlswede [5] proved a coding result for AVCs with maximal error and noiseless feedback and showed that under some conditions the capacity with feedback is equal to C_{dep}.

1.3.5 Side information and other extensions

A key resource in some communication scenarios is side information about the channel behavior [138]. This information can be available at the encoder, decoder, or both. A unified framework for studying channel coding under different assumptions about side information availability was recently proposed by Moulin and Wang in the context of distortion-attack channels [116]. In the AVC literature, the problem was systematically studied by Ahlswede and Wolfowitz [19] under randomized coding and Stambler [145] under deterministic coding.

Theorem 7 (Randomized coding with side information at the decoder [19,145])**.** *For the AVC $\mathcal{W}$, if the state sequence* $\mathbf{s}$ *is available at the decoder then the capacity under randomized coding and average error is given by*

$$\overline{C}_{r,\mathrm{CSIR}} = \max_{P \in \mathcal{P}(\mathcal{X})} \min_{s \in \mathcal{S}} I\left(P, W(\cdot|\cdot, s)\right) \ . \tag{1.68}$$

The capacity under deterministic coding and average error is also given by the same expression:

$$\overline{C}_{d,\mathrm{CSIR}} = \max_{P \in \mathcal{P}(\mathcal{X})} \min_{s \in \mathcal{S}} I\left(P, W(\cdot|\cdot, s)\right) \ . \tag{1.69}$$

The paper of Gel'fand and Pinsker [70] found the capacity of channels with iid state sequences known at the encoder. The corresponding result for AVCs was found by Ahlswede [10], who used the elimination technique [6] together with a permutation construction to show the following result for AVCs with the state sequence $\mathbf{s}$ known to the encoder.

Theorem 8 (Side information at the encoder [10])**.** *For the AVC $\mathcal{W}$, if the state sequence* $\mathbf{s}$ *is available at the encoder, the capacity for deterministic coding is the same for maximal and average error and is given by*

$$\overline{C}_d = \min_{Q \in \mathcal{P}(\mathcal{S})} \max_{P(U,S,X) \in \mathcal{P}_Q(\mathcal{U} \times \mathcal{S} \times \mathcal{X})} \left(I\left(U \ \wedge \ Y\right) - I\left(U \ \wedge \ S\right)\right) \ , \tag{1.70}$$

where U is a random variable taking values in $\mathcal{U}$ with $|\mathcal{U}| \le |\mathcal{X}| + |\mathcal{S}|$ and

$$\mathcal{P}_Q(\mathcal{U} \times \mathcal{S} \times \mathcal{X}) = \left\{ P(u,s,x) : \sum_{u,x} P(u,s,x) = Q(s) \right\} . \tag{1.71}$$

1.3.6 Error exponents

Once a capacity result has been established, the second order characterization is to find how fast the probability of error can be made to decay with the blocklength. The first results on error exponents for randomized coding were due to Stiglitz [146]. Ericson [59] established a tradeoff between the error and the amount of common randomness and showed that using an exponential (in the blocklength n) number of codebooks in a randomized code leads to an exponential decay in the average probability of error.

A more thorough study of exponents was conducted by Hughes and Thomas [88], who derived AVC versions of the random coding, sphere packing, and expurgated bounds [67]. The class of randomized codes that they consider has a simple form: a fixed constant-composition code followed by a random permutation of the n channel symbols of the codeword. This scrambling makes the jammer's input exchangeable. We will need the random coding exponent to state later results in this thesis.

Theorem 9 (Random coding exponent for AVCs under maximal error [88])**.** *Let $\mathcal{W}$ be an AVC. For any rate $R > 0$ and $\delta > 0$, and type $P \in \mathcal{P}_n(\mathcal{X})$, let (Φ, Ψ) be the randomized code of blocklength n formed by choosing $\lceil \exp(n(R - \delta)) \rceil$ codewords independently and uniformly from $\mathcal{T}_n(P)$ and letting Ψ be the maximum mutual information decoder for this set. Then there exists an $n_1(|\mathcal{X}|, |\mathcal{Y}|, |\mathcal{S}|, \delta)$ such that for blocklength $n \ge n_1$ and all $Q \in \mathcal{P}_n(\mathcal{S})$ the error is upper bounded:*

$$\max_{\mathbf{s} \in \mathcal{T}_n(Q)} \varepsilon(\mathbf{s}) \le \exp\left(-n\left[E_r(R, \mathcal{W}, P, Q) - \delta\right]\right) , \tag{1.72}$$

where

$$E_r(R, \mathcal{W}, P, Q) = \min_{\tilde{P}_{XYS} : \tilde{P}_X = P, \tilde{P}_S = Q} D\left(\tilde{P}_{XYS} \parallel W \times P \times Q\right) + \left|I(P, \tilde{P}_{Y|X}) - R\right|^+ , \tag{1.73}$$

and the distribution $W \times P \times Q$ *is given by*

$$(W \times P \times Q)(y, x, s) = W(y|x, s)P(x)Q(s) , \tag{1.74}$$

$\tilde{P}_{XYS}$ *is a joint distribution on* $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$ *with marginals and conditional distributions* $\tilde{P}_X$*,* $\tilde{P}_S$*,* $\tilde{P}_{Y|X}$*, and* $D(\cdot \parallel \cdot)$ *is the Kullback-Leibler divergence.*

When it is clear, we will suppress the dependence on $\mathcal{W}$ and define

$$E_r(R, P, \Lambda) = \min_{Q \in \mathcal{Q}(\Lambda)} E_r(R, \mathcal{W}, P, Q) . \tag{1.75}$$

1.3.7 Multiuser channels

The AVC model has also been extended to multiuser channels. The coding and capacity definitions for point-to-point channels extend readily to the case of multiple-access (MAC) and broadcast channels. There are fewer results for these channels, but the state of knowledge mirrors that for discrete memoryless channels; although the MAC is fairly well-understood, the broadcast channel is not.

Jahn [89] found the randomized coding capacity of the arbitrarily varying multiple-access channel as well as an achievable rate region for the general broadcast channel using a code based on that of Marton [109]. He showed that if the interior of the rate region of the MAC under deterministic coding is nonempty then Ahlswede's elimination arguments show that the deterministic capacity region under average error is equal to the randomized coding capacity region under maximal error.

Gubner [72] extended the program of Csiszár and Narayan to the MAC and showed that a corresponding notions of symmetrizability hold for the MAC. In particular, the capacity region has empty interior under certain symmetrizability conditions, and he conjectured that these conditions were also necessary. Ahlswede and Cai later proved the conjecture true [15]. Together with Jahn's result, this completely characterized the capacity region of the MAC without constraints.

Gubner also used a different code construction and decoding rule which allowed him to extend the MAC coding theorem to the case of state constraints [73] and in particular to additive channels [74]. For deterministic coding under average error with state constraints, time sharing between coding strategies may not be possible. To see this, suppose that the encoder uses code A for the first half of the blocklength and code B for the second half. The jammer could operate using cost 2Λ in the first half and 0 in the second half, which means the code A would have to be designed for twice the cost constraint. Time sharing is what justifies taking the convex closure of the rate region for the discrete memoryless MAC, and Gubner and Hughes showed that the rate region for the MAC under state constraints need not be convex in general [75]. For the Gaussian multiple-access AVC, La and Anantharam [100] relaxed the channel model to limit the jamming strategies, which avoids the pathological behavior in Gubner and Hughes' result.

For the broadcast channel, the only results after Jahn are in a very recent paper by Hof and Bross [82]. They found an achievable rate region for degraded message sets with and without input and state constraints using techniques similar to Csiszár and Narayan, Gubner, and Ahlswede and Cai. The symmetrizability conditions for the general broadcast channel become particularly baroque, and it is unclear whether further extensions to general AVC networks under deterministic coding will be amenable to analysis.

1.3.8 Continuous alphabets

There are a few continuous channel models which have been studied in the context of arbitrarily varying channels. A Poisson channel version of the AVC has been studied by Bross and Shamai [35] under randomized and deterministic coding. This thesis does not address this channel model. The Gaussian AVC, proposed by Hughes and Narayan [86], is an additive white Gaussian noise (AWGN) channel with a power-constrained but arbitrary additive interference signal $\mathbf{s}$. We will investigate this channel and variants in Chapter 5. Finally, a general framework for studying AVC coding problems under deterministic coding has been given by Csiszár [42].

1.4 Other models of channel uncertainty

The AVC model assumes an adversarial and non-statistical model for the state selection, which differs from some other approaches to modeling interference. The AVC can model nonstationary channels, which is important in applications in which the environment may change on a time-scale unknown to the designer. Another non-stationary channel model is the *compound channel* model [27], which also consists of a set $\{W(y|x,s) : s \in \mathcal{S}\}$. In the compound channel a single $s \in \mathcal{S}$ governs the transmission over the entire block. This can model a link whose gain is unknown but constant, for example. Although the compound channel is not stationary, it is conditionally stationary, which is helpful in proving coding theorems. Hughes studied the effect of interleaving codes for AVCs and the relationship to the compound channel [84].

A particularly well-studied channel model related to compound channels is the quasistatic fading channel

$$Y = HX + W \ , \tag{1.76}$$

where X and Y are vectors and H is a random matrix. Because the transmitter and receiver must agree upon a code and rate without advance knowledge of H, there is a chance that the channel cannot support the rate. In this case the channel is considered to be in *outage* [152]. However, in this model, H is given a probability distribution, and so the probability of outage is meaningful as a statement about H, whereas in the standard compound channel model there is no distribution on H.

One drawback to the AVC's loose modeling of the state sequence is that it does not easily encompass channels with memory or known probabilistic dynamics. For example, inter-symbol-interference (ISI) channels are popular choices for modeling communication links in which channel memory provides a source of interference. By explicitly modeling the propagation properties of the communication medium, modulation and coding schemes can help mitigate the effects of this distortion.

The finite state channel (FSC) is an example of a channel with known dynamics. The the FSC the state is governed by a finite state machine whose transitions are given [67,33]. Recent work on these channels has focused on the benefits of feedback [36,149,120]. In an AVC, by contrast, the state dynamics are not given such structure, so there is a loss with respect to codes which can take advantage of known dynamics.

The other key aspect of the AVC model is the adversarial assumption on the interference, which is also taken up in two streams of works in the information theory and communications literature. The first set of works is on a game theoretic formulation of communication in which the transmitter and jammer use the mutual information of the channel as a payoff function. In this framework, the mutual information is assumed to have an operational significance *a priori* and is used as a proxy for the actual maximum rate of reliable communication. The first game-theoretic formulations are due to Blachman [26] and Dobrushin [51]. McEliece and collaborators [31, 111, 112] also studied jamming via a mutual information game in a scalar setting. The general vector case was fully studied by Baker and Chao [22].

There is extensive work on jamming interference in the communications literature (see for example [81, 108, 113]). One model of interest in the Gaussian setting is the so-called "correlated jamming" attack, in which the interference is a random process correlated with a transmitted random process. Early work on this model was done by Başar and Başar [21]. This framework was further studied by Médard [113] and more recent results have been found for multiple access [133] and multiantenna channels [92]. These works are similar to the mutual information games but allow for the jammer to know something about the transmitted signal. This may be a more appropriate model for wireless scenarios where the jammer can eavesdrop on the communication link.

One application area for adversarial channel models is in digital watermarking. In these models, the encoder takes a vector $\mathbf{T} \in \mathcal{T}^n$ called the **covertext** and a message m and produces a **stegotext** $\mathbf{X} \in \mathcal{X}^n$. An **attacker** uses the stegotext to produce a compromised signal $\mathbf{Y} \in \mathcal{Y}^n$. The decoder must recover the message m from $\mathbf{Y}^n$. The stegotext and compromised texts must satisfy distortion constraints:

$$d_e(\mathbf{T}, \mathbf{X}) = \sum_{i=1}^{n} d_e(T_i, X_i) \leq D_e \tag{1.77}$$

$$d_a(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} d_a(X_i, Y_i) \leq D_a \ , \tag{1.78}$$

where $d_e(\cdot, \cdot)$ and $d_a(\cdot, \cdot)$ are distortion measures. In these problems the encoder and decoder are typically allowed to use common randomness.

The encoder distortion constraint and attack distortion constraint can be used to constrain the allowable conditional distributions from $\mathbf{T}$ to $\mathbf{X}$ and $\mathbf{X}$ to $\mathbf{Y}$. In particular, the attacker is similar to a jammer that has knowledge of the transmitted codeword, and the capacity is related to AVCs under the nosy noise error model. We will return to a watermarking problem in the Gaussian setting in Chapter 5.

In Chapter 3 we will show that the randomized coding capacity for the AVC with nosy noise is given by C_{dep}. The expression for is similar to a rate-distortion function. For a fixed distribution P, the jammer tries to minimize the mutual information $I\left(P, V\right)$ over all channels $V \in \mathcal{W}_{dep}(\Lambda)$. In the rate-distortion setting, the encoder attempts to minimize the mutual information $I\left(P, V\right)$ over all test channels V satisfying the distortion constraint. An operational meaning for this minimization in the setting of channel coding was recently proposed by Agarwal, Sahai, and Mitter [1]. In their model the input distribution P is fixed. For any P-typical input $\mathbf{x}$, the channel may output any vector $\mathbf{y}$ such that

$$d(\mathbf{x}, \mathbf{y}) = \sum_i d(x_i, y_i) \leq D \ , \tag{1.79}$$

for some distortion function $d(\cdot, \cdot)$ and distortion level D. They proved that the capacity for this channel under randomized coding is equal to the rate distortion function $R_d(P, D)$.

In many instances the two channel definitions coincide, so the AVC with nosy noise is the distortion-constrained channel and the capacity expression can be found from prior results. For example, for additive channels we can define the state cost in terms of a difference distortion measure. Indeed, as in the watermarking example, it may be more natural to define the channel in terms of a distortion-constrained attack. Some of our later results on AVCs with nosy noise can be applied to the distortion-constrained channel model as well. However, we will show that not all AVCs can be modeled as channels with distortion constraints.

1.5 What is in this thesis?

The results in this thesis are motivated by questions of how to design robust and adaptive coding schemes for channels that model communication in sensor networks, ad-hoc networks, and cognitive radio. We will use the AVC model and variants to isolate these issues. We will primarily derive results for point-to-point communication over AVCs with a focus on quantifying the relationship between extra resources available to the encoder and decoder and the achievable rates and error probability. In particular, we will address how the number of keys $K(n)$ is related to the error probability, how limited feedback can let the encoder and decoder adapt the rate to the actual state sequence $\mathbf{s}$ governing the channel rather than settling for the worst case, and how side information at the encoder about some of the interference can be used to gain more robustness against jamming.

Table 1.1 on page 25 showed known results for unconstrained AVCs. In this thesis we will prove some corresponding results for constrained AVCs. In particular, we will:

- find bounds on the list-decoding capacities under average and maximal error for constrained AVCs (Chapter 2),
- find the randomized coding capacity under nosy noise (Chapter 3),
- show that the randomized coding capacities for maximal error and nosy noise are achievable with limited common randomness (Chapter 3),
- provide rateless constructions for maximal error and nosy noise (Chapter 4),
- show that the randomized coding capacity for the Gaussian version of the AVC is achievable with limited common randomness (Chapter 5),
- show how additional known interference can be used to improve a threshold for deterministic coding over Gaussian AVCs (Chapter 5).

Some of the results in this thesis have appeared in the proceedings of conferences and associated preprints. The results in Chapter 2 contains some results from [128, 127]. Chapter 3 is drawn from [130, 127], and Chapter 4 is taken from [130, 129, 62, 61]. Chapter 5 is taken from the papers [125, 126, 132].

In Chapter 2 we will provide results on list decoding for AVCs with input and state constraints. For maximal error the results are entirely analogous – we can achieve rates within $C_{\text{dep}}(\Gamma, \Lambda) - O(1/L)$ with lists of size L. For average error, the story is more complicated. As in the unconstrained case, there is a sufficiently large constant list size for which list decoding achieves the randomized coding capacity $C_r(\Gamma, \Lambda)$. For smaller lists, we can achieve lower rates, and in some cases the capacity may be smaller than the randomized coding capacity. This is analogous to deterministic coding for constrained AVCs where the capacity may be positive but strictly smaller than the randomized coding capacity.

In Chapter 3 we discuss two strategies for partial derandomization of AVCs that have been used in the literature – the elimination technique and a message authentication code. We describe these two strategies and use them to prove that a key size of $O(\log n)$ bits is sufficient to achieve the randomized coding capacity for constrained AVCs. For maximal error we use the elimination technique together with a randomized code based on permutations, and for AVCs with nosy noise we use the message authentication code with the maximal error list codes of Chapter 2. As by-products we show that the randomized coding capacity under nosy noise is $\hat{C}_r(\Gamma, \Lambda) = C_{\text{dep}}(\Gamma, \Lambda)$ and that for distortion constrained channels we can achieve the rate-distortion function with $O(\log n)$ bits.

We turn next to the impact of limited feedback in Chapter 4. In particular, we adopt a model in which every c channel uses the decoder obtains an estimate of the average channel and can send the encoder one bit to terminate transmission. That is, the decoder can either send an ACK (acknowledge) or NACK (not acknowledge) to inform the encoder if it has decoded or not. This coarse feedback is enough to let the encoder and decoder adapt their rate to the actual $\mathbf{s}$ governing the channel. This is particularly important in applications where the worst case state given by an adversarial assumption on the jammer gives significantly lower rates than those in more realistic channel conditions. Under maximal error and nosy noise we can derive similar coding strategies that achieve rates close to the empirical mutual information induced by $\mathbf{s}$.

All of the previous results have been for discrete channels with finite input, output, and state alphabets, but most popular information theoretic models for wireless communications use Gaussian channels. Chapter 5 address the Gaussian AVC. For the Gaussian AVC, the deterministic coding capacity exhibits a threshold phenomenon when the jammer can symmetrize the channel. We again show that $O(\log n)$ bits can achieve the randomized coding capacity and use this result to prove an achievable rate region for the Gaussian arbitrarily varying degraded broadcast channel. Finally, we look at deterministic coding when there are two sources of interference: one from the jammer and one from another system. We assume the latter is known non-causally at the transmitter, which can then use the interference to mask its own message. Although the encoder and decoder do not share a secret key, the extra interference known to the transmitter makes it more difficult for the jammer to symmetrize the channel and enlarges the capacity region. We can apply this coding scheme to watermarking systems and a recently proposed model for cognitive radio systems.

Although the results proved here are for point-to-point channels, they are motivated by problems that arise in networking. In particular, we can think of coding at the link level for AVCs as "insulating" the link from time variation in the channel quality caused by the actions of other links in a network. At a larger level, we could think of groups of cooperating links which must be insulated from other groups of links. These considerations may become important when modeling multiple networks of users competing for the same resources.

– *The Mikado*, W.S. Gilbert and A. Sullivan

Chapter 2

List decoding for discrete AVCs

2.1 Introduction

One way of relaxing channel coding problems is to consider *list decoding* [55, 156], in which the decoder is allowed to output a list of codewords of size no more than L and an error occurs if the transmitted codeword is not in the list. List decoding can be used as a component in more complicated coding systems [66], and we will use the codes constructed in this chapter in later chapters. The pioneering work of Sudan [147] on decoding Reed-Solomon codes led to a number of improved algorithms for list-decoding Reed-Solomon codes. The Guruswami-Sudan algorithm [80] has been extended by Koetter and Vardy to take advantage of soft information [96]. More complicated constructions have improved on the standard Reed-Solomon codes to give better decoding performance [119, 78, 79]. General combinatorial bounds on the limits of list decoding have also been studied [77, 76].

For AVCs with deterministic coding, the list-decoding capacities under maximal error C_L and average error $\overline{C}_L$ for unconstrained AVCs have been investigated. For maximal error, Ahlswede [4, 11] found that for any rate $R = C_{\text{dep}} - \epsilon$ is achievable under list decoding with list size $O(\epsilon^{-1})$. For the average probability of error criterion, the list coding capacity was found independently by Blinovsky, Narayan, and Pinsker [30, 29] and Hughes [85]. They extended the notion of symmetrizability to lists and showed that for channels with $C_{\text{std}} > 0$ there exists a finite list size L_{sym} such that $\overline{C}_L = C_{\text{std}}$ for $L > L_{\text{sym}}$. Furthermore, they show that the capacity is 0 using list decoding with list size smaller than or equal to L_{sym}. That is, $\overline{C}_L = 0$ for $L \leq L_{\text{sym}}$.

In this chapter we will extend these results to channels with state and input constraints. We follow the line of argument used by Ahlswede [11] for the maximal error case and by Hughes [85] for the average error case. For maximal error, the results are similar but in the average error case they are qualitatively different. For maximal error, we show that a rate $R = C_{\text{dep}}(\Gamma, \Lambda) - \epsilon$ is achievable using list decoding with list size $L = O(\epsilon^{-1})$.

For average error we define a function $\tilde{L}_{\text{sym}}(P, \Lambda)$, called the *weak symmetrizability* for type P. If P^* is the input distribution maximizing $C_{\text{std}}(\Gamma, \Lambda)$, then we show that $\overline{C}_L(\Gamma, \Lambda) = C_{\text{std}}(\Gamma, \Lambda)$ for list sizes $L > \tilde{L}_{\text{sym}}(P^*, \Lambda)$. For smaller list sizes we can achieve positive rates using codebooks of type P if $L > \tilde{L}_{\text{sym}}(P, \Lambda)$. For the converse, we define a function $L_{\text{sym}}(P, \Lambda)$ called the *strong symmetrizability* of the channel and show that the maximum rate achievable using list decoding is upper bounded by the maximum mutual information over all types with $L_{\text{sym}}(P, \Lambda) < L$. In particular, if $L_{\text{sym}}(P^*, \Lambda) < L$, then the list-decoding capacity under average error may be strictly smaller than $C_{\text{std}}(\Gamma, \Lambda)$. The behavior for constrained AVCs is thus different from that in the unconstrained case. This result is analogous to deterministic coding under average error, for which Csiszár and Narayan [46] showed that the capacity may be positive and strictly smaller than the randomized coding capacity $C_r(\Gamma, \Lambda) = C_{\text{std}}(\Gamma, \Lambda)$.

Deterministic codes for maximal error are connected to the design of codes in algebraic coding theory. The minimum distance of a linear code provides a bound on the number of errors that can be corrected in a constrained AVC setting. The generalized minimum distance [154,153] was used by Guruswami [76] to prove results on the list coding limits for linear codes. It may be interesting to look at these results via the symmetrizability of the associated AVCs.

2.2 List Decoding for Maximal Error

As we noted in the introduction, the capacity under maximal error using deterministic codes is still not known. For some AVCs this capacity is equivalent to finding the zero-error capacity of a corresponding DMC, which means this problem may be quite difficult in general. List decoding is a way of relaxing the design requirements for the AVC. For unconstrained AVCs, the capacity under list decoding was investigated by Ahlswede [5, 11] using hypergraph coloring arguments [7, 8]. In this section we generalize his result to the constrained AVCs using constant composition codes. Our arguments do not use the hypergraph formalism.

Theorem 10 (List decoding for maximal error)**.** *Let $\mathcal{W}$ be an arbitrarily varying channel with input and state cost functions $g(x)$ and $l(s)$. Then for any $\epsilon_1 > 0$ the rate*

$$R = C_{\text{dep}}(\Gamma, \Lambda) - \epsilon_1 \tag{2.1}$$

is achievable under maximal error using list decoding with list size

$$L = O\left(\frac{1}{\epsilon_1}\right) \ . \tag{2.2}$$

That is, the capacity $C_L(\Gamma, \Lambda)$ under maximal error using list decoding with list size L is bounded

$$C_{\text{dep}}(\Gamma, \Lambda) - O(L^{-1}) \leq C_L(\Gamma, \Lambda) \leq C_{\text{dep}}(\Gamma, \Lambda) \ . \tag{2.3}$$

The proof of Theorem 10 is given in Section 2.2.4. We show in Lemma 3 that $C_{\text{dep}}(\Gamma, \Lambda) - \epsilon$ is achievable with list size L satisfying

$$L > \frac{6 \log |\mathcal{Y}|}{\epsilon} \ . \tag{2.4}$$

Example 2.1 – Bit-flipping under list decoding

Consider the bit flipping example of Example 1.4 on page 21. There we saw that

$$C_{\text{dep}}(\Lambda) = 1 - h_b(\Lambda) \ , \tag{2.5}$$

where $h_b(\cdot)$ is the binary entropy function. Figure 2.1 shows the rates achievable with the maximal error list codes. For deterministic coding under average error, Csiszár and Narayan have shown that $1 - h_b(\Lambda)$ is achievable. That is, under average error, $\overline{C}_d(\Gamma, \Lambda) = C_{\text{std}}(\Gamma, \Lambda)$ is achievable with list size 1.

2.2.1 Preliminaries

Let $V(y|x)$ be a channel. Then we denote the (V, ϵ)-shell of a length n sequence $\mathbf{x}$ by

$$T_V^{\epsilon}(\mathbf{x}) = \{\mathbf{y} \in \mathcal{Y}^n : d_{\max}\left(T_{\mathbf{yx}}, VT_{\mathbf{x}}\right) < \epsilon\} \ , \tag{2.6}$$

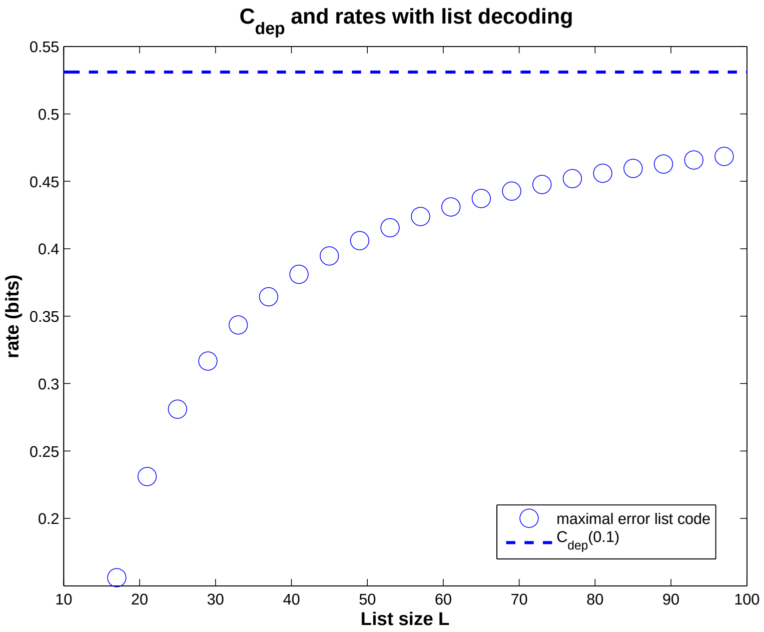


Figure 2.1: Rates achievable under maximal error with list decoding for bit-flipping channels with $\Lambda = 0.1$. The gap to capacity shrinks quite slowly with the list size.

Where $d_{\max}(\cdot, \cdot)$ is the ℓ_∞ distance between the probability distributions:

$$d_{\max}(F, G) = \max_{z} |F(z) - G(z)| \ . \tag{2.7}$$

We know the following bounds [44, Section 1.2] [85, Section III.B]:

$$|\mathcal{T}_n(P)| \geq (n+1)^{|\mathcal{X}|} \exp(nH(P)) \tag{2.8}$$

$$V^n(\{\mathbf{y} : T_{\mathbf{xy}} = P_{XY}\} | \mathbf{x}) \leq \exp(-nD(P_{XY} \ \| \ V \times P_X)) \ . \tag{2.9}$$

$$V^n(\{\mathbf{y} : T_{\mathbf{xyz}} = P_{XYZ}\} | \mathbf{z}) \leq \exp(-nI(X \ \wedge \ Y | Z)) \ . \tag{2.10}$$

We can define an AVC-version of an ϵ-shell:

$$V_{\mathbf{x},\mathbf{s}}(x,y) = \frac{1}{n} \sum_{k:x_k=x} W(y|x,s_k)$$

$$T_W^{\epsilon}(\mathbf{x}|\mathbf{s}) = \{\mathbf{y} \in \mathcal{Y}^n : d_{\max}\left(T_{\mathbf{xy}}, V_{\mathbf{x},\mathbf{s}}\right) < \epsilon\} \ . \tag{2.11}$$

We can take the union over all $\mathbf{s} \in \mathcal{S}^n(\Lambda)$ to get the set of all $\mathbf{y}$ sequences that could have been generated from $\mathbf{x}$ and a state $\mathbf{s}$ satisfying the cost constraint:

$$T_{\mathcal{W}}^{\epsilon}(\mathbf{x}) = \bigcup_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} T_W^{\epsilon}(\mathbf{x}|\mathbf{s}) \ . \tag{2.12}$$

For a received sequence $\mathbf{y}$, we must find the possible $\mathbf{x}$ sequences for which there exists a state vector $\mathbf{s} \in \mathcal{S}^n(\Lambda)$ such that $\mathbf{x}$ and $\mathbf{s}$ could have generated $\mathbf{y}$. For an input distribution $P(x)$ and channel $V(y|x)$ we can define an output distribution $P'(y)$ and "reverse channel" $V'(x|y)$ using Bayes rule:

$$P'(y) = \sum_x P(x)V(y|x) \tag{2.13}$$

$$V(y|x)P(x) = V'(x|y)P'(y) \ . \tag{2.14}$$

The decoder can use the empirical output type $T_{\mathbf{y}}$ to find a candidate set of channels $V(y|x)$ consistent with the observed sequence. We define

$$\mathcal{V}_P^{\epsilon}(\mathbf{y}) = \{V \in \mathcal{W}_{dep}(P,\Lambda) \cap \mathcal{P}_n(\mathcal{Y}|\mathcal{X}) : d_{\max}\left(P', T_{\mathbf{y}}\right) < \epsilon\} \ . \tag{2.15}$$

For a fixed $\mathbf{y}$ we will bound both the size of the set

$$T_{V'}^{\epsilon}(\mathbf{y}) = \{\mathbf{x} \in \mathcal{X}^n : d_{\max}\left(T_{\mathbf{xy}}, V'T_{\mathbf{y}}\right) < \epsilon\} \ , \tag{2.16}$$

for $V \in \mathcal{V}_P^\epsilon(\mathbf{y})$, and the size of the union of all $T_{V'}^\epsilon(\mathbf{y})$ where $V \in \mathcal{V}_P^\epsilon(\mathbf{y})$.

2.2.2 List codes with exponential list size

The following lemma provides some combinatorial bounds on the cardinalities of various typical sets and shells. This will allow us to prove Lemma 2, which constructs a list-decodable code with a list size that is exponential in the blocklength. We will use this code as the basis for constructing the code in Lemma 3, which has constant list size.

Lemma 1. *Let $\mathcal{W}$ be an AVC with state cost function $l(s)$ and cost constraint Λ. For any $P \in \mathcal{P}(\mathcal{X})$ with $\min_x P(x) > 0$, $\alpha > 0$, and any ϵ_2 satisfying $0 < \epsilon_2 < \min_a P(a)$ there exists an n sufficiently large such that the following statements all hold:*

1. *For $\mathbf{x} \in \mathcal{X}^n$ and $\mathbf{s}$ such that $l(\mathbf{s}) \leq \Lambda$ we have for some $E(\epsilon_2) > 0$:*

$$\mathbb{P}\left(T_W^{\epsilon_2}(\mathbf{x}|\mathbf{s})|\mathbf{x},\mathbf{s}\right) \geq 1 - \exp\left(-nE(\epsilon_2)\right) \ . \tag{2.17}$$

2. *For $P \in \mathcal{P}_n(\mathcal{X})$, channel $V \in \mathcal{W}_{dep}(P)$, $\mathbf{y} \in \mathcal{Y}^n$, and $\epsilon_2 > 0$,*

$$|T_{V'}^{\epsilon_2}(\mathbf{y})| \leq \exp\left(n\left(\sum_y H\left(V'(x|y)\right)T_{\mathbf{y}}(y) + O(\epsilon_2 \log \epsilon_2^{-1})\right)\right) \tag{2.18}$$

$$\left|\bigcup_{V \in \mathcal{V}_P^{\alpha\epsilon_2}(\mathbf{y})} T_{V'}^{\epsilon_2}(\mathbf{y})\right| \leq \exp\left(n\left(\max_{V \in \mathcal{V}_P^{\epsilon_2}(\mathbf{y})} \sum_y H\left(V'(x|y)\right)P'(y) + O(\epsilon_2 \log \epsilon_2^{-1})\right)\right) \tag{2.19}$$

3. *For sufficiently small* $\epsilon_2 > 0$, $\mathbf{x} \in T_P$, *and* $\mathbf{y} \in T_{\mathcal{W}}^{\epsilon_2}(\mathbf{x})$ *we have*

$$\mathbf{x} \in \bigcup_{V \in \mathcal{V}_P^{\epsilon_2}(\mathbf{y})} T_{V'}^{(|\mathcal{X}|+1)\epsilon_2}(\mathbf{y}) \tag{2.20}$$

$$\mathbf{y} \in \bigcup_{V \in \mathcal{V}_P^{|\mathcal{X}|\epsilon_2}(\mathbf{y})} T_V^{\epsilon_2}(\mathbf{x}) \ . \tag{2.21}$$

Proof. We take up the different items in turn.

1. Fix sequences $\mathbf{x}$ and $\mathbf{s}$ with $l(\mathbf{s}) \leq \Lambda$. Let $\{Y_i : i \in [n]\}$ be independent random variables with distribution $\{W(\cdot|x_i, s_i)\}$. Let $g_{(a,b)}(Y_1, \ldots, Y_n) = N(a, b|\mathbf{x}, Y_1^n)$. Then

$$\mathbb{E}[g_{(a,b)}(Y_1, \ldots, Y_n)] = \sum_{k : x_k = a} W(b|a, s_k)$$

If $\{\tilde{Y}_i\}$ are independent copies of $\{Y_i : i = 1, \ldots, n\}$, then we have

$$\left| g_{(a,b)}(Y_1, \ldots, Y_i, \ldots, Y_n) - g_{(a,b)}(Y_1, \ldots, \tilde{Y}_i, \ldots, Y_n) \right| \leq 1 \qquad a.s. \ .$$

By standard concentration inequalities [49, Corollary 2.4.14], for any $\epsilon_2 > 0$,

$$\mathbb{P}\left(\left| g_{(a,b)}(Y_1^n) - \mathbb{E}[g_{(a,b)}(Y_1^n)] \right| \geq n\epsilon_2 \right) \leq \exp\left(-nD\left(\frac{1+\epsilon_2}{2} \ \Big\| \ \frac{1}{2} \right) \right) \ . \tag{2.22}$$

Taking a union bound over all $(a, b) \in \mathcal{X} \times \mathcal{Y}$ in (2.22) shows that there exists a function $E(\epsilon_2) > 0$ so that for n sufficiently large

$$\mathbb{P}\left(T_W^{\epsilon_2}(\mathbf{x}|\mathbf{s}) | \mathbf{x}, \mathbf{s} \right) \geq 1 - \exp\left(-nE(\epsilon_2) \right) \ .$$

2. For input distribution P and channel V we can define V' via (2.14). Equation (2.18) then follows from [44, Lemma 2.13].

   To prove (2.19) note that by (2.15) there are at most $(n+1)^{|\mathcal{X}|\cdot|\mathcal{Y}}$ channels $V \in \mathcal{V}_P^{\alpha\epsilon_2}(\mathbf{y})$. For any such V we have the bound (2.18), so a union bound yields (2.19).

3. Since $\mathbf{y} \in T_{\mathcal{W}}^{\epsilon_2}(\mathbf{x})$, from (2.12) we know there exists an $\mathbf{s} \in \mathcal{S}^n(\Lambda)$ such that $\mathbf{y} \in T_{W,\mathbf{s}}^{\epsilon_2}(\mathbf{x})$. We define the channel

$$\begin{aligned} V(b|a) &= \frac{1}{N(a|\mathbf{x})} \sum_{k : x_k = a} W(b|a, s_k) \\ &= \sum_s W(b|a,s) \frac{N(a,s|\mathbf{x},\mathbf{s})}{N(a|\mathbf{x})} \ . \end{aligned}$$

   Therefore $V \in \mathcal{W}_{dep}$ and $\mathbf{y} \in T_V^{\epsilon_2}(\mathbf{x})$. We claim that $V \in \mathcal{V}_P^{|\mathcal{X}|\epsilon_2}(\mathbf{y})$.

   We can now bound $d_{\max}\left(T_{\mathbf{xy}}, P(x)V(y|x)\right)$ by using (2.11) and the fact that $P(a) = n^{-1}N(a|\mathbf{x})$:

$$\begin{aligned} d_{\max}\left(T_{\mathbf{xy}}, P(x)V(y|x)\right) &= d_{\max}\left(T_{\mathbf{xy}}, \frac{1}{n} \sum_{k : x_k = a} W(y|a, s_k)\right) \\ &\leq \epsilon_2 \ . \end{aligned} \tag{2.23}$$

   This proves that $\mathbf{y} \in T_V^{\epsilon_2}(\mathbf{x})$. We must also show that $V \in \mathcal{V}_P^{|\mathcal{X}|\epsilon_2}(\mathbf{y})$. Marginalizing (2.23) over $\mathcal{X}$ we obtain:

$$d_{\max}\left(T_{\mathbf{y}}, P'(y)\right) \leq |\mathcal{X}|\epsilon_2 \ , \tag{2.24}$$

   So $V \in \mathcal{V}_P^{|\mathcal{X}|\epsilon_2}(\mathbf{y})$, yielding (2.21).

To show (2.20), let V be a channel such that $\mathbf{y} \in T_V^{\epsilon_2}(\mathbf{x})$.

$$\begin{aligned} d_{\max}\left(T_{\mathbf{xy}}, P(x)V(y|x)\right) &= d_{\max}\left(T_{\mathbf{xy}}, T_{\mathbf{x}}V(y|x)\right) \\ &\leq \epsilon_2 \ . \end{aligned} \tag{2.25}$$

Now, marginalizing this over $\mathcal{X}$ allows us to bound the distance between $T_{\mathbf{x,y}}$ and $T_{\mathbf{y}}V'(x|y)$:

$$\begin{aligned} d_{\max}\left(T_{\mathbf{xy}}, T_{\mathbf{y}}V'(x|y)\right) &\leq d_{\max}\left(T_{\mathbf{x,y}}, P'(y)V'(x|y)\right) + d_{\max}\left(P'(y)V'(x|y), T_{\mathbf{y}}V'(x|y)\right) \\ &\leq \epsilon_2 + |\mathcal{X}|\epsilon_2 \ . \end{aligned} \tag{2.26}$$

Thus we have shown $\mathbf{x} \in T_{V'}^{(|\mathcal{X}|+1)\epsilon_2}(\mathbf{y})$.

□

Using this lemma we can prove the existence of list-decodable codes for maximal error with exponential list size. The codebook is the entire set of typical sequences T_P and the list is the union of ϵ-shells under the different state sequences whose cardinality is bounded by (2.19). The decoder outputs a list that is the union of shells.

Lemma 2. *Let $\mathcal{W}$ be an AVC with state cost function $l(s)$ and cost constraint Λ. For any $P \in \mathcal{P}(\mathcal{X})$ with $\min_x P(x) > 0$ and $\epsilon_3 > 0$ there is an n sufficiently large such that for $P \in \mathcal{P}_n(\mathcal{X})$ there is an (n, N, L) list-decodable code $\mathcal{C}$ with*

$$N \geq \exp\left(n\left(H(P(x)) - o(1)\right)\right) \tag{2.27}$$

$$L \leq \exp\left(n\left(\max_{V \in \mathcal{W}_{dep}(\Lambda)} H(V'(x|y)|P'(y)) + O(\epsilon_3 \log \epsilon_3^{-1})\right)\right) \tag{2.28}$$

$$\varepsilon_L \leq \exp(-nE(\epsilon_3)) \ , \tag{2.29}$$

where $o(1)$ *is a term that goes to* 0 *as* $n \to \infty$.

Proof. Choose the codewords of the code to be all $\mathbf{x} \in T_P$. For each channel output $\mathbf{y}$ the decoder outputs the list

$$\bigcup_{V \in \mathcal{V}_P^{\epsilon_3}(\mathbf{y})} T_{V'}^{(|\mathcal{X}|+1)\epsilon_3}(\mathbf{y}) \ . \tag{2.30}$$

Equation (2.27) follows from (2.8). The bound (2.28) on the list size follows from (2.19) in Lemma 1 part 2 with $\epsilon_2 = \epsilon_3$ and $\alpha = 1/(|\mathcal{X}| + 1)$. To bound the error in (2.29) note that with probability upper bounded by $\exp(-nE(\epsilon_3))$ we have $\mathbf{y} \in T_{\mathcal{W}}^{\epsilon_3}(\mathbf{x}|\mathbf{s})$ (by Lemma 1 part 1) and hence by Lemma 1 part 3 we have $\mathbf{x} \in T_{V'}^{(|\mathcal{X}+1)\epsilon_3}(\mathbf{y})$ for some $V \in \mathcal{V}_P^{\epsilon_3}(\mathbf{y})$. □

2.2.3 List codes with constant list size

We will now show that for any input distribution P we can construct a list decodable code that achieves a rate

$$\min_{V \in \mathcal{W}_{dep}(\Lambda)} I\left(P, V\right) - \epsilon \tag{2.31}$$

with lists of size $O(\epsilon^{-1})$. Given a small gap ϵ from $\min_{V \in \mathcal{W}_{dep}(\Lambda)} I\left(P, V\right)$, we construct the code by sampling codewords from the code of Lemma 2. By choosing a sufficiently large list size we can show that with high probability the samples will form a list-decodable code with probability of error going to 0 as the blocklength n goes to ∞.

Lemma 3. *Let* $\mathcal{W}$ *be an AVC with state cost function* $l(s)$ *and cost constraint* Λ. *For any* $\epsilon_4 > 0$ *and* $P \in \mathcal{P}(\mathcal{X})$ *with* $\max_x P(x) > 0$, *for* n *sufficiently large there exists a* *list code with codewords of type P, rate*

$$R = \min_{V \in \mathcal{W}_{dep}(P,\Lambda)} I(P,V) - \epsilon_4 \ , \tag{2.32}$$

list size

$$L < \left\lfloor \frac{6 \log |\mathcal{Y}|}{\epsilon_4} \right\rfloor + 1 \ , \tag{2.33}$$

and error

$$\varepsilon_L \leq \exp(-nE(\epsilon_4)) \ , \tag{2.34}$$

where $E(\epsilon_4) > 0$.

Proof. Fix $P \in \mathcal{P}(\mathcal{X})$ and $\epsilon_4 > 0$. For any $\epsilon_5 > 0$, Lemma 2 says that we can choose n sufficiently large such that there exists an (n, N_0, L_0) list code (ϕ_0, ψ_0) with codebook $\mathcal{C}_0 = \{(\mathbf{u}(i), D_i) : i \in [N_0]\}$ satisfying (2.27)–(2.29) with $\epsilon_3 = \epsilon_5$. We will use this code to construct an (n, N, L) list code of the desired rate and list size.

Let $\mathbf{B} = \{\mathbf{x}(j) : j \in [N]\}$ be a collection of N iid random variables uniformly distributed on the set $\mathcal{C}_0$, where $N < |\mathcal{Y}|^n$. Define an encoding map by $\phi(j) = \mathbf{x}(j)$ and a decoding map by $\psi(\mathbf{y}) = \mathbf{B} \cap \psi_0(\mathbf{y})$. Note that the size of $\psi(\mathbf{y})$ depends on $\mathbf{y}$ and may be different for each $\mathbf{y}$. We will show that for n sufficiently large and L chosen according to (2.33) we can choose R according to (2.32) so that with high probability $\psi(\mathbf{y}) \leq L$ for all $\mathbf{y}$ and random selection will produce a (n, N, L) list-decodable code of rate R and error probability upper bounded by $\exp(-nE(\epsilon_4))$.

Fix $\mathbf{y} \in \mathcal{Y}^n$. We begin by looking at the expected size of the list $\psi(\mathbf{y})$:

$$\mathbb{E}\left[|\psi(\mathbf{y})|\right] \leq \frac{L_0}{N_0} \ . \tag{2.35}$$

This expectation is over the sampling from $\mathcal{C}_0$. For a fixed $\mathbf{y}$, the size of the list $|\psi(\mathbf{y})|$ is given by the sum of indicator functions $\mathbf{1}(\mathbf{y} \in D_j)$ over codewords $\mathbf{x}(j)$ in the sampled codebook $\mathbf{B}$:

$$|\psi(\mathbf{y})| = \sum_{j=1}^{N} \mathbf{1}(\mathbf{y} \in D_j) \ . \tag{2.36}$$

Because the codewords in $\mathbf{B}$ are selected in an iid manner, we can bound the probability that $|\psi(\mathbf{y})| > L$ using Sanov's Theorem [41, Theorem 12.4.1] on the random variables $\{\mathbf{1}(\mathbf{y} \in D_j)\}$. Letting $\nu = (N+1)^2$, the theorem gives

$$\mathbb{P}\left(\frac{1}{N} \sum_{j=1}^{N} \mathbf{1}(\mathbf{y} \in D_j) > \frac{L}{N} \right) \le \exp\left(-N \left(D\left(\frac{L}{N} \,\middle\|\, \frac{L_0}{N_0} \right) \right) + \log \nu \right) \ . \tag{2.37}$$

Let G denote the term inside the exponent. Then

$$G = -L \log \frac{L/N}{L_0/N_0} - N\left(1 - \frac{L}{N}\right) \log \frac{1 - L/N}{1 - L_0/N_0} + \log \nu \ . \tag{2.38}$$

To deal with the second term we use the inequality $-(1-a)\log(1-a) \le 2a$ (for small a) on the term $(1 - L/N)\log(1 - L/N)$ and discard the small negative term $(1 - L/N)\log(1 - L_0/N_0)$.

$$G \le -L \log \frac{L/N}{L_0/N_0} + N\left(2\frac{L}{N}\right) + \log \nu \tag{2.39}$$

$$= L \log \frac{L_0}{N_0} + L \log \frac{N}{L} + 2L + \log \nu \ . \tag{2.40}$$

Now, from Lemma 2 we have

$$\log \frac{L_0}{N_0} \leq n \left(\max_{V \in \mathcal{W}_{dep}(\Lambda)} H(V'(x|y)|P'(y)) - H(P(x)) + O(\epsilon_5 \log \epsilon_5^{-1}) \right)$$
$$= -n \left(\min_{V \in \mathcal{W}_{dep}(\Lambda)} I\left(P, V\right) - O(\epsilon_5 \log \epsilon_5^{-1}) \right) \ . \tag{2.41}$$

Since $N < |\mathcal{Y}|^n$ we have

$$\log \nu = 2 \log(N+1) \tag{2.42}$$
$$\leq 2n \log |\mathcal{Y}| \ . \tag{2.43}$$

Then we have the bound:

$$G \leq -nL \left(\min_{V \in \mathcal{W}_{dep}(\Lambda)} I\left(P, V\right) - \frac{1}{n} \log \frac{N}{L} - O(\epsilon_5 \log \epsilon_5^{-1}) - \frac{2}{L} \log |\mathcal{Y}| \right) + 2L \ .$$

For any constant $\epsilon_5 > 0$ we can choose n large enough such that

$$G \leq -nL \left(\min_{V \in \mathcal{W}_{dep}(\Lambda)} I\left(P, V\right) - R - O(\epsilon_5 \log \epsilon_5^{-1}) - \frac{2}{L} \log |\mathcal{Y}| \right) \ . \tag{2.44}$$

Now we take a union bound over all $\mathbf{y}$ in (2.37) to get

$$\mathbb{P} \left(\bigcup_{\mathbf{y} \in \mathcal{Y}^n} \left\{ \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}(\mathbf{y} \in D_j) > \frac{L}{N} \right\} \right)$$
$$\leq \exp \left(-n \left(L \left(\min_{V \in \mathcal{W}_{dep}(\Lambda)} I\left(P, V\right) - R - O(\epsilon_5 \log \epsilon_5^{-1}) \right) - 3 \log |\mathcal{Y}| \right) \right) \ . \tag{2.45}$$

For any $\epsilon_4 > 0$ we can choose ϵ_5 sufficiently small so that the $\epsilon_4/2 = O(\epsilon_5 \log \epsilon_5^{-1})$ and the list size L according to:

$$L > \frac{6 \log |\mathcal{Y}|}{\epsilon_4} . \tag{2.46}$$

Then if N is chosen such that R satisfies (2.32) the exponent is positive. Then for sufficiently large n the probability that the construction yields a (n, N, L) list-decodable code can be made arbitrarily close to 1. Since the codebook constructed here is a subset of the original code of exponential list size, the maximal error bound follows from Lemma 2. □

2.2.4 Proof of Theorem 10

Using Lemma 3 with the input distribution P maximizing

$$C_{\mathrm{dep}}(\Gamma, \Lambda) = \max_{P \in \mathcal{I}(\Gamma)} \min_{V \in \mathcal{W}_{dep}(\Lambda)} I\left(P, V\right) \tag{2.47}$$

yields the result in Theorem 10.

Proof of Theorem 10. Let $P^* \in \mathcal{P}(\mathcal{X})$ be given by

$$P^* = \operatorname*{argmax}_{P \in \mathcal{I}(\Gamma)} \min_{V \in \mathcal{W}_{dep}(P,\Lambda)} I\left(P, V\right) . \tag{2.48}$$

Then $C_{\mathrm{dep}}(\Gamma, \Lambda) = I\left(P^*, V\right)$.

Fix $\epsilon_1 > 0$. If $\min_x P^*(x) > 0$ then Lemma 3 says that for n sufficiently large there exists an (n, N, L) list decodable code with error as small as we like. If $\min_x P^*(x) = 0$ then by the continuity of the mutual information for any ϵ_1 there exists a δ such that we for P with $d_{\max}\left(P, P^*\right) < \delta$ we have

$$\min_{V \in \mathcal{W}_{dep}(P,\Lambda)} I\left(P, V\right) \geq C_{\mathrm{dep}}(\Gamma, \Lambda) - \epsilon_1/2 \tag{2.49}$$

We can choose a P such that $\min_x P(x) > 0$ and $P \in \mathcal{I}(\Gamma)$, so Lemma 3 says that there exist codes achieving $\min_{V \in \mathcal{W}_{dep}(P,\Lambda)} I(P, V) - \epsilon_1/2$ with list size $O(\epsilon_1^{-1})$ whose error can be made as small as we like. □

2.3 List Decoding for Average Error

In the case where we simultaneously allow list decoding and measure performance by the average error over the codebook, we can achieve all rates below $C_{\mathrm{std}}(\Gamma, \Lambda)$ with finite list sizes. To state our main result we require some additional definitions.

Recall that for an integer m we defined $\mathcal{U}_{\mathrm{sym}}(m)$ as the set of channels $U : \mathcal{X}^m \to \mathcal{S}$ that symmetrize the AVC $\mathcal{W}$ in (1.37):

$$\mathcal{U}_{\mathrm{sym}}(m) = \Bigg\{ U(s|x^m) : V(y|x, x_1, \ldots, x_m) = \sum_s W(y|x,s) U(s|x_1, x_2, \ldots, x_m) \\ \text{is symmetric} \Bigg\} . \qquad (2.50)$$

The jammer can use a channel $U \in \mathcal{U}_{\mathrm{sym}}(m)$ to generate a state sequence from m codewords. The average channel $V = WU$ is a symmetric channel with $m+1$ inputs.

For a distribution $P \in \mathcal{P}(\mathcal{X})$ we define **strong symmetrizing cost** $\lambda_m(P)$ to be the smallest expected cost of a channel $U(s|x^m)$ that symmetrizes the AVC $\mathcal{W}$ with arbitrarily correlated inputs which have marginal distributions P:

$$\lambda_m(P) = \min_{U \in \mathcal{U}_{\mathrm{sym}}(m)} \max_{\overline{P} \in \mathcal{P}(\mathcal{X}^m) : P_i = P} \sum_{x^m} \sum_s \overline{P}(x^m) U(s|x^m) l(s) . \qquad (2.51)$$

Here $\overline{P}$ is any joint distribution on $\mathcal{X}^m$ whose marginal distributions are equal to P. We call an AVC **strongly m-symmetrizable** if $\lambda_m(P) \leq \Lambda$. We define the **strong symmetrizability** $L_{\mathrm{sym}}(P, \Lambda)$ of the channel under input P to be the largest integer

m such that $\lambda_m(P) < \Lambda$. That is,

$$L_{\text{sym}}(P, \Lambda) = \max \{m : \lambda_m(P) \leq \Lambda\} \ . \tag{2.52}$$

If a codebook has codewords of type P and $L \leq L_{\text{sym}}(P, \Lambda)$ for a list-decodable code of list size L, then the jammer can choose any L codewords from the codebook and a channel $U \in \mathcal{U}_{\text{sym}}(L)$ to generate a state sequence. From (2.51) we can show that this state sequence will satisfy the cost constraint with probability going to 1 as the blocklength increases. Lemma 6 shows that such a strategy will lead to a average probability of error that does not go to 0 with the blocklength.

We define the **weak symmetrizing cost** $\tilde{\lambda}_m(P)$ to be the smallest expected cost of a channel $U(s|x^m)$ that symmetrizes the AVC $\mathcal{W}$ with independent inputs:

$$\tilde{\lambda}_m(P) = \min_{U \in \mathcal{U}_{\text{sym}}(m)} \sum_{x^m} \sum_{s} P^m(x^m) U(s|x^m) l(s) \ , \tag{2.53}$$

where P^m is the product distribution $P \times P \times \cdots \times P$. We call an AVC **weakly m-symmetrizable** if $\tilde{\lambda}_m(P) \leq \Lambda$. Similarly, the **weak symmetrizability** $\tilde{L}_{\text{sym}}(P, \Lambda)$ is the largest integer m such that $\tilde{\lambda}_m(P) \leq \Lambda$. That is,

$$\tilde{L}_{\text{sym}}(P, \Lambda) = \max \left\{m : \tilde{\lambda}_m(P) \leq \Lambda\right\} \ . \tag{2.54}$$

For the achievability arguments, Lemma 9 shows that we can find a decoding rule that outputs a list of size no larger than $\tilde{L}_{\text{sym}}(P, \Lambda) + 1$.

Because the strong symmetrizing cost includes a maximization over the joint distribution of the inputs to the channel U, we have $\lambda_m(P) \geq \tilde{\lambda}_m(P)$ in general. This means that $L_{\text{sym}}(P, \Lambda) \leq \tilde{L}_{\text{sym}}(P, \Lambda)$.

We will define for convenience the function

$$I\left(P, \Lambda\right) = \min_{V \in \mathcal{W}_{std}(\Lambda)} I\left(P, V\right) \ . \tag{2.55}$$

We then have

$$C_{\text{std}}(\Gamma, \Lambda) = \max_{P \in \mathcal{I}(\Gamma)} I\left(P, \Lambda\right) \ . \tag{2.56}$$

Let P^* be the input distribution maximizing $C_{\text{std}}(\Gamma, \Lambda)$.

Our main result in this section is a partial characterization of the capacity of constrained AVCs under list decoding. For each list size L we can compute the distributions P for which $L > L_{\text{sym}}(P, \Lambda)$ and $L > \tilde{L}_{\text{sym}}(P, \Lambda)$. The strong symmetrizability $L_{\text{sym}}(P, \Lambda)$ gives a converse – the capacity $C_L(\Gamma, \Lambda)$ cannot exceed $I(P)$ for P satisfying $L > L_{\text{sym}}(P, \Lambda)$. The weak symmetrizability $\tilde{L}_{\text{sym}}(P, \Lambda)$ gives an achievable region – for P with $L > \tilde{L}_{\text{sym}}(P, \Lambda)$ the rates $I\left(P, \Lambda\right)$ are achievable and the capacity $C_L(\Gamma, \Lambda)$ is at least as large as $I\left(P, \Lambda\right)$. If $L > \tilde{L}_{\text{sym}}(P^*, \Lambda)$ then the capacity achieving input distribution P^* can be used and $C_L(\Gamma, \Lambda) = C_{\text{std}}(\Gamma, \Lambda)$

Theorem 11 (List decoding for average error)**.** *Let $\mathcal{W}$ be an arbitrarily varying channel with input and state cost functions $g(\cdot)$ and $l(\cdot)$. If L is such that the maximum weak symmetrizing cost $\max_{P \in \mathcal{I}(\Gamma)} \tilde{\lambda}_L(P) > \Lambda$ then we have the following lower bound on $\overline{C}_L(\Gamma, \Lambda)$.*

$$\overline{C}_L(\Gamma, \Lambda) \geq \max_{P \in \mathcal{I}(\Gamma) : \tilde{L}_{\text{sym}}(P, \Lambda) < L} I(P) \ . \tag{2.57}$$

If L is such that $\max_{P \in \mathcal{I}(\Gamma)} \lambda_L(P) < \Lambda$ then we have the following upper bound on

$\overline{C}_L(\Gamma, \Lambda)$.

$$\overline{C}_L(\Gamma, \Lambda) \leq \max_{P \in \mathcal{I}(\Gamma) : L_{\mathrm{sym}}(P, \Lambda) < L} I(P) \tag{2.58}$$

$$\tag{2.59}$$

If P^* is the capacity achieving input distribution for $C_{\mathrm{std}}(\Gamma, \Lambda)$, then for list size $L > \tilde{L}_{\mathrm{sym}}(P^*, \Lambda)$ we have

$$\overline{C}_L(\Gamma, \Lambda) = C_{\mathrm{std}}(\Gamma, \Lambda) \ . \tag{2.60}$$

The proof of this theorem is given in Section 2.3.4 and parallels that of Csiszár and Narayan [46] for constrained AVCs. The decoding rule we use is an extension of the decoding rule used by Hughes [85] to the case with constraints. To show that $I(P, \Lambda)$ is achievable for $L > \tilde{L}_{\mathrm{sym}}(P, \Lambda)$, we use the fact the a random codebook with fixed type P enjoys certain properties (Lemma 10). We then show that $L > \tilde{L}_{\mathrm{sym}}(P, \Lambda)$ implies a certain "separation" of probability distributions (Lemma 8), which we can use to show that the decoding rule will only output at most $\tilde{L}_{\mathrm{sym}}(P, \Lambda) + 1$ codewords (Lemma 9). We can then use the codebook properties to show that $I(P, \Lambda)$ is achievable with fixed input type P (Lemma 11). The converse arguments are given in Section 2.3.2.

In general the strong and weak symmetrizabilities are different, so for a given list size L we may have

$$L_{\mathrm{sym}}(P, \Lambda) < L \leq \tilde{L}_{\mathrm{sym}}(P, \Lambda) \ . \tag{2.61}$$

In this case we cannot prove that $I(P, \Lambda)$ is achievable with codebooks whose codewords have type P. A similar problem arises in the arbitrarily varying multiple access channel [72], where correlation between the two users is allowed in the converse but is not used in the achievability arguments. For list decoding we conjecture that the list decoding capacity is dictated by the weak symmetrizability, which assumes independent inputs to the symmetrizing channel.

Conjecture 1 (Capacity conjecture for $\overline{C}_L$)**.** *Let $\mathcal{W}$ be an arbitrarily varying channel with input and state cost functions $g(\cdot)$ and $l(\cdot)$. Then the list coding capacity under average error for $\mathcal{W}$ with lists of size L is*

$$\overline{C}_L(\Gamma, \Lambda) = \max_{P \in \mathcal{I}(\Gamma) : \tilde{L}_{\mathrm{sym}}(P,\Lambda) < L} I(P) \ . \tag{2.62}$$

2.3.1 Finite symmetrizability

The following theorem shows that if $I(P)$ is positive, then $\tilde{L}_{\mathrm{sym}}(P, \Lambda)$ is finite. In particular, since $I(P^*, \Lambda)$ is finite, the theorem implies that if $C_{\mathrm{std}}(\Gamma, \Lambda) > 0$, then $\tilde{L}_{\mathrm{sym}}(P^*, \Lambda) < \infty$.

Theorem 12 (Finite symmetrizability)**.** *Let $\mathcal{W}$ be an arbitrarily varying channel with input and state cost functions $g(\cdot)$ and $l(\cdot)$. If $C_{\mathrm{std}}(\Gamma, \Lambda) = 0$ then $L_{\mathrm{sym}}(P, \Lambda) = \infty$ for all $P \in \mathcal{I}(\Gamma)$. If $C_{\mathrm{std}}(\Gamma, \Lambda) > 0$ then*

$$\tilde{L}_{\mathrm{sym}}(P, \Lambda) \leq \frac{\log(\min(|\mathcal{Y}|, |\mathcal{S}|))}{I(P, \Lambda)} \tag{2.63}$$

for all P such that $I(P, \Lambda) > 0$.

Proof. Suppose $C_{\mathrm{std}}(\Gamma, \Lambda) = 0$. Then for all $P \in \mathcal{I}(\Gamma)$ we have $I(P, \Lambda) = 0$. Without loss of generality, we may take $P(x) > 0$ for all $x \in \mathcal{X}$. For such P there exists a distribution $Q(s) \in \mathcal{Q}(\Lambda)$ such that the output distribution P_Y does not depend on the input x:

$$P_Y(y) = \sum_s W(y|x,s)Q(s) \qquad \forall x \in \mathcal{X} \ .$$

That is, the input and output are independent under $Q(s)$. Let $U(s|x^L) = Q(s)$ for all x_1^L. Then the average channel

$$V(y|x,x^L) = \sum_s W(y|x,s)U(s|x^L)$$

is symmetric in $(x, x_1, \ldots, x_L)$. Furthermore, for any distribution $\overline{P}(x_1^L)$ with marginal distributions equal to P

$$\begin{aligned}\lambda_L(P) = \sum_{s,x^L} \overline{P}(x^L)U(s|x^L)l(S) &= \sum_s Q(s)l(s) \\ &\leq \Lambda \ .\end{aligned}$$

Since this holds for all L, $\tilde{\lambda}_L(P) \leq \Lambda$ for all L and thus $L_{\mathrm{sym}}(P,\Lambda) = \infty$. Because $L_{\mathrm{sym}}(P,\Lambda) \geq \tilde{L}_{\mathrm{sym}}(P,\Lambda)$, this shows that the weak symmetrizability $\tilde{L}_{\mathrm{sym}}(P,\Lambda) = \infty$ as well.

Suppose now that $C_{\mathrm{std}}(\Gamma,\Lambda) > 0$. Let P be an input distribution for which $I(P,\Lambda) > 0$. Suppose that under P the channel is weakly L-symmetrizable. Therefore there is a channel $U(s|x_1^L)$ that symmetrizes W. Let $X_1, X_2, \ldots, X_L$ be independent with distribution P and let (S, X^L) be distributed according to the joint distribution $U(S|X^L)P(X_1)\cdots P(X_L)$. Then $(X, X^L) \to (X, S) \to Y$ is a Markov chain, so by the Data Processing inequality we have

$$\begin{aligned} I\left(XS \ \wedge \ Y\right) &\geq I\left(XX^L \ \wedge \ Y\right) \\ &\geq I\left(X \ \wedge \ Y\right) + \sum_{j=1}^{L} I\left(X_j \ \wedge \ Y\right) \\ &= (L+1) I\left(X \ \wedge \ Y\right) \ , \end{aligned}$$

where the last line follows from the symmetrizability. Now, subtracting $I\left(X \ \wedge \ Y\right)$ from both sides, we obtain the bound

$$I\left(S \ \wedge \ Y | \, X\right) \geq L \cdot I\left(X \ \wedge \ Y\right) \ .$$

This gives us a bound on the list size:

$$\begin{aligned} L &\leq \frac{I\left(S \ \wedge \ Y | \, X\right)}{I\left(X \ \wedge \ Y\right)} \\ &\leq \frac{\log(\min(|\mathcal{Y}|, |\mathcal{S}|))}{I(P, \Lambda)} \ . \end{aligned}$$

Since this bound holds for all L such that the AVC is weakly L-symmetrizable under distribution P, we can substitute $\tilde{L}_{\mathrm{sym}}(P, \Lambda)$ for L to obtain the result. □

2.3.2 Converse bounds

The converse bound in Theorem 11 requires Lemmas 6 and 7 below. Lemma 6 shows that codebooks whose codewords have types $\lambda_L(P) < \Lambda$ cannot have asymptotically decreasing probability of error. Lemma 7 shows list decoding cannot achieve rates higher than the mutual information of the channel. This kind of converse is based on symmetrizability arguments in Hughes [85] and Csiszár and Narayan [46] and is similar to those found in Gubner's paper on the multiple-access channel [73].

2.3.2.1 Converse for symmetrizable types

We begin with a lemma showing that a joint distribution $\overline{P}(x_1^L)$ with marginals P_i that are all close to some P can be approximated by a joint distribution $\hat{P}(x_1^L)$ whose marginals are all equal to P. The proof is given in Section D.1.

Lemma 4 (Approximating joint distributions). *Let $\mathcal{X}$ be a finite set with $|\mathcal{X}| \geq 2$. For any $\epsilon > 0$ and probability distribution P on $\mathcal{X}$ there exists a $\delta > 0$ such that for any collection of distributions $\{P_i \in \mathcal{P}(\mathcal{X}) : i \in [L]\}$ satisfying*

$$d_{\max}\left(P_i, P\right) < \delta \qquad \forall i \tag{2.64}$$

and any joint distribution $\overline{P}(x_1, x_2, \ldots, x_L)$ with

$$\sum_{x_j : j \neq i} \overline{P}(x_1, x_2, \ldots, x_L) = P_i(x_i) \qquad \forall i,\ x_i \in \mathcal{X} \tag{2.65}$$

there exists a joint distribution $\hat{P}(x_1, x_2, \ldots, x_L)$ such that

$$\sum_{x_j : j \neq i} \hat{P}(x_1, x_2, \ldots, x_L) = P(x_i) \qquad \forall i,\ x_i \in \mathcal{X} \tag{2.66}$$

and

$$d_{\max}\left(\overline{P}, \hat{P}\right) < \epsilon \ . \tag{2.67}$$

The following lemma shows that if the codebook has codewords whose types are symmetrizable and close to a fixed symmetrizable type P, then the jammer has a strategy that keeps the error bounded away from 0.

Lemma 5. *Let $\mathcal{W}$ be an AVC with state cost function $l(\cdot)$ and constraint Λ and let L be a positive integer. Let $\epsilon > 0$ be arbitrary and suppose P is a distribution with*

$\lambda_L(P) < \Lambda - \epsilon$. Then there exists a $\delta > 0$ and n_0 such that for any (n, N, L) list code with $n \geq n_0$ and $N \geq L + 1$ whose codewords $\{\mathbf{x}(i) : i \in [N]\}$ satisfy

$$d_{\max}\left(T_{\mathbf{x}(i)}, P\right) < \delta \qquad \forall i \in [N] \tag{2.68}$$

$$\lambda_L(T_{\mathbf{x}(i)}) < \Lambda - \epsilon \qquad \forall i \in [N] \ , \tag{2.69}$$

the average error for the code is lower bounded:

$$\max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \bar{\varepsilon}_L(\mathbf{s}) > \frac{1}{L+1} - \frac{L}{N(L+1)} \ . \tag{2.70}$$

Proof. From Lemma 4 we can see that for any $\epsilon_1 > 0$ there exists a $\delta_1 > 0$ such that for any set $J \subset [N]$ of codewords with $|J| = L$ and $d_{\max}\left(T_{\mathbf{x}(j)}, P\right) < \delta_1$, we can find a joint type $\overline{P} \in \mathcal{P}(\mathcal{X}^L)$ with marginals equal to P such that the joint type $T_{\mathbf{x}(J)}$ satisfies

$$d_{\max}\left(T_{\mathbf{x}(J)}, \overline{P}\right) < \epsilon_1 \ . \tag{2.71}$$

Now let U achieve the minimum in the definition of $\lambda_L(P)$. Since $\lambda_L(P) < \Lambda - \epsilon$ we have

$$\sum_{s, x_1^L} l(s) U(s|x_1^L) T_{\mathbf{x}(J)}(x_1^L) \leq \sum_{s, x_1^L} l(s) U(s|x_1^L) \overline{P}(x_1^L) + \epsilon_1 \lambda^* |\mathcal{X}|^L \tag{2.72}$$

$$< \Lambda - \epsilon + \epsilon_1 \lambda^* |\mathcal{X}|^L \ , \tag{2.73}$$

where $\lambda^* = \max_{s \in \mathcal{S}} l(s)$. Now choose $\epsilon_1 = \epsilon / (2\lambda^* |\mathcal{X}|^L)$ so that

$$\sum_{s, x_1^L} l(s) U(s|x_1^L) T_{\mathbf{x}(J)}(x_1^L) < \Lambda - \epsilon/2 \ , \tag{2.74}$$

and choose $\delta = \delta_1$ according to Lemma 4.

The jammer will pick a $J \subset [N]$ with $|J| = L$ uniformly from all such subsets and select its state sequence according to the random variable $\mathbf{S}(J)$ with distribution

$$Q^n(\mathbf{s}) = \prod_{t=1}^{n} U(s_t | \{x_t(j) : j \in J\}) \ . \tag{2.75}$$

The expected cost of $\mathbf{S}(J)$ is

$$\begin{aligned} \frac{1}{n}\mathbb{E}[l(\mathbf{S}(J))] &= \frac{1}{n}\sum_{t=1}^{n}\sum_{\mathbf{s}} l(s_t) U(s_t | \{x_t(j) : j \in J\}) & (2.76) \\ &= \sum_{s,\tilde{x}^L} l(s) U(s | \tilde{x}_1, \ldots, \tilde{x}_L) \frac{|\{t : x_t(j) = \tilde{x}_j \ \forall j\}|}{n} & (2.77) \\ &= \sum_{s,\tilde{x}^L} l(s) U(s | \tilde{x}_1^L) T_{\mathbf{x}(J)} & (2.78) \\ &< \Lambda - \epsilon/2 \ . & (2.79) \end{aligned}$$

We can also bound the variance of $l(\mathbf{S}(J))$:

$$\mathrm{Var}\left(l(\mathbf{S}(J))\right) \le \frac{(\lambda^*)^2}{n} \ . \tag{2.80}$$

Then Chebyshev's inequality gives the bound:

$$\begin{aligned} \mathbb{P}(l(\mathbf{S}(U_J, J)) > \Lambda) &\le \frac{(\lambda^*)^2}{n(\Lambda - (\Lambda - \epsilon/2))^2} & (2.81) \\ &\le \frac{4(\lambda^*)^2}{n\epsilon^2} \ . & (2.82) \end{aligned}$$

We now need some properties of symmetrizing channels used with the random variables $\mathbf{S}(J)$. Firstly, we have:

$$\begin{aligned}\mathbb{E}\left[W^n(\mathbf{y}|\mathbf{x}(i), \mathbf{S}(J))\right] &= \sum_{\mathbf{s}} W^n(\mathbf{y}|\mathbf{x}(i), \mathbf{s}) U^n(\mathbf{s}|\{x(j) : j \in J\}) &(2.83)\\ &= \mathbb{E}\left[W^n(\mathbf{y}|\mathbf{x}(j), \mathbf{S}(J \setminus \{j\} \cup \{i\}))\right] \ . &(2.84)\end{aligned}$$

Using (2.84) we can see that for some subset $G \subset [N]$ with $|G| = L+1$:

$$\begin{aligned}\sum_{i \in G} \mathbb{E}\left[\bar{\varepsilon}_L(i, \mathbf{S}(G \setminus \{i\}))\right] &= \sum_{i \in G} \left(1 - \sum_{\mathbf{y} : i \in \psi(\mathbf{y})} \mathbb{E}\left[W^n(\mathbf{y}|\mathbf{x}_i, \mathbf{S}(G \setminus \{i\}))\right] \right) &(2.85)\\ &= L + 1 - \sum_{i \in G} \sum_{\mathbf{y} : i \in \psi(\mathbf{y})} \mathbb{E}\left[W^n(\mathbf{y}|\mathbf{x}_{i_0}, \mathbf{S}_{G \setminus \{i_0\}})\right] \ . &(2.86)\end{aligned}$$

Because each $\mathbf{y}$ can be decoded to a list of size at most L , we can get a lower bound

$$\begin{aligned}\sum_{i \in G} \mathbb{E}\left[\varepsilon(i, \mathbf{S}_{G \setminus \{i\}})\right] &\geq L + 1 - L \sum_{\mathbf{y} \in \mathcal{Y}^n} \mathbb{E}\left[W^n(\mathbf{y}|\mathbf{x}_{i_0}, \mathbf{S}_{G \setminus \{i_0\}})\right]\\ &= 1 \ . &(2.87)\end{aligned}$$

We can now begin to bound the probability of error for this jamming strategy. Let $\mathcal{J}$ be the set of all subsets of $[N]$ of size L, and let $\mathbf{J}$ be a random variable uniformly distributed on $\mathcal{J}$. We can write the expected error as

$$\mathbb{E}_{\mathbf{J}, \mathbf{S}(\mathbf{J})}\left[\bar{\varepsilon}_L(\mathbf{S}(\mathbf{J}))\right] = \frac{1}{\binom{N}{L}} \frac{1}{N} \sum_{J \in \mathcal{J}} \sum_{i=1}^{N} \mathbb{E}\left[\bar{\varepsilon}_L(i, \mathbf{S}(J))\right] \ . \tag{2.88}$$

Then we have:

$$\mathbb{E}_{\mathbf{J}, \mathbf{S}(U_{\mathbf{J}}, \mathbf{J})}\left[\bar{\varepsilon}_L(\mathbf{S}(U_{\mathbf{J}}, \mathbf{J}))\right] \geq \frac{1}{\binom{N}{L}} \frac{1}{N} \sum_{G \subset [N] : |G| = L+1} \sum_{i \in G} \mathbb{E}\left[\bar{\varepsilon}_L(i, \mathbf{S}(G \setminus \{i\}))\right] \ . \tag{2.89}$$

Now we can rewrite the inner sum using (2.84):

$$\begin{aligned}
\mathbb{E}_{\mathbf{J},\mathbf{S}(\mathbf{J})}\left[\overline{\varepsilon}_L(\mathbf{S}(\mathbf{J}))\right] &\ge \frac{\binom{N}{L+1}}{\binom{N}{L}\cdot N} && (2.90)\\
&= \frac{\binom{N}{L}\frac{N-L}{L+1}}{\binom{N}{L}\cdot N} && (2.91)\\
&= \frac{N-L}{(L+1)N} && (2.92)\\
&= \frac{1}{L+1} - \frac{L}{N(L+1)} \ . && (2.93)
\end{aligned}$$

Finally, we can add in the bound (2.82) to obtain

$$\begin{aligned}
\frac{1}{L+1} - \frac{L}{N(L+1)} &\le \mathbb{E}_{\mathbf{J},\mathbf{S}(\mathbf{J})}\left[\overline{\varepsilon}_L(\mathbf{S}(\mathbf{J}))\right] && (2.94)\\
&\le \max_{\mathbf{s}\in\mathcal{S}^n(\Lambda)} \overline{\varepsilon}_L(\mathbf{s}) + \mathbb{P}\left(l(\mathbf{S}(\mathbf{J})) > \Lambda\right) && (2.95)\\
&\le \max_{\mathbf{s}\in\mathcal{S}^n(\Lambda)} \overline{\varepsilon}_L(\mathbf{s}) \frac{4(\lambda^*)^2}{n\epsilon^2} \ . && (2.96)
\end{aligned}$$

Now, we can choose n_0 large enough such that

$$\max_{\mathbf{s}\in\mathcal{S}^n(\Lambda)} \overline{\varepsilon}_L(\mathbf{s}) > \frac{1}{L+2} - \frac{L}{N(L+1)} \ . \qquad (2.97)$$

□

Lemma 6 combines Lemma 5 with a covering argument to show that for any $\epsilon > 0$ the jammer's strategy can be extended to codebooks for which all the codewords have type P with $\lambda_L(P) < \Lambda - \epsilon$.

Lemma 6. *Let $\mathcal{W}$ be an AVC with state cost function $l(\cdot)$ and constraint Λ and let L be a positive integer. For any $\epsilon > 0$ there exists a $\nu(L, \mathcal{W}, \epsilon) > 0$ and n_0 such that for any (n, N, L) list code (ϕ, ψ) with $n \geq n_0$ and $N > L + 1$ whose codewords*

$\{\mathbf{x}(i) : i \in [N]\}$ *satisfy*

$$\lambda_L(T_{\mathbf{x}(i)}) < \Lambda - \epsilon \qquad \forall i \in [N] \ , \tag{2.98}$$

the error must satisfy

$$\max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \overline{\varepsilon}_L(\mathbf{s}) > \nu(L, \mathcal{W}, \epsilon) \ . \tag{2.99}$$

Proof. Fix $\epsilon > 0$. For each $P \in \mathcal{P}(\mathcal{X})$ from Lemma 4 we know there is a $\delta(P) > 0$ such that any joint distribution $\overline{P}$ with marginals within $\delta(P)$ of P can be approximated by a $\hat{P}$ with marginals equal to P such that $d_{\max}\left(\overline{P}, \hat{P}\right) < \epsilon$. Let

$$\mathcal{B}(P) = \{P' \in \mathcal{P}(\mathcal{X}) : d_{\max}\left(P, P'\right) < \delta(P)\} \ . \tag{2.100}$$

Then $\{\mathcal{B}(P) : P \in \mathcal{P}(\mathcal{X})\}$ is an open cover of $\mathcal{P}(\mathcal{X})$. Since $\mathcal{P}(\mathcal{X})$ is compact there is a constant r and finite subcover $\{\mathcal{B}(P_j) : j \in [r]\}$. From this finite cover we can create a partition $\{A_j : j \in [r]\}$ of $\mathcal{P}$ such that $A_j \subseteq \mathcal{B}(P_j)$ for all j.

Now consider an (n, N, L) code whose codewords $\mathcal{C}$ satisfy (2.98). Let $F_j = \{i \in [N] : T_{\mathbf{x}(i)} \in A_j\}$. We can bound the error

$$\overline{\varepsilon}_L(\mathbf{s}) = \frac{1}{Nr} \sum_{j=1}^{r} \sum_{i \in F_j} \overline{\varepsilon}_L(i, \mathbf{s}) \geq \frac{|F_j|}{Nr} \left(\frac{1}{|F_j|} \sum_{i \in F_j} \overline{\varepsilon}_L(i, \mathbf{s}) \right) \ . \tag{2.101}$$

Since $\{F_j\}$ partition the codebook, for some j we have $|F_j| \geq N/r$. From Lemma 5 the jammer can force the error to be lower bounded by

$$\max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \overline{\varepsilon}_L(\mathbf{s}) \geq \frac{1}{r^2} \left(\frac{1}{L+1} - \frac{L}{N(L+1)} \right) \ . \tag{2.102}$$

Since the constant r is a function of ϵ, $\mathcal{W}$ and L, we are done. □

2.3.2.2 Converse for rates above the mutual information

The other converse result shows that a codebook with codewords whose types are in a set $\mathcal{A}$ cannot achieve rates higher than the maximum of the mutual information $I(P, \Lambda)$ over all types $P \in \mathcal{A}$.

Lemma 7. *Let $\mathcal{W}$ be an AVC with state cost function $l(\cdot)$. Fix a subset $\mathcal{A} \subset \mathcal{P}(\mathcal{X})$. For any constant list size L and constants $\Lambda > 0$ and $\epsilon > 0$, there exists an n_0 and $\delta > 0$ such that for any (n, N, L) list-decodable code of blocklength $n \geq n_0$ with codewords whose types are in $\mathcal{A}$, the inequality*

$$\frac{1}{n} \log \frac{N}{L} \geq \max_{P \in \mathcal{A}} I(P, \Lambda) + \epsilon \tag{2.103}$$

implies

$$\max_{\mathbf{s}:\mathcal{S}^n(\Lambda)} \bar{\varepsilon}_L(\mathbf{s}) > \delta \ . \tag{2.104}$$

Proof. Fix $\epsilon > 0$. For each $P \in \mathcal{P}(\mathcal{X})$ let $Q_P^* \in \mathcal{Q}(\Lambda)$ be a jamming distribution that achieves the minimum in

$$I(P, \Lambda) = \min_{V \in \mathcal{W}_{std}(\Lambda)} I(P, V) \ . \tag{2.105}$$

Without loss of generality, we can take $\min l(s) = 0$. Let $s_0 = \operatorname{argmin}_{s \in \mathcal{S}} l(s)$. For an $\eta \in (0, 1)$, we define the new distribution

$$Q_P(s) = \begin{cases} Q_P^*(s)(1-\eta) & s \neq s_0 \\ \eta + (1-\eta) Q_P^*(s) & s = s_0 \end{cases} \tag{2.106}$$

Under $Q_P(s)$, we have $\mathbb{E}[l(s)] \leq \Lambda(1-\eta)$. Define the channel

$$(VQ_P)(y|x) = \sum_{s} W(y|x,s) Q_P(s) \ . \tag{2.107}$$

The mutual information is uniformly continuous in the channel, so for any $\epsilon > 0$ we can choose η such that

$$I\left(P, VQ_P\right) \leq \min_{V \in \mathcal{W}_{std}(\Lambda)} I\left(P, V\right) + \epsilon/2 \ . \tag{2.108}$$

Because the mutual information is also continuous in the input distribution, for any $\epsilon > 0$ there exists a $\delta' > 0$ so that for any P' with $d_{\max}\left(P, P'\right) < \delta'$ we have

$$I\left(P', VQ_P\right) \leq \min_{V \in \mathcal{W}_{std}(\Lambda)} I\left(P, V\right) + \epsilon/4 \ . \tag{2.109}$$

Now, the open balls

$$\mathcal{B}(P) = \{P' : d_{\max}\left(P, P'\right) < \delta'\} \tag{2.110}$$

are an open cover of the compact set $\mathcal{P}(\mathcal{X})$, so there exists a constant r and a finite set $\{P_j : j \in [r]\}$ such that $\{\mathcal{B}(P_j) : j \in [r]\}$ is a finite subcover. From this we can generate a partition $\{\mathcal{A}_j : j \in [r]\}$ of $\mathcal{A}$ such that $\mathcal{A}_j \subset \mathcal{B}(P_j)$ for all $j \in [r]$.

Now let $\{\mathbf{x}(1) : i \in [N]\}$ be the codewords of an (n, N, L) list-decodable code such that $T_{\mathbf{x}(i)} \in \mathcal{A}$ for all $i \in [N]$. Let $F_j = \{i : T_{\mathbf{x}(i)} \in \mathcal{A}_j\}$. We can find a $k \in [r]$ such that $|F_k| \geq N/r$. The jammer will choose its state sequence $\mathbf{s}$ according to a random variable $\mathbf{S}$ chosen iid according to Q_{P_j}. The expected error $\mathbb{E}[\bar{\varepsilon}_L(\mathbf{S})]$ under this strategy can be lower bounded :

$$\mathbb{E}[\overline{\varepsilon}_L(\mathbf{S})] \geq \frac{|F_k|}{N} \left(\frac{1}{|F_k|} \sum_{i \in F_k} \mathbb{E}[\overline{\varepsilon}_L(i, \mathbf{S})] \right) \tag{2.111}$$

$$\geq \frac{1}{r} \left(\frac{1}{|F_k|} \sum_{i \in F_k} \mathbb{E}[\overline{\varepsilon}_L(i, \mathbf{S})] \right) . \tag{2.112}$$

The term inside the parentheses is the expected average error of the sub-codebook $\{\mathbf{x}(i) : i \in F_k\}$ which has rate

$$\frac{1}{n} \log \frac{|F_k|}{L} \geq \max_{P \in \mathcal{A}} I\left(P, \Lambda\right) + \epsilon - \frac{1}{n} \log r . \tag{2.113}$$

This expected average error is equal to the average error $\overline{\varepsilon}_Q$ for the sub-codebook on the DMC VQ_{P_K}. Then we can lower bound the error

$$\max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \overline{\varepsilon}_L(\mathbf{s}) \geq \overline{\varepsilon}_Q - \mathbb{P}_Q(l(\mathbf{S}) > n\Lambda) . \tag{2.114}$$

By the strong converse for the list decoding on a DMC [117], for any $\epsilon' > 0$ and rate larger than $I\left(P, VQ_{P_k}\right) + \epsilon'$ the average error $\overline{\varepsilon}_Q$ converges to 1. Standard large deviations results show that $l(\mathbf{S})$ will satisfy the cost constraint with high probability. Therefore, for sufficiently large n we can lower bound the error by $\delta = 1/(2r)$. □

2.3.3 Achievability arguments

To show the forward part of Theorem 11 we must define a codebook and decoding rule. In Section 2.3.3.1 we describe the decoding rule, which is a modified version of the rule used by Hughes [85]. In Section 2.3.3.2 we prove a lemma (Lemma 9), which we will use to show that the decoding rule can only produce a list of size $\tilde{L}_{\mathrm{sym}}(P, \Lambda)+1$ or smaller. In Section 2.3.3.3 we prove Lemma 10, which proves the existence of a constant composition codebook of type P with useful properties. Finally, in Section 2.3.3.4 we show that this codebook can be used in conjuction with the decoding rule to achieve rates arbitrarily close to $I\left(P, \Lambda\right)$ with lists of size $\tilde{L}_{\mathrm{sym}}(P, \Lambda) + 1$.

2.3.3.1 The decoding rule

In order to describe the decoding rule we will use, we define the set

$$\mathcal{G}_\eta(\Lambda) = \{P_{XSY} \in \mathcal{P}(\mathcal{X} \times \mathcal{S} \times \mathcal{Y}) : D\left(P_{XSY} \;\|\; P_X \times P_S \times W\right) \le \eta,\ \mathbb{E}[l(s)] \le \Lambda\} \ , \tag{2.115}$$

where

$$(P_X \times P_S \times W)(x, s, y) = P_X(x) P_S(s) W(y|x, s) \ . \tag{2.116}$$

The set $\mathcal{G}_\eta(\Lambda)$ contains joint distributions which are close to those generated from the AVC $\mathcal{W}$ via independent inputs with distribution P_X and P_S.

Definition 1 (Decoding rule)**.** *Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ be a given codebook and suppose $\mathbf{y}$ was received. Let $\psi(\mathbf{y})$ denote the list decoded from $\mathbf{y}$. Then put $i \in \psi(\mathbf{y})$ if and only if there exists an $\mathbf{s} \in \mathcal{S}^n(\Lambda)$ such that*

1. *$T_{\mathbf{x}_i \mathbf{s} \mathbf{y}} \in \mathcal{G}_\eta(\Lambda)$, and*

2. *for every set of L other distinct codewords $\{\mathbf{x}_j : j \in J,\ J \subset [N] \setminus \{i\},\ |J| = L\}$ such that there exists a set $\{\mathbf{s}_j : \mathbf{s}_j \in \mathcal{S}^n(\Lambda),\ j \in J\}$ with $T_{\mathbf{x}_j \mathbf{s}_j \mathbf{y}} \in \mathcal{G}_\eta(\Lambda)$ for all $j \in J$ we have*

$$I\left(YX \ \wedge \ X^L \middle| S\right) \le \eta \ , \tag{2.117}$$

where P_{YXX^LS} is the joint type of $(\mathbf{y}, \mathbf{x}_i, \{\mathbf{x}_j : j \in J\}, \mathbf{s})$.

This is the decoding rule used by Hughes [85] modified in the natural way suggested by Csiszár and Narayan [46]. An interpretation of this rule is that the decoder outputs a list of codewords $\{\mathbf{x}_i\}$ each having a "good explanation" $\{\mathbf{s}_i\}$. A "good explanation" is a state sequence that plausibly could have generated the observed output $\mathbf{y}$ (condition 1) and makes all other L-tuples of codewords seem independent of the codeword and output (condition 2).

2.3.3.2 Guaranteeing a bounded list size

Lemma 9 will be used to show that the decoding rule cannot output a list larger than $M = \tilde{L}_{\text{sym}}(P, \Lambda) + 1$. The key is to show that no tuple of random variables (Y, X^{M+1}, S^{M+1}) can satisfy the conditions of the decoding rule. This in turn shows that for sufficiently large n, no set of $M + 1$ *codewords* can satisfy the conditions of the decoding rule. Therefore, for sufficiently large blocklengths, the decoding rule will only output M or fewer codewords.

The proof of Lemma 9 rests on Lemma 8, which guarantees a nonzero total variational distance between joint distributions on $\mathcal{X}^{M+1} \times \mathcal{Y}$ induced by the AVC. For a tuple $x^M = (x_1, x_2, \ldots, x_M)$, define $x^M_{-\{i\}}$ to be the tuple $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_M)$.

Lemma 8. *Let $\beta > 0$, $\mathcal{W}$ be an AVC with state cost function $l(\cdot)$ and constraint Λ, $P \in \mathcal{P}(\mathcal{X})$ with $I(P, \Lambda) > 0$ and $\min_x P(x) \geq \beta$, and $M = \tilde{L}_{\text{sym}}(P, \Lambda) + 1$. For any $\alpha > 0$ and every collection of distributions $\{U_i \in \mathcal{P}(\mathcal{X}^M \times \mathcal{S}) : i = 1, 2, \ldots, M\}$ such that*

$$\sum_{x^{M+1}, s} P(x_i) U_i(x^M_{-\{i\}}, s) l(s) \leq \tilde{\lambda}_M(P) - \alpha \tag{2.118}$$

for all $i = 1, 2, \ldots, M+1$, *there exists a* $\zeta > 0$ *such that*

$$\max_{j \neq i} \sum_{y, x^{M+1}} \left| \sum_s W(y|x_i, s) U_i(x^{M+1}_{-\{i\}}, s) P(x_i) - \sum_s W(y|x_j, s) U_j(x^{M+1}_{-\{j\}}, s) P(x_j) \right| \geq \zeta \ . \tag{2.119}$$

Proof. Note that the outer sum in (2.119) is over all x^{M+1}. Define the function $V_k : \mathcal{X}^{M+1} \times \mathcal{S} \to \mathbb{R}$ by:

$$V_k(x^{M+1}, s) = U_k(x^{M+1}_{-\{k\}}, s) \ . \tag{2.120}$$

Let Π_{M+1} be the set of all permutations of $[M+1]$ and for $\pi \in \Pi_{M+1}$ let π_i be the image of i under π. Then

$$\begin{aligned} &\max_{j \neq i} \sum_{y, x^{M+1}} \left| \sum_s W(y|x_i, s) V_i(x^{M+1}, s) P(x_i) - \sum_s W(y|x_j, s) V_j(x^{M+1}, s) P(x_j) \right| \\ &= \max_{j \neq i} \sum_{y, x^{M+1}} \left| \sum_s W(y|x_i, s) V_{\pi_i}(\pi(x^{M+1}), s) P(x_i) \right. \\ &\qquad\qquad \left. - \sum_s W(y|x_j, s) V_{\pi_j}(\pi(x^{M+1}), s) P(x_j) \right| \ . \end{aligned} \tag{2.121}$$

We can lower bound this by averaging over all $\pi \in \Pi_{M+1}$:

$$\begin{aligned} \max_{j \neq i} \sum_{y, x^{M+1}} \frac{1}{(M+1)!} \sum_{\pi \in \Pi_{M+1}} &\left| \sum_s W(y|x_i, s) V_{\pi_i}(\pi(x^{M+1}), s) P(x_i) \right. \\ &\qquad \left. - \sum_s W(y|x_j, s) V_{\pi_j}(\pi(x^{M+1}), s) P(x_j) \right| \ . \end{aligned} \tag{2.122}$$

Define the average

$$\begin{aligned}
\overline{V}(x_{-\{i\}}^{M+1}, s) &= \frac{1}{(M+1)!} \sum_{\pi \in \Pi_{M+1}} V_{\pi_i}(\pi(x^{M+1}), s) \\
&= \frac{1}{(M+1)!} \sum_{l=1}^{M+1} \sum_{\pi \in \Pi_{M+1} : \pi_i = l} U_l(\pi(x^{M+1})_{-\{\pi_i\}}, s) \\
&= \frac{1}{(M+1)!} \sum_{l=1}^{M+1} \sum_{\sigma \in \Pi_M} U_l(\sigma(x_{-\{i\}}^{M+1}), s) \ .
\end{aligned}$$

Note that $\overline{V}$ is a symmetric function for all s.

Now we use the convexity of $|\cdot|$ to pull the averaging inside the absolute value to get a further lower bound on (2.122) by substituting in $\overline{V}$.

$$F(\overline{V}, P) = \max_{j \neq i} \sum_{y, x^{M+1}} \Bigg| \sum_s W(y|x_i, s) \overline{V}(x_{-\{i\}}^{M+1}, s) P(x_i) \\
- \sum_s W(y|x_j, s) \overline{V}(x_{-\{j\}}^{M+1}, s) P(x_j) \Bigg| \ . \tag{2.123}$$

The function $F(\overline{V}, P)$ is continuous function on the compact set of symmetric distributions $\{\overline{V}\}$ and the set of distributions P with $\min_x P(x) \geq \beta$, so it has a minimum $\zeta = F(\overline{V}^*, P^*)$ for some $(\overline{V}^*, P^*)$. We will prove that $\zeta > 0$ by contradiction.

Suppose $F(\overline{V}^*, P^*) = 0$. Then

$$\sum_s W(y|x_i, s) \overline{V}^*(x_{-\{i\}}^{M+1}, s) P^*(x_i) = \sum_s W(y|x_j, s) \overline{V}^*(x_{-\{j\}}^{M+1}, s) P^*(x_j) \ .$$

So

$$\sum_y \sum_s W(y|x_i, s)\overline{V}^*(x_{-\{i\}}^{M+1}, s)P^*(x_i) = \sum_y \sum_s W(y|x_j, s)\overline{V}^*(x_{-\{j\}}^{M+1}, s)P^*(x_j)$$
$$\overline{V}^*(x_{-\{i\}}^{M+1})P^*(x_i) = \overline{V}^*(x_{-\{j\}}^{M+1})P^*(x_j) \ ,$$

which implies (see [85, Lemma A3]) that for all j:

$$\overline{V}^*(x_{-\{j\}}^{M+1})P^*(x_j) = P^{*(M+1)}(x^{M+1}) \ .$$

Therefore

$$\sum_s W(y|x_1, s)\overline{V}^*(s|x_2^{M+1}) \ . \tag{2.124}$$

is symmetric in $(x_1, x_2, \ldots, x_{M+1})$. Therefore $\overline{V}^*(s|x_2^{M+1}) \in \mathcal{U}_{\text{sym}}(M+1)$. From the definition of $\tilde{\lambda}_M(P)$ in (2.53) we see that

$$\sum_{x^{M+1}, s} \overline{V}^*(x_{-\{i\}}^M, s)P(x_i)l(s) \geq \tilde{\lambda}_M(P) \ . \tag{2.125}$$

But from (2.118), and the definition of $\overline{V}$ we see that the $\{U_i\}$ must be chosen such that

$$\sum_{x^{M+1}, s} \overline{V}^*(x_{-\{i\}}^M, s)P(x_i)l(s) \leq \tilde{\lambda}_M(P) - \alpha \ . \tag{2.126}$$

Therefore we have a contradiction and the minimum ζ of $F(\overline{V}, P)$ must be greater than 0. Equation (2.119) follows. □

We can use Lemma 8 to show that our decoding rule will not output a list of size larger than $\tilde{L}_{\text{sym}}(P,\Lambda)+1$. The next lemma shows that for a sufficiently small choice of the threshold η in the decoding rule given in Section 2.3.3.1 there are no random variables that can force the decoding rule to output a list that is too large.

Lemma 9. *Let $\beta > 0$, $\mathcal{W}$ be an AVC with state cost function $l(\cdot)$ and constraint Λ, $P \in \mathcal{P}(\mathcal{X})$ with $\min_x P(x) \geq \beta$, and $M = \tilde{L}_{\text{sym}}(P,\Lambda) + 1$. Then there exists an $\eta > 0$ sufficiently small such that no tuple of rv's (Y, X^{M+1}, S^{M+1}) can simultaneously satisfy*

$$\min_x P(x) \geq \beta \tag{2.127}$$

$$P_{X_i} = P \tag{2.128}$$

$$P_{YX_iS_i} \in \mathcal{G}_\eta(\Lambda) \tag{2.129}$$

$$I\left(YX_i \wedge X_{-\{i\}}^{M+1} \middle| S_i\right) \leq \eta \quad 1 \leq i \leq M+1 \tag{2.130}$$

Proof. Assume, to the contrary, that there does exist a tuple of random variables (Y, X^{M+1}, S^{M+1}) satifying (2.127)–(2.130). This will lead to a bound on a certain KL-divergence which, via Pinsker's inequality, becomes a bound on total variational distance that contradicts the conclusion of Lemma 8 with $U_i = P_{X_{-\{i\}}^{M+1}S_i}$. The assumption (2.129) shows that

$$\sum_{x^{M+1},s} P(x_i)P_{X_{-\{i\}}^{M+1}S_i}(x_{-\{i\}}^{M+1},s)l(s) \leq \Lambda < \tilde{\lambda}_M(P) \ , \tag{2.131}$$

so (2.118) holds with $\alpha = \tilde{\lambda}_M(P) - \Lambda$.

Let

$$(W \times P_{X_i} \times P_{X_{-\{i\}}^{M+1}S_i})(y, x^{M+1}, s) = W(y|x_i,s)P_{X_i}(x_i)P_{X_{-\{i\}}^{M+1}S_i}(x_{-\{i\}}^{M+1},s) \ . \tag{2.132}$$

For every i we have the following divergence bound:

$$
\begin{aligned}
& D\left(P_{YX^{M+1}S_i} \,\middle\|\, W \times P_{X_i} \times P_{X_{-\{i\}}^{M+1}S_i}\right) && (2.133) \\
&\quad = D\left(P_{YX_iS_i} \,\|\, W_i \times P_{X_i} \times P_{S_i}\right) + D\left(P_{X_{-\{i\}}^{M+1}|YX_iS_i} \,\middle\|\, P_{X_{-\{i\}}^{M+1}|S_i} \,\middle|\, P_{YX_iS_i}\right) && (2.134) \\
&\quad = D\left(P_{YX_iS_i} \,\|\, W_i \times P_{X_i} \times P_{S_i}\right) + I\left(YX_i \wedge X_{-\{i\}}^{M+1} \,\middle|\, S_i\right) && (2.135) \\
&\quad \leq 2\eta \ , && (2.136)
\end{aligned}
$$

where the last line follows from (2.129), (2.115) and (2.130).

Projecting the distributions onto $\mathcal{Y} \times \mathcal{X}^{M+1}$ cannot increase the divergence. Setting

$$
V_i(y, x_{-\{i\}}^{M+1}|x_i) = \sum_s W(y|x_i, s) P_{X_{-\{i\}}^{M+1}S_i}(x_{-\{i\}}^{M+1}, s) \ , \tag{2.137}
$$

we obtain from (2.136) the following inequality:

$$
D\left(P_{YX^{M+1}} \,\|\, V_i \times P_{X_i}\right) < 2\eta \ .
$$

To use Lemma 8 we must turn this divergence bound into a bound on a total variational distance. We can use Pinsker's inequality [44, p. 58, Problem 17] to show that the KL-divergence is an upper bound on the variational distance:

$$
\sum_{y, x^{M+1}} \left| P_{YX^{M+1}}(y, x^{M+1}) - V_i(y, x_{-\{i\}}^{M+1}|x_i) P_{X_i}(x_i) \right| < \sqrt{(4 \ln 2)\eta} \quad \forall i \in [M+1] \ . \tag{2.138}
$$

Since the bound holds for all i, we know that $V_i P_{X_i}$ is close to $V_j P_{X_j}$. The triangle inequality yields:

$$\max_{i \neq j} \sum_{y, x^{M+1}} \left| V_j(y, x_{-\{j\}}^{M+1} | x_j) P_{X_j}(x_j) - V_i(y | x_{-\{i\}}^{M+1} | x_i) P_{X_i}(x_i) \right| < 2\sqrt{(4 \ln 2)\eta} \ . \qquad (2.139)$$

By choosing η small enough this violates the conclusion of Lemma 8, which is the desired contradiction. Therefore no tuple of random variables can satisfy the conditions given in (2.127) - (2.130). □

2.3.3.3 Codebook properties

We quote Lemma 1 from Hughes [85] in Lemma 10. The proof is given in [85] and requires some large deviations results [9, 46] that have proved useful in many other coding problems for AVCs under average error [48, 85, 72, 73]. We note that these properties do not depend on the constraints.

Lemma 10 (Codebook existence)**.** *For any $L \geq 1$, $\epsilon > 0$, $R = n^{-1} \log(N/L) \geq \epsilon$, and type P, there exists an n sufficiently large and codewords $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ of blocklength n, each of type P, such that for every $\mathbf{x} \in \mathcal{X}^n$, $\mathbf{s} \in \mathcal{S}^n$, and joint type P_{XX^LS} we have the following for $k = 1, 2, \ldots, L$:*

1. *If $I\left(X \ \wedge \ S\right) \geq \epsilon$ then*

$$\frac{1}{N} \left| \{ i : T_{\mathbf{x}_i \mathbf{s}} = P_{XS} \} \right| \leq \exp(-n\epsilon/2) \ . \qquad (2.140)$$

2. *If $I\left(X \ \wedge \ X_k S\right) \geq |R - I\left(X_k \ \wedge \ S\right)|^+ + \epsilon$ then*

$$\frac{1}{N} \left| \left\{ i : T_{\mathbf{x}_i \mathbf{x}_j \mathbf{s}} = P_{XX_kS} \ \textit{for some} \ j \neq i \right\} \right| \leq \exp(-n\epsilon/2) \ . \qquad (2.141)$$

3. *Also, for any* $\mathbf{x}$

$$\left|\left\{j : T_{\mathbf{x}\mathbf{x}_j\mathbf{s}} = P_{XX_kS}\right\}\right| \le \exp\left(n\left(\left|R - I\left(X_k \ \wedge \ XS\right)\right|^+ + \epsilon\right)\right) \ . \tag{2.142}$$

4. *Moreover, if* $R < \min_k I\left(X_k \ \wedge \ S\right)$*, then* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ *can be selected to further satisfy*

$$|\{J \subset [N] : |J| = L, \ T_{\mathbf{x}_i\mathbf{x}_J\mathbf{s}} = P_{XX^LS}\}| \le \exp(n\epsilon) \ . \tag{2.143}$$

5. *If* $R < \min_k I\left(X_k \ \wedge \ S\right)$ *and* $I\left(X \ \wedge \ X^LS\right) \ge \epsilon$ *then*

$$\frac{1}{N}\left|\{i : T_{\mathbf{x}_i\mathbf{x}_J\mathbf{s}} = P_{XX^LS} \text{ for some } J \subset [N] \setminus \{i\}, \ |J| = L\}\right| \le \exp(-n\epsilon/2) \ . \tag{2.144}$$

2.3.3.4 Achievable rates for list decoding

The proof of the following lemma is nearly identical to Lemma 3 of Hughes, and a proof is included for completeness in Appendix D.2.

Lemma 11. *Let* $\mathcal{W}$ *be an AVC with state cost function* $l(\cdot)$ *and state constraint* Λ*. For any* $\epsilon_3 > 0$*,* $\beta > 0$ *and* $P \in \mathcal{P}(\mathcal{X})$ *with* $I(P, \Lambda) > 0$ *and* $\min_x P(x) \ge \beta$*, and* $M = \tilde{L}_{\mathrm{sym}}(P, \Lambda) + 1$ *there exists a positive integer* $n_0(\beta, \epsilon_3, \mathcal{W})$*,* $\delta(\beta, \epsilon_3, \mathcal{W}) > 0$ *and an* (n, N, M) *list-decodable code with* $n > n_0$ *whose codewords have constant type* P *such that*

$$R = \frac{1}{n}\log\left(\frac{N}{M}\right) > I\left(P, \Lambda\right) - \epsilon_3 \tag{2.145}$$

$$\max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \bar{\varepsilon}(\mathbf{s}) < \exp(-n\delta) \ . \tag{2.146}$$

2.3.4 Proof of Theorem 11

Proof. Define the function

$$\rho(\alpha) = \max_{P \in \mathcal{I}(\Gamma - \alpha) : \tilde{\lambda}_L(P) \geq \Lambda + \alpha} \min_{V \in \mathcal{W}_{std}(\Lambda)} I\left(P, V\right) \ . \tag{2.147}$$

We claim that $\rho(\alpha)$ is continuous for α in a small neighborhood of 0. Since $I\left(P, V\right)$ is a continuous and concave function of P, the minimum $I\left(P, \Lambda\right) = \min_{V \in \mathcal{W}_{std}(\Lambda)} I\left(P, V\right)$ is also a continuous concave function. Since $\tilde{\lambda}_L(P)$ is a the minimum of a set of linear functions it is concave in P^L and from [32, p.86] we can see that it is a concave function of P. Therefore, as in Csiszár and Narayan [46, p. 188] we can see that $\rho(\alpha)$ is concave for α such that $\{P \in \mathcal{I}(\Gamma - \alpha) > \tilde{\lambda}_L(P) \geq \Lambda + \alpha\}$ is nonempty. Since we have assumed that $\max_{P \in \mathcal{I}(\Gamma)} \tilde{\lambda}_L(P) > \Lambda$ the point $\alpha = 0$ is in this interval, so there exists a $\zeta > 0$ such that $\rho(\alpha)$ is continuous in a small neighborhood $(-\zeta, \zeta)$ of 0.

Fix α. Let P^* achieve the maximum for $\rho(\alpha)$. For any $\epsilon > 0$ there exists a $\delta > 0$ such that for any P with $d_{\max}\left(P^*, P\right) < \delta'$ we have

$$I\left(P, \Lambda\right) \geq I\left(P^*, \Lambda\right) - \epsilon \ . \tag{2.148}$$

We have then for some constant $c_1(L, \mathcal{W})$ that

$$\sum_{x \in \mathcal{X}} g(x) P(x) \leq \Gamma - \alpha + c_1(L, \mathcal{W}) \delta \ . \tag{2.149}$$

Furthermore, let $U \in \mathcal{U}_{\text{sym}}$ attain the minimum in $\tilde{\lambda}_L(P)$. Then for some constant

$c_2(L, \mathcal{W})$ we have

$$\begin{aligned}
\Lambda + \alpha &\le \tilde{\lambda}_L(P^*) && (2.150)\\
&= \min_{U^* \in \mathcal{U}_{\mathrm{sym}}} \sum_{s, x_1^L} l(s) U^*(s|x_1^L) \prod_{j=1}^{L} P^*(x_j) && (2.151)\\
&\le \sum_{s, x_1^L} l(s) U(s|x_1^L) \prod_{j=1}^{L} P^*(x_j) && (2.152)\\
&\le \tilde{\lambda}_L(P) + c_2(L, \mathcal{W})\delta \ . && (2.153)
\end{aligned}$$

Thus we can choose δ small enough such that $P \in \mathcal{I}(\Gamma)$ and $\tilde{\lambda}_L(P) \ge \Lambda + \alpha/2$.

Now we can a P with with $d_{\max}(P^*, P) < \delta'$ such that there exists a $\beta > 0$ and $\min_x P(x) \ge \beta$. Then by Lemma 11 we can find a blocklength sufficiently large and a list-decodable code with list size L such that

$$\begin{aligned}
\frac{1}{n} \log \left(\frac{N}{M} \right) &\ge I(P, \Lambda) - \epsilon && (2.154)\\
&\ge \rho(\alpha) - 2\epsilon \ , && (2.155)
\end{aligned}$$

and the error is as small as we like. Since $\rho(\alpha)$ is continuous, this shows that $R = \rho(0)$ is achievable with input constraint Γ and cost constraint Λ.

For the converse, note that for any $\alpha > 0$, Lemmas 6 and 7 show that no rate above

$$\nu(\alpha) = \max_{P \in \mathcal{I}(\Gamma) : \lambda_L(P) \ge \Lambda - \alpha} \ \min_{V \in \mathcal{W}_{std}(\Lambda)} I(P, V) \qquad (2.156)$$

can be achievable. Again, $I\left(P, \Lambda\right)$ is a continuous function of P, and

$$\begin{aligned}\mu\lambda_L(P) + (1-\mu)\lambda_L(P') &\leq \min_{U \in \mathcal{U}_{\mathrm{sym}}(L)} \max_{\overline{P}: P_i = P,\ \overline{P}': P'_i = P'} \sum_{s, x_1^L} l(s) U(s|x_1^L)(\mu\overline{P} + (1-\mu\overline{P}') \\ &\leq \min_{U \in \mathcal{U}_{\mathrm{sym}}(L)} \max_{\overline{P}'': P''_i = \mu P + (1-\mu)P'} \sum_{s, x_1^L} l(s) U(s|x_1^L)\overline{P}''(x_1^L) \\ &= \lambda_L(\mu P + (1-\mu)P') \ , \qquad (2.157)\end{aligned}$$

so $\lambda_L(P)$ is concave. Therefore $\nu(\alpha)$ is continuous when $\{P \in \mathcal{I}(\Gamma) : \lambda_L(P) \leq \Lambda - \alpha\}$ is nonempty. Since $\max_{P \in \mathcal{I}(\Gamma)} \lambda_L(P) < \Lambda$ we know $\nu(\alpha)$ is continuous at 0, which gives the converse. □

2.4 Example

We now turn to an example of an additive cost-constrained AVC. Let $\mathcal{X} = \{-1, 1\}$ and let $\mathcal{S} = \{-\sigma, -\sigma+1, \ldots, \sigma\}$ for some integer σ. The output Y of this channel is given by

$$Y = X + S \ . \qquad (2.158)$$

That is, Y is the real addition of the input and state. This is similar in spirit to the example given by Hughes [85], but is more closely related to [134], which analyzes a game between power constrained noise and an encoder with binary inputs.

We will consider two kinds of cost constraints on the jammer for this AVC. The first is an L_1 constraint:

$$l_1(s) = |s| \ . \qquad (2.159)$$

The second is an L_2 constraint:

$$l_2(s) = |s|^2 \ . \tag{2.160}$$

For each of these constraints we will compute capacities for different values of Λ.

2.4.1 Randomized coding capacity

The first question to settle is that of the randomized coding capacity $C_r(\Lambda) = C_{\text{std}}(\Lambda)$ for this channel, given by (1.50). By symmetry, we may assume that the input distribution P^* is uniform on the set $\{-1, 1\}$. Therefore we can write:

$$\begin{aligned} C_r(\Lambda) &= \min_{Q(s) \in \mathcal{P}(\mathcal{S},\Lambda)} I\left(X \ \wedge \ X + S\right) \\ &= \min_{Q(s) \in \mathcal{P}(\mathcal{S},\Lambda)} H(X+S) - H(S) \ . \end{aligned} \tag{2.161}$$

To find the random coding capacity we must minimize the mutual information $I\left(X \ \wedge \ X+S\right)$. This can be stated as the following optimization problem. Let $I(Q) = I\left(X \ \wedge \ X+S\right)$ with $Q = Q(s)$ and $P(X=1) = P(X=-1) = 1/2$. For a cost function $l(s) = |s|^\theta$, let $\Theta = (l(-A), l(-A+1), \ldots l(A))^T$. The optimization is

$$\begin{aligned} \text{minimize} \quad & I(Q) && (2.162) \\ \text{subject to} \quad & \mathbf{1}^T Q = 1 && (2.163) \\ & -Q(s) \leq 0 \quad \forall s && (2.164) \\ & \Theta^T Q - \Lambda \leq 0 \ , && (2.165) \end{aligned}$$

Since the mutual information is convex in the distribution Q, this is a convex optimization problem in the vector Q and can be solved using standard optimization techniques [32].

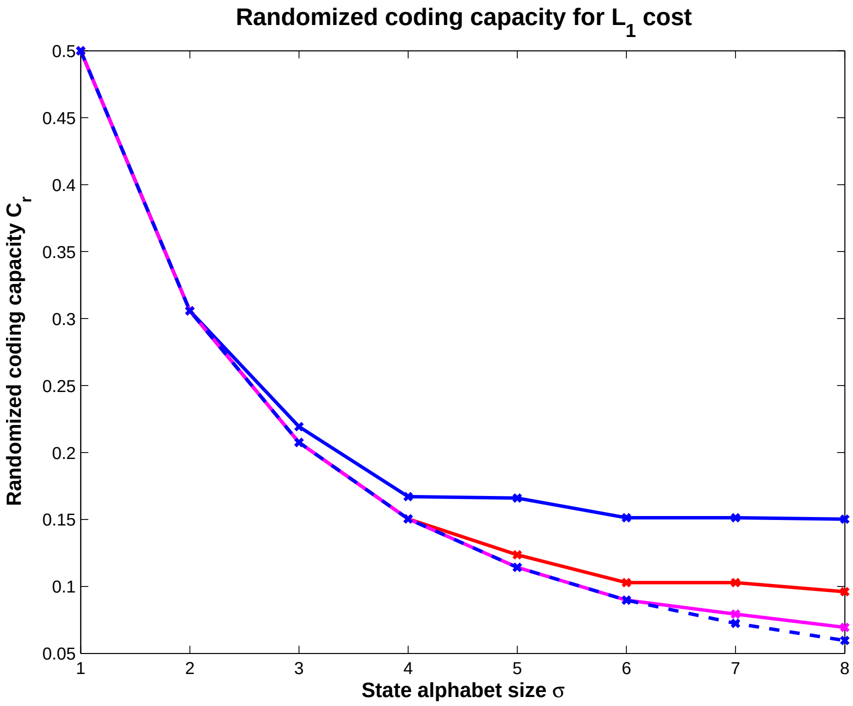


Figure 2.2: Randomized coding capacity $C_r(\Lambda)$ versus σ for $\Lambda = 1.5$, 2, and 2.5 with loss function $l_1(s)$. The dashed line is the randomized coding capacity for the unconstrained jammer.

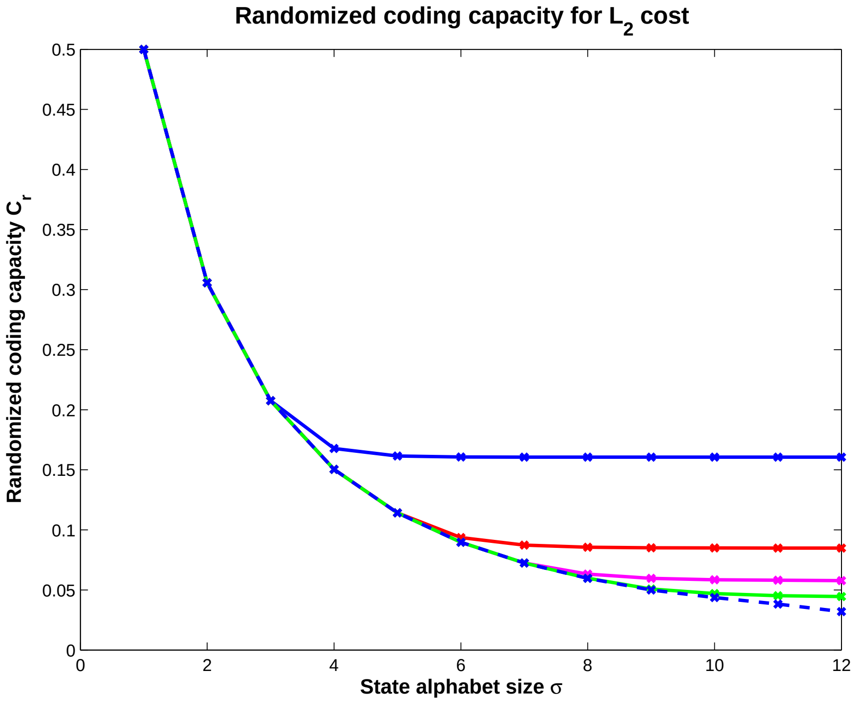


Figure 2.3: Randomized coding capacity $C_r(\Lambda)$ versus σ for $\Lambda = 4,\ 8,\ 12,\ 16$ with loss function $l_2(s)$. The dashed line is the randomized coding capacity for the unconstrained case.

By performing an optimization for each value of σ and Λ we can create the plots of the capacities for the L_1 case in Figure 2.2 and L_2 case in Figure 2.3. As the cost constraint Λ is increased, the randomized coding capacity decreases, and for smaller alphabet sizes the constraint becomes inactive. As expected, the L_2 cost constraint is more restrictive as σ increases.

2.4.2 Achievable rates for average-error list decoding

For each list size L we can compute achievable and converse bounds for the list decoding capacity $\overline{C}_L(\Lambda)$. We can achieve the randomized coding capacity $C_r(\Lambda)$ with list size larger that $\tilde{L}_{\mathrm{sym}}(P^*, \Lambda)$. For lists smaller than $L_{\mathrm{sym}}(P^*, \Lambda)$ we will not be able to achieve the randomized coding capacity. We will focus on showing achievable rates for list decoding with fixed list size L.

For each candidate list size L and input distribution P, we must determine if there exists a channel $U : \mathcal{X}^L \to \mathcal{S}$ satisfying (2.50) whose weak symmetrizing cost in (2.53) is less than the constraint Λ. For a fixed P, the set of symmetrizing channels satisfying the weak cost constraint is convex. We can further restrict our attention to symmetrizing U's that are themselves symmetric channels. To see this, fix a tuple $\overline{x} = (x_1, x_2, \ldots, x_L)$ and think of $U(s|\overline{x})$ as a length $|\mathcal{S}|$ column vector. Consider the $|\mathcal{Y}| \times |\mathcal{S}|$ transition matrix $W_{-1} = W(y| - 1, s)$:

$$W_{-1} = \left(\begin{array}{c} I_{|\mathcal{S}|} \\ \hline 0 \end{array} \right) \ . \tag{2.166}$$

For any permutation π of $[L]$, we have

$$W_{-1} U(s|\overline{x}) = \left(\begin{array}{c} U(s|\overline{x}) \\ \hline 0 \end{array} \right) = \left(\begin{array}{c} U(s|\pi\overline{x}) \\ \hline 0 \end{array} \right) = W_{-1} U(s|\pi\overline{x}) \ . \tag{2.167}$$

Therefore $U(s|\overline{x})$ can only depend on the type of $\overline{x}$.

Thus we can write the average channel as

$$\sum_s W(y|x, s) U(s|t) \ , \tag{2.168}$$

where $t \in [0, 1, \ldots, L]$ counts the number of 1's in X^L. The condition that this channel be symmetric can be rewritten as:

$$\sum_s W(y| - 1, s) U(s|t) - \sum_s W(y| + 1, s) U(s|t - 1) \qquad \forall y, t \tag{2.169}$$

where U must satisfy

$$\sum_s l(s) \sum_{t=0}^{L} \binom{L}{t} 2^{-L} U(s|t) \leq \Lambda \ . \tag{2.170}$$

We transform this minimization into a quadratic program in order to solve it more efficiently. Note that (2.169) holds if and only if

$$f(U) = \sum_y \sum_{t=1}^{L} \left(\sum_s W(y|-1,s)U(s|t) - \sum_s W(y|+1,s)U(s|t-1) \right)^2 = 0 \ . \tag{2.171}$$

To determine if the channel is L-symmetrizable, we minimize the function $f(U)$. If $\min f(U) = 0$ then the channel is symmetrizable, and if $\min f(U) > 0$ it is not. If we replace the square in (2.171) with an absolute value function, then we obtain a function similar to that in Lemma 8. Therefore order to calculate the symmetrizability of the channel, we must solve the following program:

$$\begin{aligned}
&\text{minimize} && && (2.172)\\
&\text{subject to} && \sum_s U(s|t) = 1 \quad \forall t && (2.173)\\
& && -U(s|t) \leq 0 \quad \forall s,t && (2.174)\\
& && \sum_s \sum_{t=0}^{L} \binom{L}{t} 2^{-L} l(s) U(s|t) - \Lambda \leq 0 \ . && (2.175)
\end{aligned}$$

This is a quadratic program in the channel U and we can again use fast solving techniques to find $\tilde{L}_{\text{sym}}(P,\Lambda)$ for different P and Λ.

The plot in Figure 2.4 shows $\max I(P,\Lambda)$ versus Λ for $l_2(\cdot)$ and $\sigma = 8$ under list-decoding codes of fixed list sizes. As Λ increases, the capacity-achieving input distribution with $P(X=1) = 1/2$ becomes L-symmetrizable for small L. However,

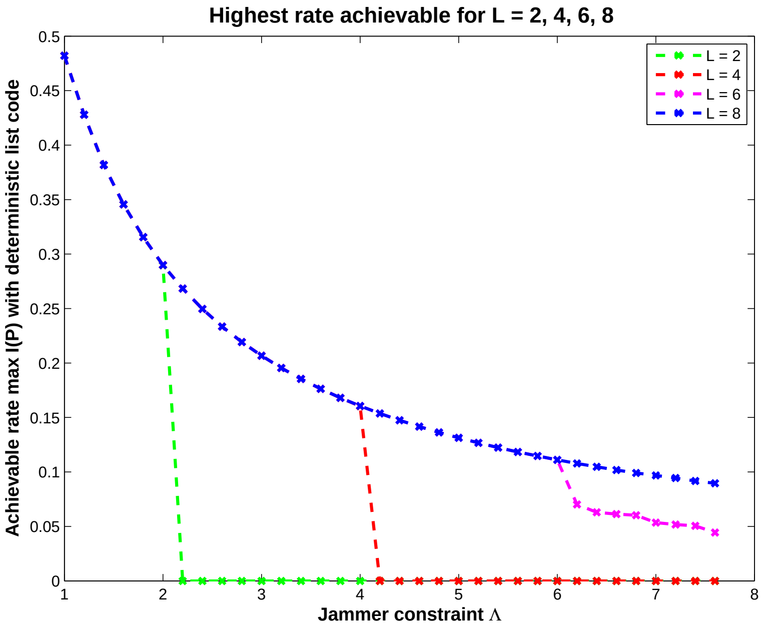


Figure 2.4: The largest value of $I(P, \Lambda)$ achievable versus Λ for $l(s) = |s|^2$, $\sigma = 8$, and for different list sizes.

suboptimal input distributions are not weakly L-symmetrizable, and list codes of size L can still achieve some rates below $C_r(\Lambda)$. In Figure 2.5 we show $\operatorname{argmax} I(P, \Lambda)$ for the distributions P that are not L-symmetrizable.

The extensive analysis in [134] found that the worst case power-constrained *noise* for binary modulation over an additive noise channel had support only on integer points. In this example, we are interested in the interplay between the list size, achievable rates, and cost constraint. In order to compute the random coding capacity we need to find the worst-case noise distribution, but this capacity is not necessarily realizable with deterministic codes. List decoding relaxes the coding problem and already approaches the performance of randomized coding for quite small list sizes. This is in contrast to the maximal error list decoding in Example 2.1 on page 49, where the list size needs to be quite large to get close to the capacity.

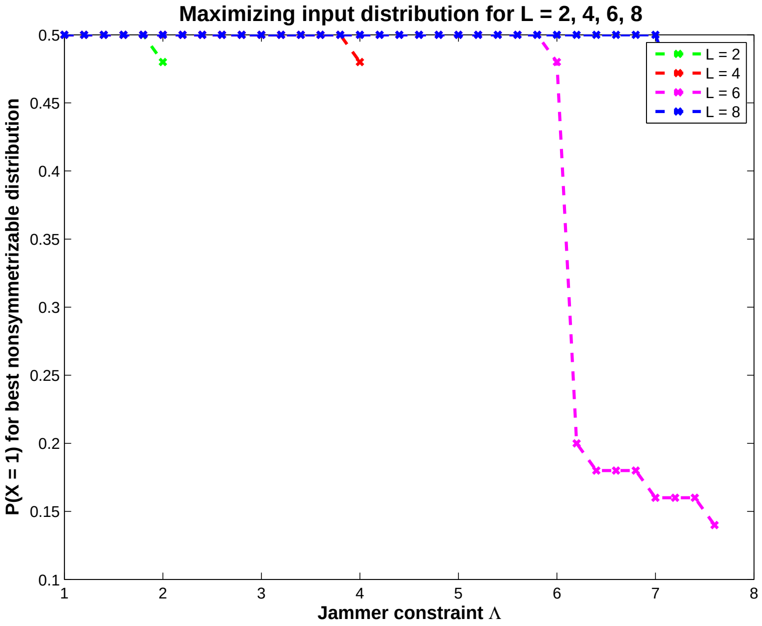


Figure 2.5: The value of $P(X=1)$ which is non-symmetrizable and achieves the highest rate versus Λ for $l(s)=|s|^2$, $\sigma=8$, and different list sizes.

2.5 Discussion

In this chapter we found bounds on the capacity for list coding over constrained AVCs under both maximal and average error. Under maximal error, list decoding with list sizes L could achieve rates within a gap of $O(L^{-1})$ of $C_{\mathrm{dep}}(\Gamma,\Lambda)$. In the next chapter we will use the results on list decoding for maximal error to find the randomized coding capacity under the nosy noise error model for general AVCs, generalizing the results of Langberg [101] on the bit-flipping channel.

Under average error, the list coding capacity behaves differently depending on whether there are constraints on the jammer. For sufficiently large L the list coding capacity equals $C_r(\Gamma,\Lambda)=C_{\mathrm{std}}(\Gamma,\Lambda)$, and for smaller L we may be able to achieve smaller rates. It may be possible to further generalize the results on average error to include multiple constraints or general alphabets. The techniques developed by

Csiszár [42] could be applicable in this approach.

If we restrict our attention to linear codes, a connection between the notion of symmetrizability for list codes and generalized Hamming weights [154, 153] has been shown by Guruswami [76] for the case of list decoding from erasures. The r-th generalized Hamming weight $d_r(\mathcal{C})$ of a code $\mathcal{C}$ is the minimum weight for the basis of an r-dimensional subcode of $\mathcal{C}$. For erasure channels, a list code $\mathcal{C}$ can correct Λn errors with a list of size L if and only if $d_r(\mathcal{C}) > \Lambda n$ for $r = 1 + \lfloor \log n \rfloor$. The converse argument is similar to that for the average-error AVC – the error pattern can simulate r codewords if $d_r(\mathcal{C}) < \Lambda n$. It would be interesting to see how strong this connection is for more general AVCs.

It's di- lem- ma it's de- li- mit it's de- luxe it's de- love- ly!

– *Anything Goes*, Cole Porter

Chapter 3

Derandomization for discrete AVCs

3.1 Introduction

Early work on arbitrarily varying channels dealt with randomized coding from a game theoretic perspective, but the engineering question of how much common randomness is needed to enable randomized coding has received less research attention. The randomized encoding scheme described by Blackwell, Breiman, and Thomasian [28] required that the encoder and decoder choose a random codebook prior to each transmission. The corresponding amount of private information needed by the encoder and decoder in order to carry out the coding scheme is therefore equal to the entropy of a random variable on the set of iid codebooks. This key size dwarfs the amount of information that is actually transmitted over the channel, which suggests that randomized coding in this form may have very little practical engineering significance.

Research interest turned to deterministic coding theorems for AVCs in the work of Kiefer, Wolfowitz, and Ahlswede [95, 19, 20]. Ahlswede proposed the "elimination technique" [6] to partially derandomize the fully randomized codes of Blackwell et al., showing that it was sufficient for the encoder and decoder to choose one of n^2 codebooks. His results imply that a key size of $2 \log n$ bits is sufficient to achieve randomized coding capacity $C_r = C_{\text{dep}}$. For unconstrained AVCs he used this result to show that the deterministic coding capacity $\overline{C}_d$ is equal to C_r, provided $\overline{C}_d > 0$. Ahlswede's technique is to treat the fully randomized code of Blackwell et al. as a codebook-valued random variable $\mathbf{B}$. He constructs a randomized code by drawing n^2 samples from $\mathbf{B}$. The probability (over the sampling) that the randomized code on n^2 codebooks also has small probability of error can be made arbitrarily close to 1. Therefore such a randomized code must exist. Note that the jammer knows the collection of n^2 deterministic codes making up this new randomized code, but does not know the key. That is, it does not know which codebook was chosen by the encoder and decoder. If $\overline{C}_d > 0$ then the encoder can choose a codebook and use a short prefix to inform the decoder of this choice.

Ericson [59] used Ahlswede's method to investigate error-exponents for AVCs. To attain exponential decay in the probability of error using this method requires a key size of $O(n)$ bits, so the amount of shared common randomness is on the same order as the data to be transmitted. This is reminiscent of Shannon's "one-time pad" [136] for cryptography. Hughes and Thomas [88, 151] used a different code ensemble to find error exponents without bounding the amount of common randomness. In this chapter we apply Ahlswede's technique to the randomized codes of Hughes and Thomas to find a range of tradeoffs between randomization and error decay. This kind of tradeoff may be more useful in engineering applications in which sharing $O(n)$ bits of key to send $O(n)$ bits of data is unreasonable.

For the nosy noise error criterion, where the jammer has access to the transmitted codeword as well as the message, the elimination technique is no longer applicable. Instead, we can adopt the approach taken by Langberg for bit-flipping AVCs [101], which uses maximal error list codes in conjunction with a combinatorial construction due to Erdös, Frankl, and Füredi [56] to show that all rates below $1 - h_b(\Lambda)$ can be achieved with only $O(\log n)$ shared random bits for the key. The combinatorial construction does not depend on the channel, so we can use the same construction together with the maximal error list codes from the previous chapter to generalize Langberg's result to all AVCs.

In this chapter we provide achievable tradeoffs between error and key size, but we do not have matching converse bounds showing that a certain amount of randomization is required to achieve a given error decay.

3.2 "Elimination" for standard AVCs

We will first describe a partial derandomization method for constrained AVCs using the "elimination technique" of Ahlswede.

3.2.1 Subsampling a random code

In the "elimination technique," we start with a randomized code that has good properties and then sample from this code to obtain a smaller ensemble of codes that still has good properties. The following lemma gives this result in a form that will be convenient for us in the sequel. A proof is included for completeness.

Lemma 12 (Elimination technique [6]). *Let J be a positive integer and let $\mathbf{C}$ be an (n, N, J) randomized code with $N = \exp(nR)$ whose expected maximal error satisfies*

$$\max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \max_i \mathbb{E}_{\mathbf{C}}[\varepsilon(i, \mathbf{s})] \leq \delta(n) \ , \tag{3.1}$$

for an AVC $\mathcal{W}$ with cost function $l(\cdot)$ and cost constraint Λ. Then for all μ satisfying:

$$\mu \log \delta(n)^{-1} - h_b(\mu) \log 2 > \frac{n}{K}(R \log 2 + \log |\mathcal{S}|) \ , \tag{3.2}$$

where $h_b(\mu)$ is the binary entropy function, we can find a (n, N, K) randomized code with maximal probability of error less than μ.

Proof. Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$ be K codebooks drawn according to the random variable $\mathbf{C}$. Let $\varepsilon(\mathcal{C}, i, \mathbf{s})$ be the error on message i for state sequence $\mathbf{s}$ under the deterministic code $\mathcal{C}$. The assumption on $\mathbf{C}$ is then that $\mathbb{E}_{\mathbf{C}}[\varepsilon(\mathcal{C}, i, \mathbf{s})] \leq \delta(n)$ for all i and $\mathbf{s} \in \mathcal{S}^n(\Lambda)$. Then we can use Bernstein's trick (exponentiating both sides and using the Markov inequality):

$$\begin{aligned}
\mathbb{P}_{\mathbf{C}}\left(\frac{1}{K} \sum_{k=1}^{K} \varepsilon(\mathcal{C}_k, i, \mathbf{s}) \geq \mu \right) &= \mathbb{P}_{\mathbf{C}}\left(\exp\left(r \sum_{k=1}^{K} \varepsilon(\mathcal{C}_k, i, \mathbf{s}) \right) \geq \exp(K r \mu) \right) \\
&\leq \exp(-K r \mu) \mathbb{E}_{\mathbf{C}} \left[\exp(r \varepsilon(\mathcal{C}_k, i, \mathbf{s}) \right]^K \\
&= \exp(-K r \mu) \mathbb{E}_{\mathbf{C}} \left[1 + \sum_{m=1}^{\infty} \frac{1}{m!} r^m \varepsilon(\mathcal{C}_k, i, \mathbf{s})^m \right]^K \\
&\leq \exp(-K r \mu) \left(1 + \delta(n) \sum_{m=1}^{\infty} \frac{1}{m!} r^m \right)^K \\
&\leq \exp(-K r \mu) \left(1 + \delta(n) e^r \right)^K \\
&= \exp\left(-K(r\mu - \log(1 + \delta(n) e^r)) \right) \ .
\end{aligned} \tag{3.3}$$

Taking a union bound over all $\mathbf{s}$ and $N = \exp(nR \log 2)$ messages we obtain:

$$\begin{aligned}
&\mathbb{P}_{\mathbf{C}}\left(\frac{1}{K} \sum_{k=1}^{K} \varepsilon(\mathcal{C}_k, i, \mathbf{s}) \geq \mu, \ \ \forall s, \ i \in [N] \right) \\
&\qquad \leq \exp\left(-K(r\mu - \log(1 + \delta(n) e^r)) + n(R \log 2 + \log |\mathcal{S}|) \right) \ ,
\end{aligned} \tag{3.4}$$

where R is measured in bits.

Therefore the probability (over the sampling of $\mathbf{C}$) that the maximal error is greater than μ for the randomized code uniformly distributed on $\{\mathcal{C}_k : k \in [K]\}$ can be made as small as we like as long as $K(r\mu - \log(1 + \delta(n)e^r))$ grows faster than n. We can optimize over r:

$$\frac{d}{dr}(r\mu - \log(1 + \delta(n)e^r)) = \mu - \frac{\delta(n)e^r}{1 + \delta(n)e^r} \ . \tag{3.5}$$

This gives $r = \log(\mu/((1-\mu)\delta))$, which yields the condition for the probability in (3.4) to go to 0:

$$K\left(\mu \log \delta^{-1} - h_b(\mu) \log 2\right) > n(R \log 2 + \log |\mathcal{S}|) \ , \tag{3.6}$$

where $h_b(\mu)$ is the the binary entropy function of μ in bits. This shows that the probability (over sampling K codebooks) that the maximal error of our new randomized code exceeds μ can be made as small as we like, and therefore a code whose error is bounded by μ must exist. □

3.2.2 Partial derandomization for standard AVCs

As noted by Ericson [59], Csiszár and Narayan [45], and Hughes and Thomas [88], the first step of the elimination technique [6] can be used to reduce the randomization needed for a randomized code. The capacity under this model is the randomized coding capacity $C_r(\Gamma, \Lambda)$ of the AVC. Previous authors were concerned with the case where the key size is exponential in the blocklength, which gives an error probability that decays exponentially in the blocklength. Here we look at more general error key-size tradeoffs which provide a range of operating points.

Theorem 13. *Let $\mathcal{W}$ be an AVC with cost function $l(\cdot)$ and cost-constraint Λ. For* *any* $\epsilon > 0$ *and* $\zeta > 0$*, there exists a sequence of* $(n, \exp(nR), K(n))$ *randomized code, where the key size* $K(n)$ *satisfies* $K(n)/n \to \infty$ *and* $n^{-1} \log(K(n)/n) \to 0$ *and the rate and error satisfy*

$$R = C_r(\Gamma, \Lambda) - \epsilon \tag{3.7}$$

$$\varepsilon(n) = \zeta \frac{n}{K(n)} \ . \tag{3.8}$$

Proof. Fix $\epsilon > 0$. Hughes and Thomas [88, Theorem 6] show that for n sufficiently large and $N = \exp(n(C_r(\Gamma, \Lambda) - \epsilon))$, there exists an $(n, N, n!)$ randomized code $\mathbf{C}$ with codewords of type P^* and maximal error

$$\varepsilon_\pi(\mathbf{C}) \leq \exp\left(-n\left(E_r(C_r(\Gamma, \Lambda) - \epsilon/2, P^*, \Lambda) - \epsilon/2\right)\right) \ , \tag{3.9}$$

where E_r is the random coding exponent given in (1.75). This randomized code uses an (n, N) deterministic code $\mathcal{C} = \{\mathbf{x}(i) : i \in [N]\}$ of constant type P^* and the keys are all permutations Π_n on $[n]$. The encoder encodes message i using permutation π_k by transmitting $\pi_k(\mathbf{x}(i))$.

Thus we can apply Lemma 12 to this code to show that there exists a set of K permutations $\{\pi_1, \ldots, \pi_K\}$ such that $\{\pi_k \mathcal{C} : k \in [K]\}$ is a randomized codebook whose error $\varepsilon(n)$ satisfies

$$\begin{aligned} \frac{K(n)}{n} \left(n\varepsilon(n)(E_r(C_r(\Gamma, \Lambda) - \epsilon/2, P^*, \Lambda) - \epsilon/2) - h_b(\varepsilon(n)) \log 2\right) \\ > \left(R \log 2 + \log |\mathcal{S}|\right) \ . \end{aligned} \tag{3.10}$$

Now let $\varepsilon(n) = \zeta n / K(n)$. We will show that for n sufficiently large (3.10) is satisfied.

First, for small $\varepsilon(n)$ we have an upper bound:

$$\begin{aligned} h_b(\varepsilon(n)) &= -\varepsilon(n) \log \varepsilon(n) - (1 - \varepsilon(n)) \log(1 - \varepsilon(n)) \qquad (3.11) \\ &\leq -\varepsilon(n) \log \varepsilon(n) + 2\varepsilon(n) \ . \qquad (3.12) \end{aligned}$$

Then:

$$\begin{aligned} &\frac{K(n)}{n} \left(n\varepsilon(n) \left(E_r \left(C_r(\Gamma, \Lambda) - \frac{\epsilon}{2}, P^*, \Lambda \right) - \frac{\epsilon}{2} \right) - h_b(\varepsilon(n)) \log 2 \right) \\ &\quad \geq \frac{K(n)}{n} \Bigg(n\varepsilon(n) \left(E_r \left(C_r(\Gamma, \Lambda) - \frac{\epsilon}{2}, P^*, \Lambda \right) - \frac{\epsilon}{2} \right) - \varepsilon(n) \log \frac{1}{\varepsilon(n)} \log 2 \\ &\qquad\qquad - 2\varepsilon(n) \log 2 \Bigg) \\ &\quad = \zeta n \left(E_r(C_r(\Gamma, \Lambda) - \epsilon/2, P^*, \Lambda) - \epsilon/2 - \frac{\log 2}{n} \log \frac{K(n)}{n\zeta} - \frac{2 \log 2}{n} \right) \ . \qquad (3.13) \end{aligned}$$

For n sufficiently large we can make this as large as we like, so we can satisfy (3.10).

□

We can use this theorem to characterize different key-error tradeoffs:

- Suppose $K(n) = n^{\alpha}$ for $\alpha > 1$. Then the condition (3.8) becomes:

$$\varepsilon(n) = O(n^{1-\alpha}) \ . \qquad (3.14)$$

- Suppose $K(n) = \exp(\beta n)$. Reexamining (3.13) shows that the theorem holds if $\beta \log 2 < E_r(R, P^*, \Lambda) - \epsilon$. Then (3.8) becomes:

$$\varepsilon(n) = O(\exp(-\beta' n)) \ , \qquad (3.15)$$

  where $\beta' < \beta$.

- Suppose $K(n) = \exp(n^{\gamma})$ for $\gamma \in (0,1)$. Then (3.8) shows:

$$\varepsilon(n) = O(\exp(-n^{\gamma})) \ . \tag{3.16}$$

3.3 Channels with nosy "noise"

We now turn to the nosy noise model, in which the jammer can choose $\mathbf{s}$ with knowledge of the transmitted codeword $\mathbf{x}$. For completeness, we provide a proof of the combinatorial construction used by Langberg in his proof of the bit-flipping AVC [56,101]. This can be used with our results on maximal error list decoding from the previous chapter to show that the randomized coding capacity under this error criterion is

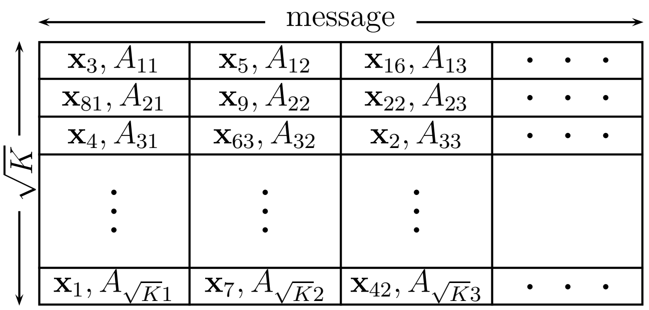
$$\hat{C}_r(\Gamma, \Lambda) = C_{\text{dep}}(\Gamma, \Lambda) = \max_{P \in \mathcal{I}(\Gamma)} \min_{V \in \mathcal{W}_{dep}(P,\Lambda)} I\left(P, V\right) \ . \tag{3.17}$$

As an application, we can partially derandomize the construction of Agarwal, Sahai, and Mitter [1] for distortion-constrained channels.

3.3.1 Derandomization via list-decodable codes

The construction given in the following Lemma has been used by Langberg [101] and Smith [143] to construct randomized codes for constrained bit-flipping AVCs in which the codeword is known to the jammer. By using our new list codes we can construct such randomized codes for general AVCs.

Lemma 13 (Message Authentication [56, 101])**.** *Let $\mathcal{W}$ be an AVC and suppose we are given an (n, N, L) deterministic list-decodable code and probability of error ϵ. For key size $K(n)$ where $K(n)$ is a power of a prime there exists an $(n, N/\sqrt{K(n)}, K(n))$*



$\mathbf{x}_3, A_{11}$	$\mathbf{x}_5, A_{12}$	$\mathbf{x}_{16}, A_{13}$	$\cdots$
$\mathbf{x}_{81}, A_{21}$	$\mathbf{x}_9, A_{22}$	$\mathbf{x}_{22}, A_{23}$	$\cdots$
$\mathbf{x}_4, A_{31}$	$\mathbf{x}_{63}, A_{32}$	$\mathbf{x}_2, A_{33}$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	
$\mathbf{x}_1, A_{\sqrt{K}1}$	$\mathbf{x}_7, A_{\sqrt{K}2}$	$\mathbf{x}_{42}, A_{\sqrt{K}3}$	$\cdots$

Figure 3.1: Constructing a randomized code from a list-decodable code. We put the codewords of the list code into a $\sqrt{K} \times N/\sqrt{K}$ table. Each column has a partition of the set of K keys into sets A_{ij} of $\sqrt{K}$ keys each. The intersection of the key sets is small.

randomized code with probability of error $\epsilon + \epsilon'$, *where*

$$\epsilon' = \frac{2L \log N(n)}{\sqrt{K(n)} \log K(n)} \ . \tag{3.18}$$

Proof. Let $\mathcal{C}_L = \{\mathbf{x}(l) : l \in [N]\}$ be the codebook of the (n, N, L) list-decodable code. To construct a randomized code from $\mathcal{C}_L$, we will associate a key k with a subset $\mathcal{C}_k$ of the codewords of $\mathcal{C}_L$. We first place the codewords in $\mathcal{C}_L$ arbitrarily in an array as shown in Figure 3.1 and then we will associate a set of keys to each codeword. The array has $\sqrt{K}$ rows and $N/\sqrt{K}$ columns. In column j we will make a partition of the set of K keys into $\sqrt{K}$ sets $\{A_{ij} : i \in [\sqrt{K}]\}$, where $|A_{ij}| = \sqrt{K}$ for all i and j. To encode message j using key k in the $(n, N/\sqrt{K(n)}, K(n))$ randomized code, we find the i such that $k \in A_{ij}$ and output the codeword associated to the set of keys A_{ij}.

Let $R' = n^{-1} \log(N/\sqrt{K})$ and assume that $\sqrt{K}$ is a power of a prime number. Let i and z be elements of the finite field $GF(\sqrt{K})$ with $\sqrt{K}$ elements, and let the key be given by the pair (i, z). For a positive integer d to be chosen later, let $\{f_j(\cdot) : j = 1, 2, \ldots, 2^{nR'}\}$ of $2^{nR'}$ be a set of distinct monic polynomials of degree

$d-1$ over $GF(\sqrt{K})$. Then let

$$A_{ij} = \left\{ (i, z f_j(z) + i) : z = 1, 2, \ldots, \sqrt{K} \right\} . \tag{3.19}$$

Since i acts as a constant shift of the polynomial $zf_j(z)$, it is clear that for each j the collection $\{A_{ij} : i = 1, 2, \ldots, \sqrt{K}\}$ is a partition of the set of all keys. Furthermore, for $j' \neq j$ we have $|A_{ij} \cap A_{ij'}| \le d$, since $f_j(z) = f_{j'}(z)$ for at most d values of z.

For codeword $\mathbf{x}(l)$, let $I(l)$ and $J(l)$ be the row and column index for the position of $\mathbf{x}(l)$ in the array. The encoder takes a message j and key (i, z) and outputs the codeword of the list code in the (i, j)-th position in the table. The decoder for the randomized code first decodes using the list code $\mathcal{C}_L$ to find a list of at most L candidate codewords $\{\mathbf{x}_{l(1)}, \mathbf{x}_{l(2)}, \ldots, \mathbf{x}_{l(L)}\}$. These codewords have associated key sets $\{A_{I(l),J(l)} : l \in [L]\}$ given by the table. If a unique $m \in [L]$ exists such that $I(l(m)) = i$ and $(i, z) \in A_{I(l(m)),J(l(m))}$, then the decoder outputs the message $J(l(m))$ associated with codeword $\mathbf{x}_{l(m)}$. If no m can be found or m is not unique then it declares an error.

There are two possible decoding errors. If the list code has a decoding error then the correct codeword will not be in the list and so the decoder for the randomized code will fail. This happens with probability smaller than ϵ by the assumptions on the list code. If the transmitted codeword is in the list produced by the list decoder, then we will have an error if there is is another $m' \in [L]$ for which $I(l(m')) = i$ and $(i, z) \in A_{I(l(m')),J(l(m'))}$. We know $|A_{ij} \cap A_{ij'}| \le d$, so there are at most Ld values of (i, z) for which this can happen. Since the jammer knows i and there are $\sqrt{K}$ values for z, the probability that the key cannot disambiguate the list is at most $\epsilon' = Ld/\sqrt{K}$. The total error probability is then bounded by $\epsilon + \epsilon'$.

The last part is to choose d appropriately. There are $\sqrt{K}^{d-1}$ monic polynomials of degree $d-1$ over $GF(\sqrt{K})$, so we need

$$\sqrt{K}^{d-1} \geq \frac{N}{\sqrt{K}} \ . \tag{3.20}$$

This in turn implies

$$d \geq \frac{\log N}{\log \sqrt{K}} \ . \tag{3.21}$$

Substituting this into the expression for ϵ' in the previous paragraph we obtain (3.18).

□

3.3.2 The capacity of channels with nosy noise

For AVCs with nosy noise, the state can depend on the transmitted codeword. By using the list-decodable codes for cost constrained AVCs from Theorem 10 and combining them with a message authentication scheme used by Langberg [101], we can construct randomized codes for this channel with limited common randomness.

Theorem 14. *Let $\mathcal{W}$ be an AVC with input and state cost functions $g(\cdot)$ and $l(\cdot)$ and cost constraints Γ and Λ. For any $\epsilon > 0$, there exists an n sufficiently large such that the sequence of rate-key size pairs $(R, K(n))$ is achievable with error $\hat{\varepsilon}_r(n)$, where $K(n) \leq \exp(n\epsilon)$ and*

$$R = C_{\text{dep}}(\Lambda, \Gamma) - \epsilon \tag{3.22}$$

$$\hat{\varepsilon}(n) \leq \exp(-n\hat{E}(\epsilon)) + \frac{12nC_{\text{dep}}(\Lambda,\Gamma)\log|\mathcal{Y}|}{\epsilon\sqrt{K(n)}\log K(n)} \ , \tag{3.23}$$

where $\hat{E}(a) > 0$ for $a > 0$. Here $C_{\text{dep}}(\Lambda, \Gamma)$ is given by (1.39) and is the randomized *coding capacity of the AVC with nosy noise:*

$$\hat{C}_r(\Lambda, \Gamma) = C_{\text{dep}}(\Lambda, \Gamma) \ . \tag{3.24}$$

Proof. We can use the previous lemma with our result on list codes to achieve the desired tradeoff. Using Lemma 3, for any $\epsilon_1(n) > 0$ we can choose an $(n, N(n), L)$ codebook with

$$L = \left\lfloor \frac{6 \log |\mathcal{Y}|}{\epsilon_1(n)} \right\rfloor + 1 \tag{3.25}$$

$$N(n) = L \exp(n(C_{\text{dep}}(\Lambda, \Gamma) - \epsilon_1(n))) \ , \tag{3.26}$$

and error

$$\varepsilon_L \leq \exp(-nE(\epsilon_1(n))) \ . \tag{3.27}$$

We can use Lemma 13 to construct an $(n, N(n)/\sqrt{K(n)}, K(n))$ randomized code with error probability

$$\hat{\varepsilon} \leq \exp(-nE(\epsilon_1(n))) + \frac{2L \log N(n)}{\sqrt{K(n)} \log K(n)} \tag{3.28}$$

$$< \exp(-nE(\epsilon_1(n))) + \frac{12nC_{\text{dep}}(\Lambda, \Gamma) \log |\mathcal{Y}|}{\epsilon_1(n) \sqrt{K(n)} \log K(n)} \ . \tag{3.29}$$

The rate of this randomized code is

$$R = \frac{1}{n} \log \frac{N(n)}{\sqrt{K(n)}} \tag{3.30}$$

$$= C_{\text{dep}}(\Lambda, \Gamma) - \epsilon_1(n) - \frac{1}{n} \log \frac{\sqrt{K(n)}}{L} \ . \tag{3.31}$$

For any $\epsilon > 0$ and $K(n) \leq \exp(n\epsilon)$ we can choose $\epsilon_1(n)$ small enough so that $R =$

$C_{\mathrm{dep}}(\Lambda, \Gamma) - \epsilon$.

Finally, we note that the jammer can choose a memoryless strategy $U(s|x) \in \mathcal{U}(P, \Lambda)$. Choosing the worst U yields a discrete memoryless channel whose capacity is $C_{\mathrm{dep}}(\Lambda, \Gamma)$, and therefore the randomized coding capacity for this channel is given by $C_{\mathrm{dep}}(\Lambda, \Gamma)$. □

This theorem gives some tradeoffs between error decay, key size, and rate loss. We could also phrase the result by fixing $K(n)$ first and finding the corresponding expressions. Now we look at some examples of $K(n)$ scalings and the associated error probability.

- Suppose $K(n) = n^{\alpha}$. Then the condition (3.23) becomes:

$$\hat{\varepsilon}(n) = O(n^{1-\alpha/2}) \ . \tag{3.32}$$

- Suppose $K(n) = \exp(\beta n)$. Then the condition (3.23) becomes:

$$\hat{\varepsilon}(n) \leq \exp(-nE(\epsilon)) + \frac{12 n C_{\mathrm{dep}}(\Lambda, \Gamma) \log |\mathcal{Y}|}{\epsilon \beta n \exp(n\beta/2)} \ . \tag{3.33}$$

  Therefore $\hat{\varepsilon}(n) = O(\exp(-n \min(E(\epsilon), \beta/2)))$.

- Suppose $K(n) = \exp(\beta n^{\gamma})$ for $\gamma \in (0, 1)$. Then the condition (3.23) becomes:

$$\hat{\varepsilon}(n) \leq \exp(-nE(\epsilon)) + \frac{12 n C_{\mathrm{dep}}(\Lambda, \Gamma) \log |\mathcal{Y}|}{\epsilon \beta n^{\gamma} \exp(n^{\gamma}\beta/2)} \ . \tag{3.34}$$

  Therefore $\hat{\varepsilon}(n) = O(\exp(-n^{\gamma}))$.

3.3.3 A connection to rate distortion

In Section 1.4 we described a channel model used by Agarwal, Sahai, and Mitter [1] in which there is a distortion function $d(x, y)$ between the input x and output y. The channel is allowed to make any mapping from an input codeword $\mathbf{x}$ to output $\mathbf{y}$ subject to an average distortion constraint D. The input is fixed to have distribution P. They then show that the capacity of this channel is the rate distortion function $R(P, D)$ for a source distributed according to P with distortion measure $d(x, y)$.

In an AVC the channel at each time is drawn from a bag of channels indexed by a *state*, whereas in the rate-distortion model model the channel can be thought of as taking an *action* mapping $x \to y$. Rather than putting a cost on a state, the constraint is on the mapping. It is straightforward to create an AVC whose randomized coding capacity under the "nosy noise" error model is equal to the rate distortion function $R(p, D)$. An analogous result was claimed independently by Moulin and Wang [116], but their construction is not a well-defined AVC and they did not give a proof.

Theorem 15. *Let $P \in \mathcal{P}(\mathcal{X})$ be a given input distribution, $d(\cdot, \cdot)$ be a distortion measure on $\mathcal{X} \times \mathcal{Y}$, and D a given distortion constraint. Then there exists an AVC $\mathcal{W} = \{W(z|x, s) : s \in \mathcal{S}\}$ with input in $\mathcal{X}$, output in $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$, state set $\mathcal{S} = \mathcal{X} \times \mathcal{Y}$, cost function $l(\cdot)$, and cost constraint D such that*

$$\min_{V \in \mathcal{W}_{dep}(D)} I\left(P, V\right) = R_d(P, D) \ . \tag{3.35}$$

That is, the rate distortion function for P is the maximum rate achievable over this AVC with input distribution P under the nosy noise error criterion.

Proof. Define the output set $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$ and state set $\mathcal{S} = \mathcal{X} \times \mathcal{Y}$. For $s = (s_x, s_y) \in \mathcal{S}$, define the cost function $l(s) = l(s_x, s_y) = d(s_x, s_y)$. Finally, define the AVC $\mathcal{W} =$

$\{W(z|x,s) : s \in \mathcal{S}\}$ to be

$$W(z|x,(s_x,s_y)) = \begin{cases} \mathbf{1}(s_y) & s_x = x \\ \mathbf{1}(x) & s_x \neq x \end{cases} \tag{3.36}$$

In this AVC, the jammer "wastes" some of its cost constraint by choosing an (s_x, s_y) such that $s_x \neq x$, because choosing such s reveals the input value x to the decoder. In order to show (3.35) we will show that optimizing $I(P,V)$ over $V \in \mathcal{W}_{dep}(D)$ gives the minimizing test channel for the rate distortion problem with source P and distortion $d(x,y)$.

Let $U \in \mathcal{U}(P,D)$ be a jamming channel and let $V \in \mathcal{W}_{dep}(D)$ be the average channel under this jamming strategy:

$$V(z|x) = \sum_{s_x,s_y} W(z|x,(s_x,s_y))U((s_x,s_y)|x) \ . \tag{3.37}$$

We can simplify V using the definition of W. Note that for an input x, the output z is either x or in $\mathcal{Y}$. Thus we can write:

$$V(x|x) = \sum_{s_x \neq x} \sum_{s_y} U((s_x,s_y)|x) \tag{3.38}$$

$$V(y|x) = U(x,y|x) \qquad y \in \mathcal{Y} \ . \tag{3.39}$$

Now suppose $\tilde{V}(y|x)$ is a test channel satisfying the distortion constraint:

$$\sum_{x,y} \tilde{V}(y|x)P(x)d(x,y) \leq D \ . \tag{3.40}$$

We define $\tilde{U}(s_x, s_y|x) = \tilde{V}(s_y|x)$ for $s_x = x$ and 0 otherwise. This shows that the set

$\mathcal{W}_{dep}(D)$ contains all valid test channels, so

$$\min_{V \in \mathcal{W}_{dep}(D)} I\left(P, V\right) \leq R_d(P, D) \ . \tag{3.41}$$

Let X and Z denote random variables for the input and output of the AVC $\mathcal{W}$ and let $A = \mathbf{1}(Z \in \mathcal{Y})$. Then we can lower bound the mutual information $I\left(X \ \wedge \ Z\right)$ using the nonnegativity of mutual information:

$$\begin{aligned} I\left(X \ \wedge \ Z\right) &= I\left(X \ \wedge \ Z, A\right) & (3.42)\\ &= I\left(X \ \wedge \ Z|\, A\right) + I\left(X \ \wedge \ A\right) & (3.43)\\ &\geq I\left(X \ \wedge \ Z|\, A = 0\right) \mathbb{P}(A=0) + I\left(X \ \wedge \ Z|\, A=1\right) \mathbb{P}(A=1) \ . & (3.44) \end{aligned}$$

Now, when $A = 0$ we have $Z = X$, so $I\left(X \ \wedge \ Z|\, A = 0\right) = H(X)$. When $A = 1$ we have

$$\mathbb{P}(X = x, Z = y | A = 1) = \frac{1}{\mathbb{P}(A=1)} U(x, y|x) P(x) \ . \tag{3.45}$$

Therefore the conditional joint distribution satisfies the distortion constraint:

$$\begin{aligned} \sum_{x,y} d(x,y) \mathbb{P}(X = x, Z = y | A = 1) &= \frac{1}{\mathbb{P}(A=1)} \sum_{x,y} d(x,y) U(x,y|x) P(x) & (3.46)\\ &\leq \frac{1}{\mathbb{P}(A=1)} D \ . & (3.47) \end{aligned}$$

Therefore one term of (3.44) can be lower bounded by a rate distortion function:

$$I\left(X \ \wedge \ Z|\, A = 1\right) \geq R_d\left(P, \frac{1}{\mathbb{P}(A=1)} D\right) \ . \tag{3.48}$$

Similarly, we can also lower bound the other term:

$$I\left(X \ \wedge \ Z \middle| \, A=0\right) \geq R_d\left(P, \frac{1}{\mathbb{P}(A=0)} D\right) \ . \tag{3.49}$$

Subsituting (3.48) and (3.49) into (3.44) and using the fact that the rate distortion function is convex in the distortion, we see:

$$I\left(X \ \wedge \ Z\right) \geq R_d\left(P, \frac{1}{\mathbb{P}(A=0)} D\right) \mathbb{P}(A=0) + R_d\left(P, \frac{1}{\mathbb{P}(A=1)} D\right) \mathbb{P}(A=1) \tag{3.50}$$

$$\geq R_d(P, D) \ . \tag{3.51}$$

Since the upper and lower bounds on $I\left(X \ \wedge \ Z\right)$ match, the mutual information of the AVC is equal to the rate-distortion function. □

The preceding theorem shows that any distortion constrained channel can be thought of as an AVC. It may be tempting to think that any AVC can be modeled by a distortion-constrained channel. However, this is not true, as shown by the following example.

Example 3.1 – An AVC that cannot be modeled as a distortion-constrained channel

Consider the AVC with binary inputs and outputs and ternary state $\mathcal{S} = \{0, 1, 2\}$. Under state 0 the channel is a BSC with crossover probability a, under 1 it is a Z-channel with parameter b, and under 2 it is an "S-channel" (reversed Z-channel) with parameter c:

$$V(y|x,0) = \begin{pmatrix} 1-a & a \\ a & 1-a \end{pmatrix} \tag{3.52}$$

$$V(y|x,1) = \begin{pmatrix} 1 & 0 \\ b & 1-b \end{pmatrix} \tag{3.53}$$

$$V(y|x,2) = \begin{pmatrix} 1-c & c \\ 0 & 1 \end{pmatrix} . \tag{3.54}$$

We can assign values to a, b, and c such that this channel is not equivalent to a distortion-constrained channel. The details are somewhat tedious and relegated to Section C.2 .

Because we can construct an equivalent AVC for distortion constrained channels, we can use codes for this AVC under nosy noise on the distortion constrained channel. The following two results follow immediately from out earlier results on list decoding and partial derandomization.

Corollary 1. *Let $P \in \mathcal{P}(\mathcal{X})$ be a given input distribution, $d(\cdot,\cdot)$ be a distortion measure on $\mathcal{X} \times \mathcal{Y}$, and D a given distortion constraint. Then for any $\epsilon > 0$ there is an n sufficiently large and a list code with codeword of type P, rate*

$$R = R_d(P,D) - \epsilon \ , \tag{3.55}$$

list size

$$L < \left\lceil \frac{6 \log |\mathcal{Y}|}{\epsilon} \right\rceil + 1 \ , \tag{3.56}$$

and error

$$\varepsilon_L \leq \exp(-nE(\epsilon)) \ . \tag{3.57}$$

Proof. Lemma 3 gives the list decoding result for the AVC constructed in Theorem 15. □

Corollary 2. *Let $P \in \mathcal{P}(\mathcal{X})$ be an input distribution, $d(\cdot, \cdot)$ a distortion measure on $\mathcal{X} \times \mathcal{Y}$, and D a distortion constraint for the channel given by Agarwal, Sahai, and Mitter. For any $\epsilon > 0$, there exists an n sufficiently large such that the sequence of rate-key size pairs $(R(n), K(n))$ is achievable with error $\hat{\varepsilon}_r(n)$, where*

$$R(n) = R_d(P, D) - \epsilon \tag{3.58}$$

$$\hat{\varepsilon}(n) \leq \exp(-nE(\epsilon)) + \frac{12nR_d(P, D) \log |\mathcal{Y}|}{\epsilon\sqrt{K(n)} \log K(n)} \ , \tag{3.59}$$

where $R_d(P, D)$ is the rate distortion function.

Proof. It is clear that Theorem 14 shows that for any P there exists a channel code of composition P such that

$$g(P) = \min_{V \in \mathcal{W}_{dep}(P,\Lambda)} I\left(P, V\right) \tag{3.60}$$

is achievable for the AVC. We can use this code on the AVC equivalent of the Agarwal-Sahai-Mitter channel. From the definition of their channel, the output will always be in the set $\mathcal{Y}$, so we can restrict the decoder in the AVC equivalent to outputs in $\mathcal{Y}$. From Theorem 15 this code can achieve any rates below $R_d(P, D)$. □

3.4 Examples

We can now revisit some of our earlier examples from Chapter 1 and interpret the plots of C_{std} and C_{dep} in terms of the operational quantities C_r and $\hat{C}_r$. For the bit flipping channel in example 1.4 we saw that for $\Lambda \leq 1/2$:

$$C_r(\Lambda) = C_{\text{std}}(\Lambda) = 1 - h_b(\Lambda) \tag{3.61}$$

$$\hat{C}_r(\Lambda) = C_{\text{dep}}(\Lambda) = 1 - h_b(\Lambda) \ . \tag{3.62}$$

$$\tag{3.63}$$

Thus for this channel there is no difference between the capacities under maximal error and nosy noise:

$$C_r(\Lambda) = \hat{C}_r(\Lambda) = 1 - h_b(\Lambda) \ . \tag{3.64}$$

For other channels there may be a gap between C_{std} and C_{dep}.

3.4.1 A real adder channel

In Example 1.5 we showed that for the binary-input binary-state real adder channel $C_{\text{dep}}(\Gamma, \Lambda) < C_{\text{std}}(\Gamma, \Lambda)$ in general. These two information quantities are equal to the randomized coding capacities under the nosy noise and maximal error criteria. We reproduce in Figure 3.2 the plot from Figure 1.7 to illustrate this gap. As the constraint Λ on the jammer increases the gap between the two capacities $C_r(\Gamma, \Lambda)$ and $\hat{C}_r(\Gamma, \Lambda)$ widens.

Figure 3.3 shows a plot of $\min_{V \in \mathcal{W}_{std}(\Lambda)} I\left(P, V\right)$ for fixed input distribution $(1 - p, p)$. As we can see, for $\Lambda \leq 1/2$ the capacity-achieving input distribution is always $p = 1/2$.

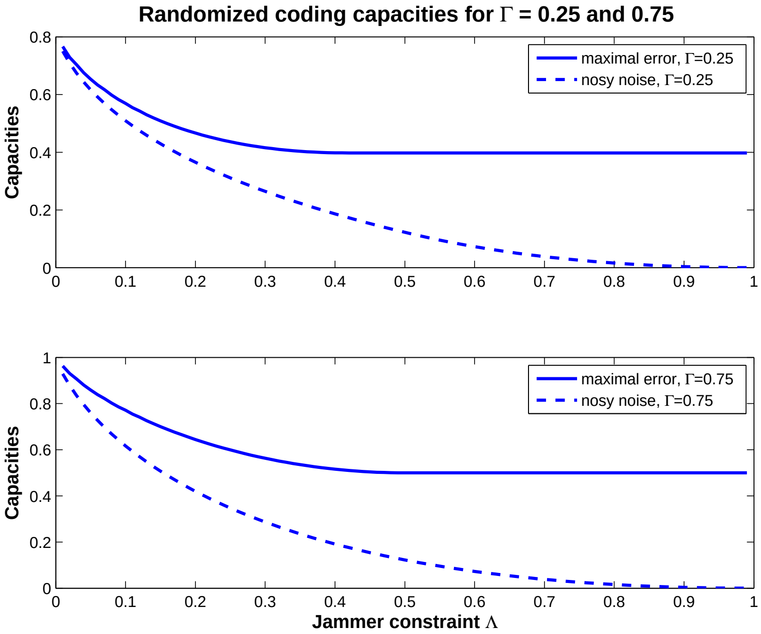


Figure 3.2: For the example in Section 3.4.1, the randomized capacities $C_r(\Gamma, \Lambda)$ (solid line) under maximal error and $\hat{C}_r(\Gamma, \Lambda)$ (dashed line) under nosy noise.

In the case where the state can depend on the input, the situation is less rosy – for an input distribution $P = (1 - p, p)$, if $\Lambda \geq 1 - p$ the jammer can change every 0 in the transmitted codeword into a 1, making the output sequence all 1's and zeroing the capacity. For smaller p the randomized coding capacity may be larger, as shown in Figure 3.4. Another feature of this channel is that the mutual information does not have a saddle point independent of Λ, so the capacity-achieving input distribution will shift as a function of Λ.

3.4.2 BSC mixed with Z-channels

Consider a channel with binary inputs and binary outputs and three states $\mathcal{S} = \{0, 1, 2\}$. For $S = 0$ the channel is a binary symmetric channel with crossover proba-

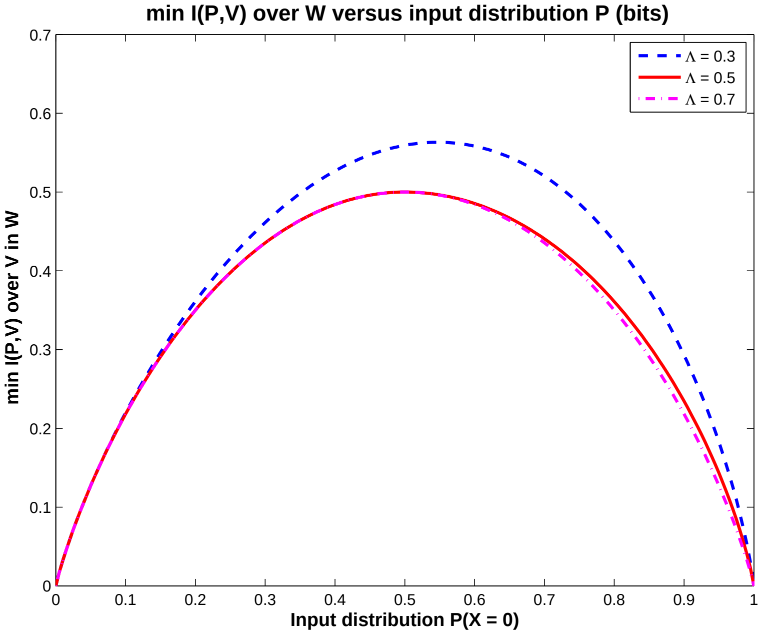


Figure 3.3: For the example in Section 3.4.1, a plot of $\min_{V \in \mathcal{W}_{std}(\Lambda)} I(P, V)$ for $\Lambda = 0.3, 0.5, 0.7$. The unconstrained capacity is $1/2$ and corresponds to $p = 1/2$. For $\Lambda \geq 1/2$ we also have $C_r(\Lambda) = 1/2$.

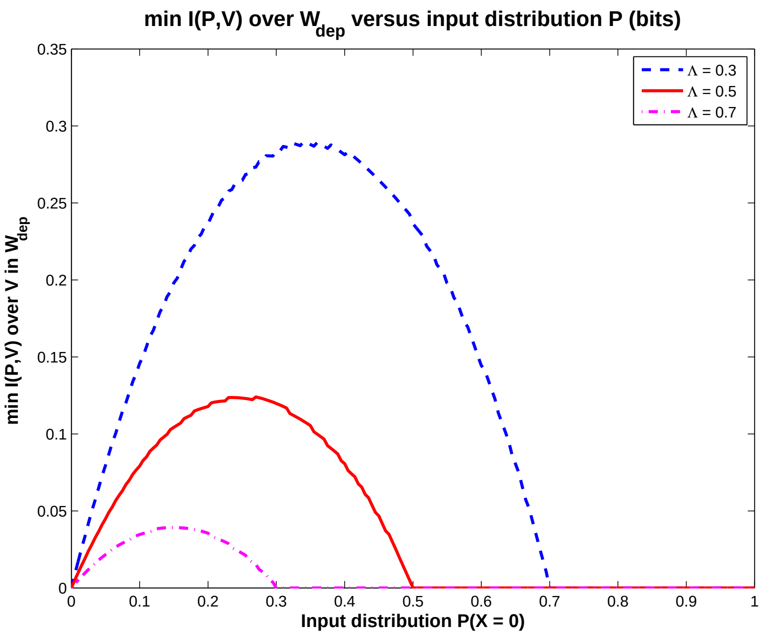


Figure 3.4: For the example in Section 3.4.1, a plot of $\min_{V \in \mathcal{W}_{dep}(\Lambda)} I\left(P, V\right)$ for $\Lambda = 0.3, 0.5, 0.7$. For large values of $\mathbb{P}(X = 0)$ the capacity is 0 and as Λ increases this zero region grows. In addition, as Λ changes the capacity achieving input distribution changes.

bility a, for $S = 1$ it is a Z-channel with crossover b:

$$W(y|x, 1) = \begin{pmatrix} 1 & 0 \\ b & 1 - b \end{pmatrix} , \tag{3.65}$$

and for $S = 2$ it is an “S-channel” with crossover c:

$$W(y|x, 1) = \begin{pmatrix} 1 - c & c \\ 0 & 1 \end{pmatrix} . \tag{3.66}$$

The S-channel maps 1 to 1 with probability 1 and 0 to 1 with probability c. We will set $l(0) = 0$ and let $l(1)$ and $l(2)$ be arbitrary.

This approach to modeling a channel assumes that the overall channel is a mixture of channels with known characteristics. Figure 3.5 shows a plot of the minimum mutual information $\min_{V \in \mathcal{W}_{dep}(\Lambda)} I(P, V)$ versus P for this example. The plot shows that for different values of Λ the capacity achieving distribution can shift when the state can depend on the input.

3.5 Discussion

We saw in this section two simple strategies for creating randomized codes with different key sizes from other codes. For standard AVCs we can sample codes from a randomized code with large key size to obtain randomized codes with smaller key size. The error probability decays roughly inversely with the number of keys. For AVCs with nosy noise we could use list codes to construct a randomized code whose error decays inversely with the square root of the number of keys.

Being able to find randomized codes with small key size may be important in applications in which secure common randomness is a scarce resource. Implementing

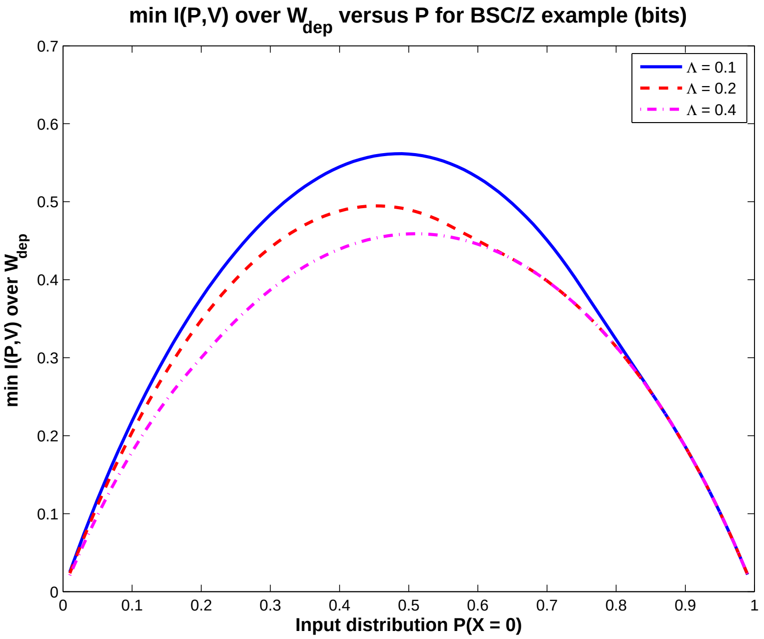


Figure 3.5: For the example in Section 3.4.2, the mutual information $\min_{V \in \mathcal{W}_{dep}(\Lambda)} I(P, V)$ for $\Lambda = 0.1, 0.2, 0.4$ with $a = 0.01$, $b = 0.1$, and $c = 0.15$, and $l(1) = 0.5$ and $l(2) = 0.6$.

randomized code constructions may involve a computational overhead that scales with the key size. Finally, by limiting the amount common randomness, one can use a secure feedback channel of negligible rate to enable randomized coding. In the next chapter we will investigate another benefit of limited feedback in the form of adapting the rate to the empirical channel.

Ich weiß nicht was ein Mensch ist,
Ich ken- ne nur sei- nen Preis.

– *Song Von Der Ware*, Hanns Eisler (lyrics by Bertolt Brecht)

Chapter 4

Limited feedback and rateless coding

In his 1972 paper on broadcast channels [40], Thomas Cover compared certain broadcasting problems to “giving a lecture to a group of disparate backgrounds and aptitudes.” Consider the following related scenario: at a certain progressive university, a professor embarked on a controversial new lecturing technique for her morning class. Each student’s alertness would wax and wane during the course of the lecture, due to such factors as the amount they had slept, their activities the previous night, or whether they had drunk any coffee. Since the professor could not tell how fast the students would learn, she decided that they should be the best judge of their own wakefulness. Each student was given an identical note sheet of fixed size. During the lecture, she attempted to teach the students a fixed amount of information. The students were to take notes as they understood the material; once their sheet was full, they could leave. The professor would look up at the class every minute to see if anyone was left, and the lecture would end only when every student had voluntarily left the room. The professor’s goal was to design her lecture to allow each student to spend just as much time in class as was necessary for him or her to learn the material.

Consider first the problem of lecturing to a single student for a fixed length of time. A simple information-theoretic approach might model the lecture as encoding a *message* into a sequence of *facts* (the codeword) which is then conveyed via a memoryless channel $W(y|x)$ to the student. This model does capture the facts that the student's attention varies over time and that he has some knowledge of his own wakefulness. We can instead use a channel $W(y|x, s)$ with state $s \in \mathcal{S}$ and *partial information about the state sequence* available to the decoder. There are now at least two different models we can pursue, depending on whether or not we assume the student's ignorance is independent of the facts being presented[1]. In one case, we can model the channel as a standard arbitrarily varying channel (AVC), and in the other as an AVC with nosy noise.

We can think of designing the professor's lecture as designing a *rateless code* for an arbitrarily varying channel with partial state information at the decoder. A single transmitter (the professor) wishes to communicate a common message (the information) to a group of receivers (the students) over channels with varying states (the complicated factors). The channels are unknown to the transmitter, but partially known to the decoder (the student's measure of their own wakefulness). The transmitter and receivers share a secret key (the note sheet) that is independent of the message. The transmitter encodes the message into a codeword (the lecture) such that that receivers can decode (leave the room) at a rate compatible with their state information.

Rateless codes are used to communicate over time varying channels when a low-rate feedback link can be used by the decoder to terminate the transmission. Figure 4.1 shows a diagram of such a communication channel. Rateless codes were first studied in the context of the erasure channel [107, 141] and later discrete memoryless

[1] In practice, this may be a difficult distinction to make!

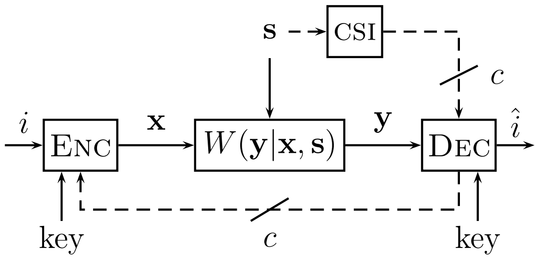


Figure 4.1: A rateless communication system. The encoder and decoder share a source of common randomness. Every c channel uses the decoder receives partial information about the channel state and can feed back a single bit.

compound channels [142, 54, 150]. Draper, Frey, and Kschischang [53] investigated rateless coding over AVCs under average error with full state information at the decoder. Coding for channels with full feedback was considered by Shayevitz and Feder [140] under an "individual sequence" assumption on the state. Our approach here is to assume only partial state information at the decoder, which results in rates lower than that of Draper et al. [53].

In this chapter we will construct partially derandomized rateless codes for AVC-like channel models. For the cost-constrained AVC under maximal error, we construct a rateless code based on thinning codewords from a codebook of large blocklength. Our coding construction may be used to partially derandomize a recently proposed construction by Eswaran, Sarwate, Sahai, and Gastpar [62]. In the case where the state sequence may depend on the transmitted codeword (nosy noise), we use a new model of partial CSI in which the decoder is given a set of channels in which the true empirical channel lies. By concatenating list-decodable codes, we find a new decoding strategy that experiences negligible loss when the channel estimates are good.

Our code constructions use a maximum blocklength n and nominal rate ρ under "worst-case" channel conditions but achieve rates higher than ρ under better channel conditions. This variation in rate is accomplished by decoding earlier, thereby shortening the effective blocklength. Our constructions rely on three parameters that are functions of the worst-case blocklength n : the number of messages $N(n)$, the key size $K(n)$ and the chunk size $c(n)$. The encoder transmits one of $N(n)$ *messages* using a randomized code construction of *key size* $K(n)$. The final parameter is the *chunk size* $c(n)$. After every $c(n)$ channel uses, the decoder is given an estimate of the channel and can feed back a single bit to terminate the transmission.

In our rateless code constructions we are interested in the case where the chunk size c is sublinear in n, the number of messages N is exponential in n, and the key size K is subexponential in n:

$$\frac{c(n)}{n} \to 0 \tag{4.1}$$

$$\frac{1}{n} \log N(n) \to \rho \tag{4.2}$$

$$\frac{1}{n} \log K(n) \to 0 \ . \tag{4.3}$$

We will choose the minimum rate ρ to be arbitrarily close to the randomized coding capacity for an AVC with cost constraint Λ. The coding strategy for a rateless code involves making a decision to terminate transmission based on the partial state information received at the decoder. This induces a decoding time $\mathbf{M}$ that is a function of the state sequence $\mathbf{s} \in \mathcal{S}^n(\Lambda)$ and the state information.

Our results will be phrased as follows : there exists a code such that for any $\mathbf{s} \in \mathcal{S}^n(\Lambda)$, the decoding rule at time $\mathbf{M}$ results in a small probability of error. Suppose we ignore the state information and feedback and just use one of the randomized codes from Theorems 13 or 14 in Chapter 3. For these codes $\mathbf{M} = n$ almost surely, so they always operate at the minimum rate. A regular block code for the AVC can therefore be seen as a limiting case of a rateless code. In this chapter we provide decoding rules and codes that use the side information to decode earlier. In the case where the side information is very accurate, we can bound the gap between the rate we achieve and the empirical mutual information of the channel.

4.1 AVCs with periodic channel state information and feedback

4.1.1 Channel model

We will model our time-varying channel by an arbitrarily varying channel $\mathcal{W} = \{W(y|x,s) : s \in \mathcal{S}\}$ with finite input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and constrained state sequence [45]. In point-to-point fixed blocklength channel coding problems, we assume a constraint Λ on the average state cost, so that

$$l(\mathbf{s}) \leq n\Lambda \quad \text{a.s. .} \tag{4.4}$$

Recall that $\mathcal{S}^n(\Lambda) = \{\mathbf{s} : l(\mathbf{s}) \leq n\Lambda\}$ is the set of sequences with average cost less than or equal to Λ.

In this chapter, we use the following notation: for an input distribution P and set of channels $\mathcal{V}$ we write

$$I\left(P, \mathcal{V}\right) = \min_{V \in \mathcal{V}} I\left(P, V\right) \ . \tag{4.5}$$

This will allow us to write the results in a more compact form.

4.1.2 Rateless codes

A rateless code is a variable-length coding strategy that uses periodic single-bit active feedback from the decoder to terminate the decoding. In our problem formulation, the decoder is periodically given side information consisting of an estimate of the channel. Our construction will use a parameter $c = c(n)$ called the chunk size. The decoder receives an estimate of the channel and can feed back a single bit at integer multiples of c. To be more precise, after the (mc)-th channel symbol is received, the following events occur:

1. the decoder receives an estimate of the channel state $(s_{(m-1)c+1}, \ldots, s_{mc})$;
2. the decoder makes a decision whether to decode;
3. the decoder feeds back a single bit to the encoder to signal whether or not it has decoded.

We denote the partial side information (channel estimate) given to the decoder after the m-th chunk by $\mathcal{V}_m$, which takes values in a set $\mathbf{V}(c)$. Let $\mathbf{y}_1^r$ denote $(y_1, y_2, \ldots, y_r)$ and define $\mathbf{y}^{(mc)} = (y_{(m-1)c+1}, \ldots, y_{mc})$, and similarly for $\mathbf{x}^{(mc)}$ and $\mathbf{s}^{(mc)}$.

A (c, N, K) randomized rateless code is set of maps $\{(\Phi_m, \tau_m, \Psi_m) : m = 1, 2, \ldots\}$:

$$\Phi_m : [N] \times [K] \times \{0,1\}^{m-1} \to \mathcal{X}^c \tag{4.6}$$

$$\tau_m : \mathcal{Y}^{mc} \times \mathbf{V}(c)^m \times [K] \to \{0,1\} \tag{4.7}$$

$$\Psi_m : \mathcal{Y}^{mc} \times \mathbf{V}(c)^m \times [K] \to [N] \ . \tag{4.8}$$

To encode chunk m, the encoding function Φ_m uses the message in $[N]$, key in $[K]$, and past feedback bits in $\{0,1\}^{m-1}$ to choose a vector of c channel inputs.

In the constructions given in this chapter, the decoder uses the function τ_m to decide whether to feed back a 1 to tell the encoder to terminate transmission or a 0 to continue. When $\tau_m(\cdot) = 1$, then the decoder uses the function Ψ_m to decode the message. Because the transmission terminates after the decoder feeds back a 1, the binary feedback signal $\{0,1\}^{m-1}$ is equal to all 0's at all times prior to terminating the scheme. In our constructions we can therefore disregard the input $\{0,1\}^{m-1}$ to the encoder Φ_m.

The decision function τ_m defines a random variable, called the *decoding time* $\mathbf{M}$ of the rateless code:

$$\mathbf{M} = \min\left\{m : \tau_m(\mathbf{y}_1^{mc}, \mathcal{V}_1^m, k) = 1\right\} \ . \tag{4.9}$$

Let $\mathcal{M} = \{M_*, M_*+1, \ldots, M^*\}$ be the smallest interval containing the support of $\mathbf{M}$. The set of possible rates for the rateless code are given by $\{(mc)^{-1} \log N : m \in \mathcal{M}\}$.

We can define decoding regions for the rateless code at a decoding time $\mathbf{M} = M$. Note that if $\mathbf{M} = M$ we have $\tau_M(Y_1^{Mc}, \mathcal{V}_1^M, k) = 1$. For message i, key k and side information vector $\mathcal{V}_1^M$ we can define a decoding region:

$$D_{i,k}(\mathcal{V}_1^M) = \left\{Y_1^{Mc} : \tau_M(Y_1^{Mc}, \mathcal{V}_1^M, k) = 1,\ \Psi_M(Y_1^{Mc}, \mathcal{V}_1^M, k) = i\right\} \ . \tag{4.10}$$

The *maximal* and *nosy noise error* for a (c, N, K) rateless code at decoding time $\mathbf{M} = M$ are, respectively,

$$\varepsilon(M, \mathbf{s}, \mathcal{V}_1^M) = \max_{i \in [N]} \frac{1}{K} \sum_{k=1}^{K} \left(1 - W^{Mc}\left(D_{i,k}(\mathcal{V}_1^M) \middle| \Phi_1^M(i,k), \mathbf{s}_1^{Mc}\right)\right) \tag{4.11}$$

$$\hat{\varepsilon}(M, J, \mathcal{V}_1^M) = \max_{i \in [N]} \frac{1}{K} \sum_{k=1}^{K} \left(1 - W^{Mc}\left(D_{i,k}(\mathcal{V}_1^M) \middle|\ \Phi_1^M(i,k), J_M(i, \Phi_1^M(i,k))\right)\right) \ . \tag{4.12}$$

Here $J = (J_1, \ldots, J_M)$ and $J_M : [N] \times \mathcal{X}^{Mc} \to \mathcal{S}^{Mc}$ is the jammer's strategy. Note that in these error definitions we do not take the maximum over all $\mathbf{s}$ or $\mathbf{J}$, because the rate and error at which we decode will depend on the realized state sequence.

This is in contrast to the equations (1.24) and (1.25).

4.1.3 Partial channel state information

In our rateless coding model, the decoder is given some partial information about the channel after each chunk of c channel uses. We model this side information as a subset of channels $\mathcal{V}_m$ that contains the true average channel in the m-th chunk. The channel state information $\mathcal{V}_m$ takes values in a set $\mathbf{V}(c)$. Under maximal error we let $\mathbf{V}(c)$ be a collection of subsets of $\mathcal{W}_{std}(\Lambda) \cap \mathcal{P}_c(\mathcal{Y}|\mathcal{X})$, and under nosy noise let it be a collection of subsets of $\mathcal{W}_{dep}(P, \Lambda) \cap \mathcal{P}_c(\mathcal{Y}|\mathcal{X})$ for a fixed input distribution P.

Suppose that during the m-th chunk of channel uses $\{(m-1)c+1, \ldots mc\}$ the channel inputs were $\mathbf{x}^{(mc)}$ and the state was $\mathbf{s}^{(mc)}$. Under the maximal error criterion, we define the average channel under $\mathbf{s}$ during the m-th chunk by

$$V_m(y|x) = \frac{1}{c} \sum_{t=(m-1)c+1}^{mc} W(y|x, s_t) \ . \tag{4.13}$$

Under the nosy noise criterion we define the average channel under under $\mathbf{x}$ and $\mathbf{s}$ by

$$V_m(y|x) = \frac{1}{N(x|\mathbf{x}^{(mc)})} \sum_{t=(m-1)c+1}^{mc} W(y|x_t, s_t) \mathbf{1}(x_t = x) \ . \tag{4.14}$$

The model we will use for our channel state information is that the decoder receives a subset of channels $\mathcal{V}_m \in \mathbf{V}(c)$, where $V_m(y|x) \in \mathcal{V}_m$. Here $\mathbf{V}(c) = \mathcal{W}_{std}(\Lambda) \cap \mathcal{P}_c(\mathcal{Y}|\mathcal{X})$ for maximal error and $\mathbf{V}(c) = \mathcal{W}_{dep}(\Lambda) \cap \mathcal{P}_c(\mathcal{Y}|\mathcal{X})$ for nosy noise. In order for our results to hold we need a polynomial upper bound on the size of $\mathbf{V}(c)$. We will assume

$$|\mathbf{V}(c)| \leq c^v \ , \tag{4.15}$$

for some $v < \infty$.

In Section 4.2 we provide a rateless code construction for AVCs under maximal error in the special case where the side information $\mathcal{V}_m$ at the decoder is an estimate of the state cost during the m-th chunk. Let

$$\lambda_m = \frac{1}{c} \sum_{t=(m-1)c+1}^{mc} l(s_t) \ . \tag{4.16}$$

We will model the CSI after the m-th chunk as a measurement $\hat{\lambda}_m \in \mathbb{R}^+$ with the property that

$$\lambda_m \leq \hat{\lambda}_m \ . \tag{4.17}$$

That is, the decoder obtains an estimate of the average state cost in the chunk. The number of possible values for λ_m is at most $(c+1)^{|\mathcal{S}|}$, which is an upper bound on the number of types on $\mathcal{S}$ with denominator c:

$$\lambda_m \in \ell(c) = \{\lambda : \exists \mathbf{s} \in \mathcal{S}^c \ \ s.t. \ \ l(\mathbf{s}) = c\lambda\} \ . \tag{4.18}$$

We can then set

$$\mathbf{V}(\,c) = \{\mathcal{W}_{std}(\lambda) \cap \mathcal{P}_c(\mathcal{Y}|\mathcal{X}) : \lambda \in \ell(c)\} \ \ . \tag{4.19}$$

Note that $\mathbf{V}(c)$ satisfies the polynomial cardinality bound. Now we can set

$$\mathcal{V}_m = \max_{\lambda \in \ell(c) : \lambda \leq \hat{\lambda}_m} \mathcal{W}_{std}(\lambda) \cap \mathcal{P}_c(\mathcal{Y}|\mathcal{X}) \ . \tag{4.20}$$

Therefore cost information about the empirical channel fits within the general framework of partial side information.

In Section 4.3 we provide a rateless code construction for AVCs under the nosy noise error criterion. For those codes, we will not assume any particular structure on the side information beyond the polynomial cardinality bound on $\mathbf{V}(c)$ given in (4.15). That is, we will model the side information after chunk m as a subset $\mathcal{V}_m \subset \mathcal{W}_{dep}$ with the property that

$$\frac{1}{c} \sum_{t \in \{(m-1)c+1, \ldots, mc\} : x_t = x} W(y|x, s_t) \in \mathcal{V}_m \ . \tag{4.21}$$

4.2 Rateless coding for standard AVCs

In this section we construct a rateless coding scheme for the standard AVC model with cost constraints that uses cost estimates at the decoder to opportunistically decode when it has received "enough" channel symbols.

4.2.1 The result

For the standard AVC model, we can use the construction of Csiszár and Narayan [45] as a basis for constructing a randomized rateless code with unbounded key size. By using the elimination technique to partially derandomize this construction we can reduce the key size and establish a tradeoff between the randomization and error.

We will define some additional notation. Let

$$\Lambda_M = \frac{1}{M} \sum_{m=1}^{M} \lambda_m \tag{4.22}$$

$$\hat{\Lambda}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\lambda}_m \tag{4.23}$$

be the true and estimated cost for the state sequence $\mathbf{s}_1^{Mc}$.

Our main result is the following theorem, which provides a rateless code construction. The proof of this theorem is given in Section 4.2.4.

Theorem 16. *Let $\mathcal{W}$ be an AVC. For any $\epsilon > 0$ and input type $P \in \mathcal{P}(\mathcal{X})$ with $\min_x P(x) > 0$, there exists an n sufficiently large and an $(c(n), N(n), K(n))$ rateless code with $M^* = n/c$ for which, given*

$$l(\mathbf{s}_1^{mc}) \leq mc \cdot \hat{\Lambda}_m \qquad \forall m \ , \tag{4.24}$$

the decoding time $\mathbf{M}$ is given by

$$\mathbf{M} = \min \left\{ M : \frac{1}{Mc(n)} \log N(n) \leq I\left(P, \mathcal{W}_{std}(\hat{\Lambda}_M)\right) - 2\frac{M^*(n)}{M}\epsilon \right\} \ , \tag{4.25}$$

and for all $(\mathbf{s}, \hat{\lambda}_1^{\mathbf{M}})$, the maximal error $\varepsilon(\mathbf{M}, \mathbf{s}, \mathcal{V}_1^{\mathbf{M}})$ of this code satisfies

$$\varepsilon(\mathbf{M}, \mathbf{s}, \mathcal{V}_1^{\mathbf{M}}) = O\left(\frac{\mathbf{M}c(n)}{K(n)}\right) \ . \tag{4.26}$$

Example 4.1 – Bit-flipping (mod-two adder)

Consider the mod-two additive AVC described in Example 1.4 on page 21. For this example, we can think of the partial side information $\hat{\lambda}_m$ as an estimate of the empirical Hamming weight of the state sequence $\mathbf{s}^{(mc)}$ such that

$$\hat{\lambda}_m \geq \frac{1}{c} \sum_{t=(m-1)c+1}^{mc} l(s_t) \ . \tag{4.27}$$

The receiver tracks the empirical weight of the state sequence to compute an empirical crossover probability $\hat{\Lambda}_M$. Theorem 16 says there is a rateless code that can decode as soon as the estimated empirical mutual information $Mc(1 - h_b(\hat{\Lambda}_M))$ exceeds the size of the message ($\log N$ bits).

4.2.2 Coding strategy

Our scheme uses a fixed maximum blocklength n and we will express other parameters as functions of n. It can be described in three steps:

Algorithm I : Rateless coding for standard AVCs

1. The encoder and decoder choose a key $k \in [K(n)]$ to use for their transmission using common randomness. The encoder chooses a message $i \in [N(n)]$ to transmit and maps it into a codeword $\mathbf{x}(i, k) \in \mathcal{X}^n$.

2. If $\tau_{m-1}(\mathbf{y}_1^{(m-1)c}, \hat{\lambda}_1^{m-1}, k) = 0$, the encoder transmits $\mathbf{x}^{(mc)}(i, k)$ in channel uses $(m-1)c + 1, (m-1)c + 2, \ldots, mc$.

3. The decoder receives channel outputs $\mathbf{y}^{(mc)}$ and an estimate $\hat{\lambda}_m$ of the state cost in the m-th chunk such that

$$\lambda_m = \frac{1}{c} \sum_{t=(m-1)c+1}^{mc} l(s_t) \leq \hat{\lambda}_m \ . \tag{4.28}$$

   Define the decision function

$$\tau_m(\mathbf{y}_1^{mc}, \hat{\lambda}_1^m, k) = \mathbf{1}\left(\frac{\log N}{mc} < I\left(P, \mathcal{W}_{std}(\hat{\Lambda}_m)\right) - \delta(M)\right) \ . \tag{4.29}$$

   If $\tau_m(\cdot) = 1$ then the decoder attempts to decode the received sequence, sets $\hat{i} = \Psi_m(\mathbf{y}_1^{mc}, k)$, and feeds back a 1 to terminate transmission. Otherwise, the decoder feeds back a 0 and we return to step 2) to send chunk $m + 1$.

Our code relies on the existence of a set of codewords $\{\mathbf{x}(i, k)\}$ which, when truncated to blocklength mc, form a good randomized code for an AVC satisfying a given cost constraint. The key to our construction is that the condition checked by the decision function (4.29) is sufficient to guarantee that the decoding error will be small.

4.2.3 Codebook construction

Our codebook will consist of codewords drawn uniformly from the set

$$\mathcal{B}(n,c) = (\mathcal{T}_c(P))^{n/c} = \underbrace{\mathcal{T}_c(P) \times \mathcal{T}_c(P) \times \cdots \mathcal{T}_c(P)}_{n/c \text{ times}} . \tag{4.30}$$

That is, the codewords are formed by concatenating constant-composition chunks of length c. For a fixed number $N(n)$ of messages to be transmitted, the minimum rate supportable by this codebook is $\rho = n^{-1} \log N$, so $M^* = n/c$ is the largest element of the set of decoding times $\mathcal{M}$. The maximum rate is given by the randomized coding capacity with cost constraint 0 on the state. We choose M_* such that

$$\frac{\rho n}{M_* c} = C_r(0) - \delta(M_*) , \tag{4.31}$$

for some small constant $\delta(M_*)$.

To make our results easier to state, for n, M, c, ρ, and δ let us define $\tilde{\Lambda}_M$ to satisfy

$$\rho_M \overset{\Delta}{=} \frac{n}{Mc}\rho = I\left(P, \mathcal{W}_{std}(\tilde{\Lambda}_M)\right) - \delta(M) . \tag{4.32}$$

That is, $\tilde{\Lambda}_M$ is the cost constraint for which the empirical rate is equal to the empirical mutual information on an AVC with input type P.

To summarize, over channel uses 1 to Mc the true state cost is Λ_M, the channel estimate at the decoder is $\hat{\Lambda}_M$, and the codebook decoding rule is designed for an AVC with cost constraint $\tilde{\Lambda}_M$. The decision rule checks to see if

$$I\left(P, \mathcal{W}_{std}(\tilde{\Lambda}_M)\right) < I\left(P, \mathcal{W}_{std}(\hat{\Lambda}_M)\right) , \tag{4.33}$$

which implies that $\tilde{\Lambda}_M \geq \hat{\Lambda}_M$. By assumption,

$$\hat{\Lambda}_M \geq \Lambda_M \ , \tag{4.34}$$

so $I\left(P, \mathcal{W}_{std}(\hat{\Lambda}_M)\right) \leq I\left(P, \mathcal{W}_{std}(\Lambda_M)\right)$. Therefore if $\tau_M(\mathbf{y}_1^{Mc}, \hat{\lambda}_1^M, k) = 1$ then we also have

$$\frac{\log N}{Mc} < I\left(P, \mathcal{W}_{std}(\Lambda_M)\right) - \delta(M) \ . \tag{4.35}$$

That is, we know that if the decoding time $\mathbf{M}$ is equal to M,

$$\tilde{\Lambda}_M \geq \hat{\Lambda}_M \geq \Lambda_M \ . \tag{4.36}$$

Our strategy has two sources of loss – one from the inaccuracy in the channel estimates that makes a gap between the true cost Λ_M and the estimates $\hat{\Lambda}_M$, and one from the precision of the decoding rates $\{\rho_m\}$ that generates a gap between the codebook's designed cost $\tilde{\Lambda}_M$ and the estimate $\hat{\Lambda}_M$ used in the decision rule.

Lemma 14 (Fully randomized rateless codebook)**.** *Let $\mathcal{W}$ be an AVC with cost function $l(\cdot)$ and let $\tilde{\Lambda}_M$ be given by (4.32). For any $\epsilon > 0$ and input distribution $P \in \mathcal{P}(\mathcal{X})$ there exists a blocklength n sufficiently large and $c(n)$ with $c^{-1} \log n \to 0$ such that the randomized codebook $\{\mathbf{X}(i) : i \in [N]\}$ of size $\exp(n\rho)$ uniformly distributed on $\mathcal{B}(n, c)$ has the following property: for any $M \in \mathcal{M}$, this codebook truncated to blocklength $n = Mc$ is the codebook of an $(Mc, N(n))$ randomized code of rate*

$$\rho_M = \frac{n}{Mc}\rho \tag{4.37}$$

and maximum probability of error

$$\delta_M \leq \exp\left(-Mc\left(E_r\left(\rho_M + 2\frac{M^*}{M}\epsilon, P, \tilde{\Lambda}_M\right) - 2\epsilon\right)\right) \ . \tag{4.38}$$

Proof. Fix $\epsilon > 0$. We will prove that for each $M \in \mathcal{M}$ there exists a randomized codebook $\mathbf{C}_M$ of blocklength Mc with the specified error for the AVC with cost constraint $\tilde{\Lambda}_M$. The distribution of the codebook $\mathbf{C}_M$ will be the same as the distribution of the codebook $\mathbf{C}_{M^*}$ of blocklength M^*c truncated to blocklength c.

Standard randomized codebook. Fix M and let $\mathbf{A}_M$ be a randomized codebook of A codewords drawn uniformly from the constant composition set $\mathcal{T}_{Mc}(P)$. From Hughes and Thomas [88, Theorem 1] we have the following error bound on message i with $R_M = (Mc)^{-1} \log A$ for an AVC with cost constraint $\tilde{\Lambda}_M$:

$$\delta_M(\mathbf{A}_M, i) \leq \exp\left(-Mc\left(E_r(R_M + \epsilon, P, \tilde{\Lambda}_M) - \epsilon\right)\right) \ , \qquad \stackrel{\Delta}{=} \zeta_M \tag{4.39}$$

Therefore we have the same bound on the average error

$$\frac{1}{A}\sum_{i=1}^{A} \delta_M(\mathbf{A}_M, i) \leq \zeta_M \ . \tag{4.40}$$

Expurgation. Let $\mathbf{B}_M$ be a random variable formed by expurgating all codewords not in the set $(\mathcal{T}_c(P))^M$. That is, we keep only those codewords which are piecewise constant composition with composition P. We write $\mathbf{B}_M = \mathbf{A}_M \cap (\mathcal{T}_c(P))^M$. Note that a realization of $\mathbf{B}_M$ has a variable number of codewords. We declare an *encoding error* if the number of codewords in $\mathbf{B}_M$ is smaller than B for some number

B. We use a combinatorial bound from Lemma 24 in Appendix B:

$$\begin{aligned}\frac{|\mathcal{T}_c(P)|^M}{|\mathcal{T}_{Mc}(P)|} &\geq \exp(-M\log(Mc)\eta(P)) \\ &\triangleq \gamma_M \ , \end{aligned} \tag{4.41}$$

where $\eta(P) < \infty$ is a positive constant.

Since $\mathbf{A}_M$ is formed by iid draws from $\mathcal{T}_{Mc}(P)$, the codebook size $|\mathbf{B}_M|$ is the sum of A Bernoulli random variables with parameter greater than γ_M. Hence we can use Sanov's theorem [41]:

$$\mathbb{P}\left(|\mathbf{B}_M| \leq B\right) \leq (A+1)^2 \exp\left(-A \cdot D\left(B/A \ \| \ \gamma_M\right)\right) \ . \tag{4.42}$$

We can bound the exponent by using the inequality $-(1-a)\log(1-a) \leq 2a$ for small a and discarding the small positive term $-(1-B/A)\log\gamma_M$:

$$\begin{aligned} A \cdot D\left(B/A \ \| \ \gamma_M\right) &= B \log\frac{B/A}{\gamma_M} + A(1-B/A)\log\frac{1-B/A}{1-\gamma_M} && (4.43) \\ &\geq B\log\frac{B/A}{\gamma_M} - 2B \ . && (4.44)\end{aligned}$$

If we let $B/A = \beta_M\gamma_M$ then

$$\mathbb{P}\left(|\mathbf{B}_M| \leq B\right) \leq \exp\left(-A\gamma_M\beta_M(\log\beta_M - 2) + 2\log(A+1)\right) \ . \tag{4.45}$$

Since $A = O(\exp(Mc))$ the probability of encoder error is much smaller than the decoding error bound ζ_M.

The encoder using $\mathbf{B}_M$ now operates as follows : it draws a realization of a codebook and declares an error if the realization contains fewer than B codewords. If there is no encoding error it transmits the i-th codeword in the codebook for message

$i \in [B]$. Note that this construction uses only B codewords for each codebook. The average error on the fraction $B/A = \beta_M \gamma_M$ of preserved codewords can be at most A/B times the original average error:

$$\frac{1}{B} \sum_{i=1}^{B} \delta_M(\mathbf{B}_M, i) \leq \frac{A}{B} \zeta_M \tag{4.46}$$

$$= \frac{\zeta_M}{\beta_M \gamma_M} \ . \tag{4.47}$$

Permutation. We now form our random codebook $\mathbf{C}_M$ by taking the codebook induced by encoder using $\mathbf{B}_M$ and permuting the message index. The encoder using $\mathbf{C}_M$ takes a message i, randomly chosen permutation π on [B], and a codebook $\mathcal{B}$ from $\mathbf{B}_M$ and outputs the codeword $\pi(i)$ from $\mathcal{B}$. The maximal error for a message i in this codebook is given by

$$\delta_M(\mathbf{C}_M, i) = \frac{1}{B!} \sum_{\pi} \delta_M(\mathbf{B}_M, \pi(i)) \tag{4.48}$$

$$= \frac{1}{B} \sum_{i=1}^{B} \delta_M(\mathbf{B}_M, i) \tag{4.49}$$

$$\leq \frac{\zeta_M}{\beta_M \gamma_M} \tag{4.50}$$

$$= \beta_M^{-1} \exp\left(-Mc \left(E_r(R_M + \epsilon, P, \tilde{\Lambda}_M) - \epsilon - \frac{\log(Mc)}{c} \eta(P) \right) \right) \ . \tag{4.51}$$

For each $M \in \mathcal{M}$ we can construct a randomized codebook $\mathbf{C}_M$ as described above.

Nesting. Now consider the codebook $\mathbf{C}_{M^*}$ of blocklength $n = M^* c$ and set the size of the codebook to $\exp(n\rho)$. We can write the rate of our original codebook $\mathbf{A}_{M^*}$ as

$$R_{M^*} = \rho - \frac{1}{M^* c} \log(\beta_{M^*} \gamma_{M^*}) \ . \tag{4.52}$$

Truncating $\mathbf{C}_{M^*}$ to blocklength Mc for $M \in \mathcal{M}$ gives a code of rate

$$R_M = \frac{n}{Mc}\rho - \frac{1}{Mc}\log(\beta_{M^*}\gamma_{M^*}) \ . \tag{4.53}$$

Therefore if we truncate $\mathbf{C}_{M^*}$ to blocklength Mc, the resulting randomized code is identically distributed to $\mathbf{C}_M$ with $\beta_M = \beta_{M^*}\gamma_M/\gamma_{M^*}$. For n and c sufficiently large we obtain the following inequality from (4.41):

$$\frac{\log n}{c}\eta(P) - \frac{1}{Mc}\log\beta_{M^*} < \epsilon \ . \tag{4.54}$$

We can write the error as

$$\delta_M(\mathbf{C}_{M^*}, i) \leq \exp\left(-Mc\left(E_r(\rho_M + \delta(M), P, \tilde{\Lambda}_M) - 2\epsilon\right)\right) \ , \tag{4.55}$$

where

$$\delta(M) = \frac{M^*}{M}2\epsilon \ . \tag{4.56}$$

The last thing to do is choose $\tilde{\Lambda}_M$ to make the exponent positive. We need

$$\rho_M + \delta(M) < I\left(P, \mathcal{W}_{std}(\tilde{\Lambda}_M)\right) \ . \tag{4.57}$$

But this is clear from (4.32) and (4.33). □

With the previous lemma in hand, we can apply Lemma 12 to derandomize the randomized code.

Lemma 15 (Derandomized rateless codebook). *Let $\mathcal{W}$ be an AVC with cost function $l(\cdot)$. For any $\epsilon > 0$ and input type $P \in \mathcal{P}(\mathcal{X})$, there is an n sufficiently large, $c(n)$ and an $(n, N(n), K(n))$ randomized code for maximal error such that the codewords* *truncated to blocklength Mc is an (Mc, N, K) randomized code for maximal error with rate and error at blocklength Mc given by*

$$\rho_M = I(P, \mathcal{W}_{std}(\tilde{\Lambda}_M)) - \frac{M^*}{M} 2\epsilon \tag{4.58}$$

$$\delta_M = O\left(\frac{Mc}{K(n)}\right) , \tag{4.59}$$

where $c(n)^{-1} \log n \to 0$.

Proof. Let $\mathbf{C}$ be the codebook-valued random variable that is the randomized code from Lemma 14. For each M, let $\mathbf{C}_M$ be the the codebook truncated to blocklength Mc. We know that $\mathbf{C}_M$ forms a good randomized codebook with error (4.38) for the AVC with cost constraint $\tilde{\Lambda}_M$ in (4.58). Let us write ν for the upper bound in (4.38).

We can now draw K codebooks sampled uniformly from $\mathbf{C}_{M^*}$. Since $\mathbf{C}_{M^*}$ truncated to blocklength Mc is $\mathbf{C}_M$, this sampling induces a sampling on $\mathbf{C}_M$ for each M. For any M, Lemma 12 shows that the probability that the randomized codebook formed from the samples has error larger than δ_M can be driven to 0 as long as

$$\frac{K(n)}{Mc}\left(-Mc\delta_M\left(E_r\left(\rho_M + 2\frac{M^*}{M}\epsilon, P, \tilde{\Lambda}_M\right) - 2\epsilon\right) - h_b(\delta_M)\log 2\right) \\ > \left(\left(\rho_M + 2\frac{M^*}{M}\epsilon\right)\log 2 + \log|\mathcal{S}|\right) . \tag{4.60}$$

As in the proof of Theorem 13 we can choose δ_M to satisfy (4.59). □

4.2.4 Proof of Theorem 16

Proof. Fix $\epsilon > 0$. Then Lemma 15 gives a codebook with the desired error as long as the $\mathbf{M}$ given by the stopping rule guarantees that the state sequence will have cost no more than $\tilde{\Lambda}_M$. If $\hat{\lambda}_m$ satisfies (4.28) then we can see from the stopping rule that

$$l\left(\mathbf{s}_1^{Mc}\right) = Mc\Lambda_M \leq Mc\hat{\Lambda}_M \leq Mc\tilde{\Lambda}_M \ . \tag{4.61}$$

Therefore Lemma 15 gives the desired error bound at the decoding time $\mathbf{M}$. □

4.2.5 Loss from channel estimation

We now address the efficiency of our scheme and show that the loss in rate can be made arbitrarily small if the channel estimates are within an arbitrarily small gap η of the true channel cost.

Corollary 3. *Let $\epsilon > 0$ be given and consider the $(c(n), N(n), K(n))$ rateless code constructed in Lemma 15. Let $\eta > 0$ and suppose that for all m,*

$$\hat{\lambda}_m - \lambda_m \leq \eta \ . \tag{4.62}$$

Then

$$\left| \frac{1}{\mathbf{M}c} \log N - I\left(P, \mathcal{W}_{std}(\Lambda_{\mathbf{M}})\right) \right| \leq 2\frac{M^*}{M}\epsilon + f(\eta) \ . \tag{4.63}$$

where $f(\eta) = O(\eta \log \eta^{-1})$.

Proof. Note that the condition (4.62) implies that $\hat{\Lambda}_{\mathbf{M}} - \Lambda_{\mathbf{M}} \leq \eta$. Let $\hat{Q}(s) \in \mathcal{Q}(\hat{\Lambda}_{\mathbf{M}})$ be a distribution on $\mathcal{S}$ that minimizes $I\left(P, \mathcal{W}_{std}(\hat{\Lambda}_{\mathbf{M}})\right)$. We will find a distribution $Q \in \mathcal{Q}(\Lambda_{\mathbf{M}})$ such that $d_{\max}\left(\hat{Q}, Q\right)$ is small.

Let $s_1 = \operatorname{argmin} l(s)$ and $s_2 = \operatorname{argmin}_{l(s) \neq l(s_1)} l(s)$. Given a gap in cost η and a distribution $\hat{Q}$, we can construct a distribution Q by setting $Q(s_1) = \hat{Q}(s_1) + \mu$ and removing mass μ from elements s with higher cost than s_1. The largest μ that this can incur can be bounded:

$$d_{\max}\left(\hat{Q}, Q\right) \leq \mu = \frac{\eta}{l(s_2) - l(s_1)} \; . \tag{4.64}$$

Let $\hat{V} \in \mathcal{W}_{std}(\hat{\Lambda}_{\mathbf{M}})$ and $V \in \mathcal{W}_{std}(\Lambda_{\mathbf{M}})$ be the channels corresponding to $\hat{Q}$ and Q. Given the bound on $d_{\max}\left(\hat{Q}, Q\right)$, we have

$$\begin{aligned} d_{\max}\left(\hat{V}, V\right) &= \max_{x,y} \left| \sum_s W(y|x,s)\hat{Q}(s) - \sum_s W(y|x,s)Q(s) \right| && (4.65) \\ &\leq \max_{x,y} \sum_s W(y|x,s) \cdot |\hat{Q}(s) - Q(s)| && (4.66) \\ &\leq |\mathcal{S}| d_{\max}\left(\hat{Q}, Q\right) \; . && (4.67) \end{aligned}$$

Let $g(\eta) = |\mathcal{S}|\eta/(l(s_2) - l(s_1))$ Now, using a bound from Lemma 23 in Appendix B, we can bound the mutual information gap:

$$\begin{aligned} \left| I\left(P, V\right) - I\left(P, \hat{V}\right) \right| &= I\left(P, V\right) - I\left(P, \hat{V}\right) \\ &\leq 2(|\mathcal{Y}| - 1)h_b(g(\eta)) + 2(|\mathcal{Y}| - 1)\log(|\mathcal{Y} - 1))g(\eta) && (4.68) \\ &\triangleq f(\eta) \; . && (4.69) \end{aligned}$$

Now from the definition of the decoding time in (4.25) we have:

$$\begin{aligned} \left| \frac{1}{\mathbf{M}c} \log N - I\left(P, \mathcal{W}_{std}(\Lambda_{\mathbf{M}})\right) \right| &\leq I\left(P, \mathcal{W}_{std}(\Lambda_{\mathbf{M}})\right) - I\left(P, \mathcal{W}_{std}(\hat{\Lambda}_{\mathbf{M}})\right) + 2\frac{M^*}{M}\epsilon \\ &\leq I\left(P, V\right) - I\left(P, \hat{V}\right) + 2\frac{M^*}{M}\epsilon \\ &\leq 2\frac{M^*}{M}\epsilon + f(\eta) \; . && (4.70) \end{aligned}$$

□

4.3 Rateless coding for AVCs with nosy noise

We now turn to a rateless coding construction for AVCs where the state sequence can depend on the transmitted codeword but we have partial side information about the empirical channel at the decoder. This construction is based on the list-decodable codes of Theorem 14.

4.3.1 The result

In the case of nosy noise, the state can depend on the transmitted codeword. We can also use a codebook of concatenated constant-composition components to make a rateless code for this model.

Theorem 17. *For any $\rho > 0$, $\epsilon > 0$, $\delta > 0$, and input type $P \in \mathcal{P}(\mathcal{X})$ with $\min_x P(x) > 0$, for n sufficiently large, there is an $(c(n), N(n), K(n))$ rateless code with $N = \lceil \exp(n\rho)/K(n) \rceil$ for which, given that the channel*

$$V_m(y|x) = \frac{1}{N(x|\mathbf{x}^{(mc)})} \sum_{t=(m-1)c+1}^{mc} W(y|x_t, s_t) \mathbf{1}(x_t = x) \qquad (4.71)$$

is in $\mathcal{V}_m$ for all m, the decoding time $\mathbf{M}$ is given by

$$\mathbf{M} = \min \left\{ M : \frac{1}{Mc} \log N \le \frac{1}{M} \sum_{m=1}^{M} I\left(P, \mathcal{V}_m\right) - \epsilon \right\} . \qquad (4.72)$$

and for all such $\mathcal{V}_1^{\mathbf{M}}$ and jamming strategies J, the error for this code satisfies

$$\hat{\varepsilon}(\mathbf{M}, J, \mathcal{V}_1^{\mathbf{M}}) \le \mathbf{M} \exp(-cE(\epsilon)) + \frac{24 n\rho \log |\mathcal{Y}|}{\epsilon \sqrt{K} \log K} , \qquad (4.73)$$

where $E(\epsilon) > 0$.

The proof of this theorem is given in Section 4.3.4. The theorem says that there exists a rateless code which can be decoded as soon as the empirical mutual information $c\sum_{m=1}^{M} I\left(P, \mathcal{V}_m\right)$ is enough to sustain the $\log N$ bits for the message. This threshold is sufficient to guarantee decoding with small probability of error because the codebook is designed for an AVC with nosy noise.

4.3.2 Coding strategy

In order to make our decoder opportunistic, we explicitly use information about the output sequence $\mathbf{y}$ at the decoder together with the side information $\mathcal{V}_m$. For $\delta > 0$ and distribution $P \in \mathcal{P}(\mathcal{X})$, given the m-th chunk of channel outputs $\mathbf{y}^{(mc)}$ and the side information set $\mathcal{V}_m$ let

$$\mathcal{V}_m(\mathbf{y}^{(mc)}, \epsilon) = \left\{ V \in \mathcal{V}_m : d_{\max}\left(T_{\mathbf{y}^{(mc)}}, \sum_x P(x)V(y|x) \right) < \delta \right\} . \tag{4.74}$$

Although $\mathcal{V}_m(\mathbf{y}^{(mc)}, \delta)$ depends on P, in our construction P is fixed so we do not make this dependence explicit. The coding algorithm depends on two additional constants δ and ϵ.

Algorithm II : Rateless coding for nosy "noise"

1. Using common randomness, the encoder and decoder choose a key $k \in [K]$ to use. The encoder chooses a message $i \in [N]$ to transmit and maps it into a codeword $\mathbf{x}(i, k) \in \mathcal{X}^n$.

2. If $\tau_{m-1}(\cdot) = 0$, the encoder transmits $\mathbf{x}^{(mc)}$ in channel uses $(m-1)c+1, (m-1)c+2, \ldots, mc$.

3. The decoder receives channel outputs $\mathbf{y}^{(mc)}$ and the channel state information

set $\mathcal{V}_m$ and calculates the set of possible channels $\mathcal{V}_m(\mathbf{y}^{(mc)}, \delta)$. If

$$\frac{\log N}{mc} < \frac{1}{m} \sum_{i=1}^{m} I\left(P, \mathcal{V}_m(\mathbf{y}^{(mc)}, \delta)\right) - mc\epsilon \tag{4.75}$$

for some i, then the decoder sets $\tau_m(\cdot) = 1$ and attempts to decode. Otherwise, it feeds back a 0 and we return to step 2) for chunk $m+1$.

For each chunk, the decoder looks over all channels in the side information set consistent with what it received, and takes the worst-case mutual information. The average of these worst-case mutual informations is our estimate of the empirical mutual information of the channel.

4.3.3 Codebook construction

The codebook we use is again sampled from $\mathcal{B}(n, c)$ given in (4.30):

$$\mathcal{B}(n,c) = (\mathcal{T}_c(P))^{n/c} = \underbrace{\mathcal{T}_c(P) \times \mathcal{T}_c(P) \times \cdots \mathcal{T}_c(P)}_{n/c \text{ times}} . \tag{4.76}$$

For the AVC with nosy noise we use a two-step decoding process. Given a decoding time, the decoder list-decodes the received sequence using the partial side information. It then uses the key to disambiguate the list using the same message authentication scheme as in Theorem 14.

Lemma 16 (Exponential list decoding with variable side information)**.** *Let $\mathcal{W}$ be an AVC. For any $\delta > 0$ and $\xi > 0$, $P \in \mathcal{P}(\mathcal{X})$ with $\min_x P(x) > 0$ there is a c sufficiently large such that for any $\mathcal{V} \in \mathbf{V}(c)$ the set $\mathcal{T}_c(P)$ is an $(c, N, L(\mathcal{V}))$ list-decodable code* *for the AVC $\mathcal{W}$ under maximal error with*

$$N = |\mathcal{T}_c(P)| \geq \exp\left(c(H(X) - \xi)\right) \tag{4.77}$$

$$L(\mathcal{V}) \leq \exp\left(c\left(\max_{V \in \mathcal{V}(\mathbf{y}_1^c, \delta)} H(X|Y) + \xi\right)\right) \ , \tag{4.78}$$

and error

$$\varepsilon_L \leq \exp(-c \cdot E_1(\xi)) \ , \tag{4.79}$$

where $H(X)$ is calculated with respect to the distribution $P(x)$ and for $V \in \mathcal{V}(\mathbf{y}_1^c, \delta)$ the conditional entropy $H(X|Y)$ is with respect to the distribution $P(x)V(y|x)$, and $E_1(\xi) > 0$.

Proof. Fix $\xi > 0$ and $\delta > 0$. For an input distribution $P(x)$ and channel $V(y|x)$, let $P'(y)$ be the marginal distribution on $\mathcal{Y}$ and $V'(x|y)$ be the channel such that $P(x)V(y|x) = P'(y)V'(x|y)$. Our decoder will output the set

$$\mathcal{L}(\mathbf{y}_1^c) = \bigcup_{V \in \mathcal{V}(\mathbf{y}_1^c, \delta)} T_{V'}^{(|\mathcal{X}|+1)\xi}(\mathbf{y}) \ . \tag{4.80}$$

The size of this set is, by a union bound, upper bounded by (4.78). The proof of Lemma 1 on page 53 shows that the probability that either $\mathbf{x} \notin \mathcal{L}(\mathbf{y}_1^c)$ or

$$\mathbf{y} \notin \bigcup_{V \in \mathcal{V}(\mathbf{y}_1^c, \delta)} T_V^{\xi}(\mathbf{x}) \tag{4.81}$$

is upper bounded by

$$\varepsilon_L(\mathcal{W}) \leq \exp(-c \cdot E_L(\mathcal{W}, \xi)) \ . \tag{4.82}$$

For c sufficiently large, the size of this list can be bounded by (4.78), and the error probability is still bounded by

$$\varepsilon_L(\mathcal{W}) \leq \exp(-c \cdot E_L(\mathcal{W}, \xi)) \ . \tag{4.83}$$

Thus, with probability exponential in c, this set will contain the transmitted $\mathbf{x} \in T_P^c$. Taking a union bound over the $|\mathbf{V}(c)| = c^v$ possible values of the side information $\mathcal{V}_c$ shows that

$$\varepsilon_L \leq \exp(-c \cdot E_L(\mathcal{W}, \xi) - v \log c) \ , \tag{4.84}$$

which gives the exponent $E_1(\xi)$. □

With the previous lemma as a basic building block, we can create nested list-decodable codes where c is chosen to be large enough to satisfy the conditions of Lemma 16. The codebooks we will consider are

$$\mathcal{B}(Mc, c) = (\mathcal{T}_c(P))^M = \underbrace{\mathcal{T}_c(P) \times \mathcal{T}_c(P) \times \cdots \mathcal{T}_c(P)}_{M \ \text{times}} \ . \tag{4.85}$$

We will fix a number ρ and a set of $N = \exp(n\rho)$ codewords for our rateless code construction. Let $\mathcal{M} = \{M_*, M_* + 1, \ldots, M^*\}$, where $M^* = n/c$ and

$$M_* = (n\rho)/(c \min\{|\mathcal{X}|, |\mathcal{Y}|\}) \ . \tag{4.86}$$

The set $\mathcal{M}$ is the set of possible decoding times for our code.

Lemma 17 (Concatenated exponential list codes)**.** *Let $\mathcal{W}$ be an AVC. For any $\delta > 0$ and $\xi > 0$, $P \in \mathcal{P}(\mathcal{X})$ with $\min_x P(x) > 0$, and $\mathcal{V}_1^M = (\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_M) \in \mathbf{V}(c)^M$, there is a c sufficiently large such that the set $\mathcal{B}(Mc, c)$ is an $(Mc, N_M, L(\mathcal{V}_1^M))$ list-*

decodable code with

$$N_M \geq \exp\left(Mc(H(X) - M\xi)\right) \tag{4.87}$$

$$L(\mathcal{V}_1^M) \leq \exp\left(c\left(\sum_{m=1}^{M} \max_{V \in \mathcal{V}_m(\mathbf{y}^{(mc)},\delta)} H(X_m|Y_m) + M\xi\right)\right) , \tag{4.88}$$

and maximal probability of error

$$\varepsilon_L \leq M \exp(-cE_2(\xi)) , \tag{4.89}$$

where $H(X)$ is calculated with respect to the distribution $P(x)$ and for a channel $V \in \mathcal{V}_m(\mathbf{y}^{(mc)}, \delta)$ the conditional entropy $H(X|Y)$ is with respect to the distribution $P(x)V(y|x)$, and $E_2(\xi) > 0$.

Proof. Choose c large enough to satisfy the conditions of Lemma 2. Our decoder will operate by list decoding each chunk separately. Let L_m be the list size guaranteed by Lemma 2 for the m-th chunk. Then the corresponding upper bound in $\prod_{m=1}^{M} L_m$ is the desired the upper bound on $L(\mathcal{V}_1^M)$. The probability of the list in each chunk not containing the corresponding transmitted chunk can be upper bounded:

$$\varepsilon_L \leq M \exp(-cE_1(\xi)) . \tag{4.90}$$

As long as c grows faster than $\log M$ the decoding error will still decay exponentially with the chunk size c. □

Our codebook is constructed by sampling codewords from the codebook $\mathcal{B}(n, c) = \mathcal{B}(M^* c, c)$. Truncating this set to blocklength Mc gives $\mathcal{B}(Mc, c)$. We want to show that for each M the sampled codewords can be used in a (Mc, N, L) list decodable code for constants L and N not depending on M. We can define for each truncation M, output sequence $\mathbf{y}_1^{Mc}$, and side information sequence $(\mathcal{V}_1, \ldots, \mathcal{V}_M)$ a "decoding bin”

$$B(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M) \subset \mathcal{X}^{Mc} \ , \tag{4.91}$$

which is the list given by the code in Lemma 17. The size of each bin is

$$|B(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M)| \leq \exp\left(c \left(\sum_{m=1}^{M} \max_{V \in \mathcal{V}_m(\mathbf{y}^{(mc)}, \delta)} H(X_m|Y_m) + M\xi \right) \right) \ . \tag{4.92}$$

Lemma 18 (Concatenated codes with constant list size)**.** *Let $\mathcal{W}$ be an AVC. For any $\epsilon > 0$, $P \in \mathcal{P}(\mathcal{X})$ with $\min_x P(x) > 0$, there is an n large enough, constant L, minimum rate $\rho > 0$, and a set of codewords $\{\mathbf{x}(j) : j \in [N]\}$ with $N = \exp(n\rho)$ such that given any CSI sequence $(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_{M^*})$ and channel output with decoding time M given by (4.75), the truncated codebook $\{\mathbf{x}_1^{Mc}(j) : j \in [N]\}$ is an (Mc, N, L) list decodable code for constant list size L satisfying*

$$L \geq \frac{12 \log |\mathcal{Y}|}{\epsilon} \ , \tag{4.93}$$

and maximal probability of decoding error

$$\varepsilon_L(M) \leq M \exp(-cE(\epsilon)) \ , \tag{4.94}$$

where $E(\epsilon) > 0$.

Proof. (Proof of Lemma 18) Fix $\epsilon > 0$. We begin with the codebook $\mathcal{B}(M^*c, c)$. Note that the truncation of this codebook to blocklength Mc for $M \in \mathcal{M}$ is the codebook in Lemma 17. Let $\{\mathbf{Z}_j : j \in [N]\}$ be $N = \exp(n\rho)$ random variables distributed uniformly on the set $\mathcal{B}(M^*c, c)$.

For any $\delta > 0$ and $\xi > 0$ we can choose $c(n)$ sufficiently large so that for any M, $\mathbf{y}_1^{Mc}$, and $\mathcal{V}_1^M \in \mathbf{V}(c)^M$ that satisfy the conditions of the decoding rule in (4.75) we have

$$\begin{aligned}
&\mathbb{P}(Z_j \in B(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M)) && (4.95)\\
&\quad \le \frac{|B(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M)|}{\exp\left(Mc\left(H(X) - \xi\right)\right)} && (4.96)\\
&\quad \le \exp\left(-c\sum_{m=1}^{M}\left(H(X) - \max_{V\in\mathcal{V}_m(\mathbf{y}^{(mc)},\delta)} H(X_m|Y_m)T_{\mathbf{y}^{(mc)}}\right) + 2Mc\xi\right) && (4.97)\\
&\quad \le \exp\left(-c\sum_{m=1}^{M} I\left(P, \mathcal{V}_m(\mathbf{y}^{(mc)}, \delta)\right) + 2Mc\xi\right) . && (4.98)
\end{aligned}$$

Let

$$G = \exp\left(-c\sum_{m=1}^{M} I\left(P, \mathcal{V}_m(\mathbf{y}^{(mc)}, \delta)\right) + 2Mc\xi\right) . \tag{4.99}$$

From our stopping rule, we know that $(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M)$ satisfies:

$$n\rho < c\sum_{m=1}^{M} I\left(P, \mathcal{V}_m(\mathbf{y}^{(mc)}, \delta)\right) - Mc\epsilon . \tag{4.100}$$

The random variable $\mathbf{1}(Z_i \in B(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M))$ is Bernoulli with parameter smaller than G, so we can bound the probability that L of the N codewords land in the set $B(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M)$ using Sanov's theorem:

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(Z_i \in B(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M)) > L/N\right) \le (N+1)^2 \exp\left(-ND\left(L/N \parallel G\right)\right) . \tag{4.101}$$

The exponent can be written as

$$L\log\left(\frac{L/N}{G}\right) + N(1 - L/N)\log\left(\frac{1 - L/N}{1 - G}\right) . \tag{4.102}$$

To deal with the $(1-L/N)\log((1-L/N)/(1-G))$ term we use the inequality $-(1-a)\log(1-a) \leq 2a$ (for small a) on the term $(1-L/N)\log(1-L/N)$ and discard the small positive term $-(1-L/N)\log(1-G)$:

$$\begin{aligned} ND\left(L/N \parallel G\right) &\geq L\log\left(\frac{L/N}{G}\right) - N2(L/N) && (4.103)\\ &= L\log\left(\frac{L/N}{G}\right) - 2L && (4.104)\\ &= L\left(-n\rho + c\sum_{m=1}^{M} I\left(P, \mathcal{V}_m(\mathbf{y}^{(mc)}, \delta)\right) - 2Mc\xi\right) + L\log L - 2L && (4.105)\\ &> L\left(Mc\epsilon - 2Mc\xi\right) + L\log L - 2L\ . && (4.106)\end{aligned}$$

For large enough n we can upper bound $(N+1)^2 \leq 2n\rho + L$. For large enough L, $L\log L > 3L$, so we can ignore those terms as well. This gives the bound

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(Z_i \in B(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M)) > L/N\right) && (4.107)\\ &\qquad\leq \exp\left(-LMc\left(\epsilon - 2\xi\right) + 2n\rho - L\log L + 3L\right) && (4.108)\\ &\qquad\leq \exp\left(-LMc\left(\epsilon - 2\xi\right) + 2n\rho\right)\ . && (4.109)\end{aligned}$$

Now the number of decoding bins $B(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M)$ can be bounded by

$$\left|\left\{B(M, \mathbf{y}_1^{Mc}, \mathcal{V}_1^M) : \mathcal{V}_1^M \in \mathbf{V}(c)^M,\ \mathbf{y}_1^{Mc} \in \mathcal{Y}^{Mc}\right\}\right| \leq |\mathcal{Y}|^{Mc} c^{Mv}\ . \qquad (4.110)$$

Therefore we can take a union bound:

$$\mathbb{P}\left(\bigcup_{\{B(M,\mathbf{y}_1^{Mc},\mathcal{V}_1^M)\}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(Z_i \in B(M,\mathbf{y}_1^{Mc},\mathcal{V}_1^M)) > L/N \right\} \right) \tag{4.111}$$

$$\leq \exp\left(-LMc\left(\epsilon - 2\xi\right) + Mc \log|\mathcal{Y}| + Mv\log c + 2n\rho\right) \ . \tag{4.112}$$

From (4.86) we know ρ is sufficiently small such that $n\rho \leq Mc\log|\mathcal{Y}|$ for all M. Then we can choose n and c sufficiently large such that the upper bound becomes:

$$\exp\left(-LMc\left(\epsilon - 2\xi\right) + 4Mc\log|\mathcal{Y}|\right) \ . \tag{4.113}$$

If $\epsilon > 2\xi$ then we can choose

$$L > \frac{4\log|\mathcal{Y}|}{(\epsilon - 2\xi)} \ , \tag{4.114}$$

to guarantee that subsampling will yield a good list-decodable code for all $M \in \{M_*, \ldots, M^*\}$. Choosing $\xi = \epsilon/3$ and $E(\epsilon) = E_2(\epsilon/3)$ yields the result. □

4.3.4 Proof of Theorem 17

Proof. We will use the codebook from Lemma 18. Since the set of messages of fixed size N, we use the construction of Lemma 13. This makes the code, when decoded at Mc an $(Mc, \exp(n\rho)/\sqrt{K(n)}, K(n))$ randomized code with probability of error

$$\hat{\varepsilon}(M,\mathbf{s}) \leq M\exp(-cE(\epsilon)) + \frac{2Ln\rho}{\sqrt{K}\log K} \ . \tag{4.115}$$

Then we can use choose $L = 12(\log|\mathcal{Y}|)/\epsilon$ to get

$$\hat{\varepsilon}(M, \mathbf{s}) \le M \exp(-cE(\epsilon)) + \frac{24n\rho \log|\mathcal{Y}|}{\epsilon\sqrt{K}\log K} \ . \tag{4.116}$$

□

4.4 An application to individual sequence channels

In some cases we can apply the partially derandomized code in Section 4.2 to the scheme proposed by Eswaran, Sarwate, Sahai, and Gastpar for communicating over a channel with an individual state sequence [62]. We will focus on the binary modulo-additive case:

$$\mathbf{y} = \mathbf{x} \oplus \mathbf{s} \ , \tag{4.117}$$

for which the empirical frequency of 1's in $\mathbf{s}$ is equal to the cost in the AVC setting. In this section we will describe the coding strategy and show how the common randomness resources required for the scheme can be reduced using a randomized rateless code for the AVC.

In the individual sequence model, the channel state $\mathbf{s}$ is fixed prior to transmission but is otherwise arbitrary. As in a rateless code, the goal is to achieve a rate close to that given by the channel averaged over $\mathbf{s}$. The coding scheme used by Eswaran et al. achieves the "empirical capacity" for a model similar to Shayevitz and Feder's but using a limited feedback strategy. The strategy adapts strategies for reducing feedback [123, 124, 122] that were originally derived in the context of error exponents. The methodology is inspired by Hybrid ARQ [144]. The decoder uses the feedback link to terminate rounds that are too noisy but otherwise attempts to correct the error in less noisy rounds.

The encoder attempts to send k bits over the channel during a variable length *round*. The encoder sends *chunks* of the codeword to the decoder, after which the decoder feeds back a decision as to whether it can decode. The encoder and decoder use common randomness to choose a set of randomly chosen *training* positions during which the encoder sends a fixed message. The decoder uses the training positions to estimate the channel. If the number of bits than can be transmitted over a channel with the estimated *empirical mutual information* exceeds k, then the decoder attempts to decode. This combination of training-based channel estimation and robust decoding exploiting the limited feedback yields rates asymptotically equal to those with advance knowledge of the average channel.

Eswaran et al. use a coding strategy that operates over a total blocklength of N. The state sequence $\mathbf{s}$ of length N is also fixed prior to transmission. Because the channel quality is not known in advance, the number of bits which can be reliably transmitted during this block is also unknown. The encoder attempts to send $k(N)$ bits at a time using a rateless code. Once the k bits have been received successfully, it attempts to send the next k bits. Within each chunk of the rateless code, some channel uses are reserved for "training" which can allow the decoder to estimate the empirical channel, which in this case is the empirical fraction of 1's in the state sequence.

Corollary 4 (Binary additive individual sequence channels [62])**.** *For the binary modulo-additive channel with an individual noise sequence, there is a coding strategy that with probability* $1 - \varepsilon(N)$ *achieves the rate*

$$R \geq 1 - h_b(p) - \rho(N) \ , \tag{4.118}$$

with feedback rate

$$R_{\mathrm{fb}} = \nu(N) \ . \tag{4.119}$$

and $h_b(\cdot)$ *is the binary entropy function. Furthermore, as* $N \to \infty$ *we have* $\rho(N) \to 0$, $\nu(N) \to 0$, *and* $\varepsilon(N) \to 0$.

4.4.1 The coding algorithm

We divide the blocklength N into *chunks* of length $c = c(N)$. Feedback occurs at the end of chunks with three possible messages: "BAD NOISE", "DECODED", and "KEEP GOING".

The encoder attempts to send $k = k(N)$ bits over several chunks comprising a *round*. Let $V_m = (m-1)c + 1, (m-1)c + 2, \ldots, mc$ be the time indices in the m-th chunk within a round. For each chunk m, the decoder and encoder choose $t = t(N)$ *training positions* T_m (via common randomness) during which a known sequence is transmitted to enable the decoder to estimate the empirical channel. The remaining time indices $U_m = V_m \backslash T_m$ are used to transmit the codeword. Let $\mathcal{V}_m = V_1, \ldots, V_m$, $\mathcal{T}_m = T_1, \ldots, T_m$, and $\mathcal{U}_m = U_1, \ldots, U_m$ be the time indices up to the m-th chunk for the round, training, and codeword positions, respectively.

Fix an distribution $P(x) \in \mathcal{P}_c(\mathcal{X})$. The encoder and decoder choose a random codebook of type P for each round. In a round, the encoder divides the codebook into segments of length $c - t$ and transmits the n-th segment over the $c - t$ non-training positions in U_m.

The decoder uses the training positions to estimate the empirical noise distribution in that chunk. After each chunk the decoder will decide either (a) that the empirical noise is too bad and tell the encoder to terminate the round and start over, (b) to decode the k bits and tell the encoder to terminate the round, or (c) that it cannot

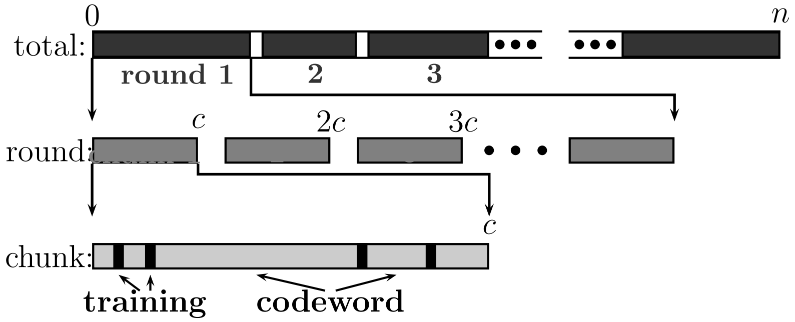


Figure 4.2: After each chunk of length b feedback can be sent. Rounds end by decoding a message or declaring the noise to be bad.

decode yet and tell the encoder to send another chunk.

A formal description of the coding strategy follows, and an illustration is provided in Figure 4.2. At the beginning of round r, the encoder and decoder use common randomness to choose a random codebook of type P to be used in that round. Let $\mathbf{x}(r)$ denote the codeword to be sent in round r. Let $\mathbf{s}_m = \mathbf{s}_{m(r)}$ denote the state sequence during the m-th chunk of round r. We will suppress the dependence on r for simplicity. For a set of indices $\mathcal{J} = \{j_1, j_2, \ldots\}$, we will let $\mathbf{s}(\mathcal{J}) = (z_{j_1}, z_{j_2}, \ldots, z_{j_{|\mathcal{J}|}})$, so that $\mathbf{s}(T_m)$ is the state vector during the training in chunk m, $\mathbf{s}(\mathcal{T}_m)$ is the state during all the training positions, $\mathbf{s}(\mathcal{U}_m)$ is the state during all the non-training positions, and $\mathbf{s}(\mathcal{V}_m)$ is the state vector of the current round up to the m-th chunk.

For each round, the following steps are repeated for each chunk:

1. The encoder and decoder choose t positions T_m to use for the training in chunk m using common randomness.

2. The encoder transmits the chunk. At times $j \in T_m$ the encoder sends 0. In the $c-t$ remaining positions the encoder sends $\mathbf{x}(\{(m-1)(c-t)+1, \ldots, m(c-t)\})$, which are the next $c-t$ entries in the codeword corresponding to the k bits to be sent in the current round.

3. The decoder estimates the empirical channel $W_{\mathbf{z}(\mathcal{U}_m)}(y|x)$ in chunk m and the empirical channel over the round so far:

$$\hat{p}_{\mathrm{emp}}^{(m)} = \frac{|\mathcal{X}|}{t} \cdot |\{j \in T_m(x) : y_j = 1\}|$$
$$\tilde{p}_{\mathrm{emp}}^{(m)} = \frac{1}{m} \sum_{i=1}^{m} \hat{p}_{\mathrm{emp}}^{(i)} \ .$$

4. The decoder makes a decision based on $\tilde{p}_{\mathrm{emp}}^{(m)}$ and m:

   (a) if

   $$1 - h_b(\tilde{p}_{\mathrm{emp}}^{(m)}) < \tau(N) \ , \tag{4.120}$$

   then the decoder feeds back "BAD NOISE" and the round is terminated without decoding the k bits. In the next round, the encoder will attempt to resend the k bits from this round.

   (b) if

   $$\frac{k}{(c-t) \times m} < 1 - h_b(\tilde{p}_{\mathrm{emp}}^{(m)}) - \epsilon_1(N) \ , \tag{4.121}$$

   then the decoder decodes, feeds back "DECODED," and the encoder starts a new round.

   (c) otherwise the decoder feeds back "KEEP GOING" and goes to 2).

The coding strategy uses $\log 3$ bits of feedback per chunk for the decision messages ("BAD NOISE," "DECODED," and "KEEP GOING"). Letting b get large with N causes the feedback rate to go to zero.

4.4.2 Application of rateless code and resource analysis

We can now apply the code construction from Lemma 15 to partially derandomize the codebook for a single round of the overall strategy of Eswaran et al. The decoding threshold (4.121) defines the minimum rate ρ as $\epsilon_1(N)$. We can use the rateless codebook with $\exp(k(N))$ messages and chunk size $c(N)$. The cost information available to the decoder is given by the training estimates:

$$\hat{\lambda}_m = \hat{p}_{\text{emp}}^{(m)} \tag{4.122}$$

$$\hat{\Lambda}_M = \hat{p}_{\text{emp}}^{(M)} \ . \tag{4.123}$$

The coding strategy uses common randomness to choose the training positions for each chunk and the codebook for each round. The codebook from Lemma 15 is decodable with small probability of error at blocklength mc as long as we can guarantee that the true cost is larger than the codebook's designed cost:

$$\Lambda_M \leq \tilde{\Lambda}_M \ . \tag{4.124}$$

Eswaran et al. [62] proved that with probability going to 1, the channel estimates are close to the true channel:

$$\left| \lambda_M - \hat{\lambda}_M \right| \leq \epsilon \ . \tag{4.125}$$

By increasing ϵ_{26} we can guarantee that for some $\epsilon' > \epsilon$ we have

$$\tilde{\Lambda}_M > \hat{\lambda}_M + \epsilon' > \lambda_M \ . \tag{4.126}$$

Thus we can use the partially derandomized codebook to limit the common randomness required to operate this scheme of Eswaran et al.

We would like to quantify the amount of common randomness required by the scheme. There are two parts of the scheme which require common randomness: the choice of training positions and the choice of codebook.

- For each chunk, the encoder and decoder must agree upon $t(N)$ training positions, which takes $t(N) \log c(N)$ bits per chunk, or

  $$\frac{t(N)N}{c(N)} \log c(B) \ . \tag{4.127}$$

  bits over the whole scheme.

- The encoder and decoder must agree on a codebook to use for each round. The code from Section 4.2 can be used with $K(N)$ bits per round, where $K(N) = \Omega(\log k(N))$. Because new training positions are selected for each chunk, a new codebook is not needed when the round is terminated due to excess noise. Since there are at most $N/k(N)$ rounds, the scheme needs

  $$\frac{N}{k(N)} K(n) \tag{4.128}$$

  bits to choose the codebooks.

Thus, as we can see, the total amount of common randomness needed for the algorithm is

$$\frac{t(N)N}{c(N)} \log c(B) + \frac{N}{k(N)} K(n) \tag{4.129}$$

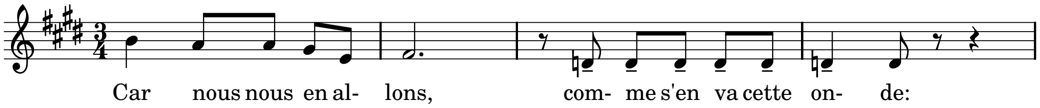
bits, which can be made sublinear in N as long as $c(N)$ and $k(N)$ grow fast enough. In particular, with the choices given in [62] of $k(N) = N^{1/2}$, $c(N) = N^{1/4}$, and $t(N) = N^{1/8}$ we can choose $K(N)$ to grow slower than $N^{1/2}$.

4.5 Conclusions

We saw in this chapter that the derandomization strategies of the previous chapter extend to the study of rateless coding when the decoder can obtain an estimate of the empirical channel. Under the standard AVC model, we studied the case where decoder obtains an estimate of the average state cost over a "chunk." We constructed a coding strategy and bound its loss from the true empirical mutual information in terms of the channel estimation error. For channels with the more robust nosy noise model, we constructed a strategy that list decodes on a chunk-by-chunk basis. This architecture may be interesting for more practical code constructions, given the recent research interest on list decoding with soft information [96].

An additional application of the results and techniques of this chapter is to the problem of communicating over channels with individual state sequences as studied in [62]. The common randomness required to use our code constructions can be generated from zero-rate noiseless feedback from the decoder to the encoder. In the scheme presented in [62], the partial channel state information is generated by training sequences in the forward link.

Finally, although the results in this chapter are for finite alphabets, extensions to continuous alphabets and the Gaussian AVC setting [86, 87, 48] should be possible using appropriate approximation techniques. An interesting rateless code using lattice constructions has been proposed by Erez et al. in [58], and it would be interesting to see if that approach can work for more robust arbitrarily varying channel models.



– *Beau Soir*, Claude Debussy (lyrics by Paul Bourget)

Chapter 5

Continuous AVCs : the Gaussian case

5.1 Introduction and channel model

In this chapter we study an AVC with continuous-alphabets known as the Gaussian AVC (GAVC). In the GAVC an additive white Gaussian noise (AWGN) channel is modified by adding a power-constrained jamming interference signal. This channel model was first proposed by Hughes and Narayan [86]. As in the discrete AVC with constraints [45], the capacity is well defined when the power constraints Γ and Λ on the input and jammer are required to hold almost surely. The randomized coding capacity under maximal error $C_r(\Gamma, \Lambda)$ is the same as that of an AWGN channel treating the jammer as more Gaussian noise. Csiszár and Narayan [48] showed that the deterministic coding capacity under average error $\overline{C}_d(\Gamma, \Lambda)$ exhibits a threshold due to symmetrizability. If $\Gamma \leq \Lambda$ then the channel is symmetrizable and $\overline{C}_d(\Gamma, \Lambda) = 0$. Otherwise $\overline{C}_d(\Gamma, \Lambda)$ is equal to $C_r(\Gamma, \Lambda)$ – that is, $\overline{C}_d(\Gamma, \Lambda)$ is given by the AWGN capacity treating the jammer as additional Gaussian noise.

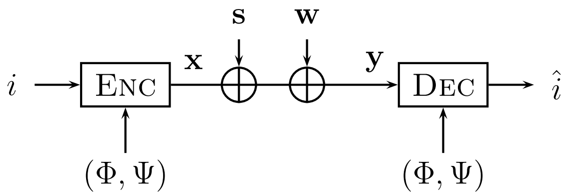


Figure 5.1: The Gaussian arbitrarily varying channel under randomized coding.

For the vector Gaussian channel, the randomized coding capacity was found by Hughes and Narayan [87] and the deterministic coding capacity by Csiszár [42]. For multiuser Gaussian AVCs, the only results are due to La and Anantharam [100], who found the capacity of a modified arbitrarily varying multiple access channel. Their channel model puts additional constraints on the jammer, which allows the transmitters to time-share between rates.

The Gaussian AVC is shown in Figure 5.1. For an input sequence $\mathbf{x} \in \mathbb{R}^n$ the output of the Gaussian AVC is given by

$$\mathbf{y} = \mathbf{x} + \mathbf{s} + \mathbf{w} \ . \tag{5.1}$$

The input is corrupted by iid additive white Gaussian noise $\mathbf{w}$ with variance σ^2 and an unknown interference vector $\mathbf{s}$. The input signal $\mathbf{x}$ and jammer signal $\mathbf{s}$ are constrained in power:

$$\frac{1}{n} \|\mathbf{x}\|^2 \le \Gamma \tag{5.2}$$

$$\frac{1}{n} \|\mathbf{s}\|^2 \le \Lambda \ . \tag{5.3}$$

Randomized coding for this channel was first studied by Hughes and Narayan [86].

A (n, N) **deterministic code** $\mathcal{C}$ satisfying the input constraint Γ is a pair of maps

(ϕ, ψ) with

$$\phi : [N] \to \mathbb{R}^n \tag{5.4}$$

$$\psi : \mathbb{R}^n \to [N] \ , \tag{5.5}$$

such that for all $i \in [N]$ we have $\|\phi(i)\|^2 \leq n\Gamma$. A (n, N) **randomized code $\mathbf{C}$** satisfying the input constraint Γ is a random variable taking on values in the set of (n, N) deterministic codes. It is written as a pair of random maps (Φ, Ψ) where each realization is an (n, N) deterministic code satisfying the constraint Γ. If (Φ, Ψ) almost surely takes values in a set of K codes, then we call this an (n, N, K) randomized code.

The maximal probability of error for randomized coding and average probability of error for deterministic coding with jamming signal $\mathbf{s}$, respectively are

$$\varepsilon(\mathbf{C}, \mathbf{s}) = \max_i \mathbb{E}_{\mathbf{C}} \left[\mathbb{P}_{\mathbf{w}} \left(\psi(\mathbf{x}_i + \mathbf{s} + \mathbf{w}) \neq i \right) \right] \tag{5.6}$$

$$\overline{\varepsilon}(\mathcal{C}, \mathbf{s}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_{\mathbf{w}} \left(\psi(\mathbf{x}_i + \mathbf{s} + \mathbf{w}) \neq i \right) \ . \tag{5.7}$$

The maximal and average error are given by maximizing these quantities over the state $\mathbf{s}$:

$$\varepsilon(\mathbf{C}) = \max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \varepsilon(\mathbf{C}, \mathbf{s}) \tag{5.8}$$

$$\overline{\varepsilon}(\mathcal{C}) = \max_{\mathbf{s} \in \mathcal{S}^n(\Lambda)} \overline{\varepsilon}(\mathcal{C}, \mathbf{s}) \ . \tag{5.9}$$

As in the discrete case, under maximal error we can think of the state $\mathbf{s}$ as depending on the transmitted message. For average error $\mathbf{s}$ may depend on the codebook $\mathcal{C}$ but not the message.

We say a rate R is **achievable** under maximal error with randomized coding if there exists a sequence of $(n, \exp(nR))$ randomized codes whose maximal error goes to 0 as $n \to \infty$. The randomized coding capacity under maximal error $C_r(\Gamma, \Lambda)$ is the supremum of the achievable rates under maximal error with randomized coding. Similarly, we say R is achievable under average error with deterministic coding if there is a sequence of $(n, \exp(nR))$ deterministic codes whose average error goes to 0 as $n \to \infty$. The deterministic coding capacity under average error $\overline{C}_d(\Gamma, \Lambda)$ is the supremum of achievable rates under average error with deterministic coding.

Hughes and Narayan [86] showed that if the input and jammer are both bounded in power almost surely and the random variable $\mathbf{C}$ is unconstrained, then the capacity is equal to that of an additive white Gaussian noise (AWGN) channel with the jammer treated as additive noise:

$$C_r(\Gamma, \Lambda) = \frac{1}{2} \log \left(1 + \frac{\Gamma}{\Lambda + \sigma^2} \right) \ . \tag{5.10}$$

Csiszár and Narayan [48] showed that for deterministic codes, the capacity is equal to (5.10) if and only if the encoder has a higher power limit than the jammer:

$$\overline{C}_d(\Gamma, \Lambda) = \begin{cases} 0 & \Gamma \le \Lambda \\ \frac{1}{2} \log \left(1 + \frac{\Gamma}{\Lambda + \sigma^2} \right) & \Gamma > \Lambda \ . \end{cases} \tag{5.11}$$

This condition is equivalent to non-symmetrizability. If $\Lambda \ge \Gamma$ then there exists a channel $U(\mathbf{S}|\mathbf{X}')$ such that

$$V(\mathbf{Y}|\mathbf{X}, \mathbf{X}') = \int_{\mathbb{R}^n} W(\mathbf{Y}|\mathbf{X}, \mathbf{S}) U(\mathbf{S}|\mathbf{X}') d\mathbf{S} \ . \tag{5.12}$$

is symmetric in $\mathbf{X}$ and $\mathbf{X}'$. The channel $U(\mathbf{S}|\mathbf{X}')$ corresponds to choosing $\mathbf{S} = \mathbf{X}' + \mathbf{Z}$, where $\mathbf{X}'$ is a randomly chosen codeword from the deterministic code and $\mathbf{Z}$ is an iid

Gaussian vector with variance $\Lambda - \Gamma$.

The capacity-achieving codebooks used for the GAVC are formed by drawing vectors $\mathbf{x}$ from the uniform distribution on the sphere of radius $\sqrt{n\Gamma}$. For fully randomized coding, it is straightforward to show that any jammer input $\mathbf{s}$ is bad for a vanishingly small fraction of such codebooks. Because of the power constraints on the input and jammer, a prefix-based scheme such as the elimination technique [6] cannot be used to convert a randomized code for the GAVC into a deterministic one. This is because the jammer can pool its power to jam the prefix, rendering the decoder incapable of decoding that part of the message. Under deterministic coding, more careful geometric arguments are needed to show that for $\Gamma > \Lambda$ there exists a codebook which is decodable with small probability of error for all $\mathbf{s}$.

In this chapter we explore a few aspects of the GAVC. We first derive results analogous to those in Chapter 3 and show that a modest amount of common randomness is sufficient to achieve the randomized coding capacity. This result has applications to the arbitrarily varying degraded Gaussian broadcast channel. For deterministic coding we propose a new channel model in which there are two sources of interference, one of which is known to the transmitter but not the jammer. We propose a "dirty paper" scheme that exploits the known interference signal. These are strategies that have found uses in applications from digital watermarking to multiantenna broadcasting. We will apply our construction to an information-theoretic model for cognitive radio systems.

5.2 Partial derandomization for the GAVC

Our goal in this section is to quantify the *amount* of randomization, or *key size*, that is needed to obtain the randomized coding capacity. Our main result is that as in the discrete case, we can use a sub-exponential number (in the blocklength n)

of codebooks to obtain an asymptotically decreasing upper bound on the average probability of error. Although the probability of error decreases for modest key sizes, in order to get an exponential decrease in the probability of error our construction requires an exponential number of codebooks. This gives a characterization of the achievable error decay as a function of the common randomness available between the encoder and decoder in the same spirit as Chapter 3.

5.2.1 Rotated codebooks : a randomized code construction

The class of randomized codes we consider can be built in two steps. As in the construction in [88], we "modulate" a single Gaussian codebook. Let $N = \exp(nR)$ and M be an arbitrary integer.

1. Let $\mathcal{B} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of N vectors on the sphere of radius $\sqrt{n\Gamma}$. We can choose this set to have small maximal error for both the AWGN channel with noise variance $\Lambda + \sigma^2$ and the channel with additive noise $\mathbf{V}_n + \mathbf{w}$, where $\mathbf{V}_n$ is uniform on the sphere of radius $\sqrt{n\Lambda}$ and $\mathbf{w}$ is iid Gaussian noise with variance σ^2.

2. Let $\{U_k : k = 1, 2, \ldots, K\}$ be $n \times n$ unitary matrices generated uniformly from the set of all unitary matrices. Without loss of generality we take $U_1 = I$.

3. The randomized code is uniform on the set $\{U_k \mathcal{B} : k = 1, 2, \ldots, K\}$. To send message i, the encoder draws an integer k uniformly from $\{1, 2, \ldots, K\}$ and encodes its message as $U_k \mathbf{x}_i$.

4. The decoder knows k and chooses the codeword in $\mathcal{B}_k$ that minimizes the distance to the received vector:

$$\phi(\mathbf{y}, k) = \operatorname*{argmin}_{j} \|\mathbf{y} - U_k \mathbf{x}_j\| \ . \tag{5.13}$$

For all rates below $(1/2)\log(1+\Gamma/(\Lambda+\sigma^2))$ we can choose the codebook $\mathcal{B}$ to have exponentially decaying probability of error [139,67,102] for both the AWGN channel with noise variance $\Lambda+\sigma^2$ and the channel with additive noise $\mathbf{V}_n + \mathbf{w}$:

$$\varepsilon(\mathcal{B}) \le \exp(-nE(n^{-1}\log N)) \ . \tag{5.14}$$

We can use this result to get a lower bound on the pairwise distance between any two codewords. This lower bound will be useful in the proof of Theorem 18 below.

Consider two codewords $\mathbf{x}_i$ and $\mathbf{x}_j$ from the codebook. Let $\gamma > 0$ be half the distance between them:

$$\|\mathbf{x}_i - \mathbf{x}_j\| = 2\gamma \ . \tag{5.15}$$

Suppose that we transmit $\mathbf{x}_i$ over an AWGN channel with noise variance $\Lambda+\sigma^2$. Then the probability of error for message i can be lower bounded by the chance that the noise in the direction of $\mathbf{x}_j - \mathbf{x}_i$ is larger than γ. Since the noise is iid, the error can be bounded by the integral of a Gaussian density [152]:

$$\begin{aligned}
\varepsilon(i) &\ge \frac{1}{\sqrt{2\pi(\Lambda+\sigma^2)}} \int_{\gamma}^{\infty} \exp\left(-\frac{1}{2(\Lambda+\sigma^2)} z^2 \right) dz & (5.16)\\
&> \sqrt{\frac{\Lambda+\sigma^2}{2\pi\gamma^2}} \left(1 - \frac{\Lambda+\sigma^2}{\gamma^2} \right) \exp(-\gamma^2/2) \ . & (5.17)
\end{aligned}$$

Therefore there exists a $\mu > 0$ such that for sufficiently large n we have $\gamma > (\mu/2)\sqrt{n}$ for some $\mu > 0$, which means that

$$\|\mathbf{x}_i - \mathbf{x}_j\| > \mu\sqrt{n} \ . \tag{5.18}$$

We will first prove our result for no noise, so $\sigma^2 = 0$. The extension to the case with $\sigma^2 > 0$ is straightforward. The proof entails showing that there exists a randomized code as described above that can achieve the randomized-coding capacity for the AVC.

Theorem 18. *Let $K(n)$ be chosen such that $K(n)/n \to \infty$ and $n^{-1} \log(K(n)/n) \to 0$. For input power constraint Γ, jammer power constraint Λ, and $\zeta > 0$ there is an n sufficiently large and an $(n, N, K(n))$ randomized code for the GAVC with $\sigma^2 = 0$ of rate $R < C_r(\Gamma, \Lambda)$, where*

$$C_r(\Gamma, \Lambda) = \frac{1}{2} \log \left(1 + \frac{\Gamma}{\Lambda} \right) \ , \tag{5.19}$$

whose error satisfies

$$\varepsilon(n) = \zeta \frac{n}{K(n)} \ . \tag{5.20}$$

That is, the randomized coding capacity $C_r(\Gamma, \Lambda)$ is achievable using codes whose key size is superlinear in the blocklength.

Proof. Fix a rate $R < C_r(\Gamma, \Lambda)$. We will suppress the dependence of $K(n)$ on n in the proof. We need to show that for n sufficiently large, there exists a codebook $\mathcal{B}$ and K unitary matrices $\{U_k\}$ such that the probability of error is bounded for any choice of $\mathbf{s}$. To do this we will first show that if $\mathbf{s}$ lies in a dense subset of the $\sqrt{n\Lambda}$ sphere, then the event that the average error for K randomly chosen matrices $\{U_k\}$ is too large has probability exponentially small in K. Therefore we can choose a collection $\{U_k\}$ that satisfies the probability of error bound for any $\mathbf{s}$.

Consider the codebook $\mathcal{B}$ of N vectors from the sphere of radius $\sqrt{n\Gamma}$. We know that the expected performance of this code is good for an additive noise channel with noise $\mathbf{V}_n$ distributed uniformly on the sphere of radius $\sqrt{n\Lambda}$. That is, for any $\delta > 0$

there exists an n sufficiently large such that

$$\max_{i \in [N]} \mathbb{E}_{\mathbf{V}_n} \left[\varepsilon(i, \mathbf{V}_n) \right] < \exp(-nE(R)) \tag{5.21}$$

$$\triangleq \delta \ . \tag{5.22}$$

Suppose that we sample K points $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K$ independently from the distribution of $\mathbf{V}_n$. Then using the same argument from Lemma 12 on page 99 we can see that

$$\mathbb{P}_{\mathbf{V}_n} \left(\frac{1}{K} \sum_{k=1}^{K} \varepsilon(i, \mathbf{v}_k) \geq t \right) = \exp\left(-K(t \log \delta^{-1} - h_b(t) \log 2) \right) \ . \tag{5.23}$$

We can take a union bound over all $i \in [N]$:

$$\mathbb{P}_{\mathbf{V}_n} \left(\bigcup_{i \in [N]} \left\{ \frac{1}{K} \sum_{k=1}^{K} \varepsilon(i, \mathbf{v}_k) \geq t \right\} \right) = \exp\left(-K(t \log \delta^{-1} - h_b(t) \log 2) + \log N \right) \ . \tag{5.24}$$

Thus the probability that the collection of points $\{\mathbf{v}_m\}$ induces an error probability that exceeds t is exponentially small in K.

Now consider drawing K unitary matrices $\{U_k : k = 1, 2, \ldots, K\}$ uniformly. For a fixed $\mathbf{v}$, the points

$$\mathbf{v}_k = U_k^{-1} \mathbf{v} \tag{5.25}$$

are uniform samples from $\mathbf{V}_n$. Unlike Lemma 12, we cannot take a union bound over all $\mathbf{s}$, so we require an approximation argument. Let $\{\mathbf{a}_m : m = 1, 2, \ldots M\}$ be a set of vectors on the sphere of radius $\sqrt{n\Lambda}$. Then another union bound yields the following:

$$\mathbb{P}\left(\bigcup_{m=1}^{M} \bigcup_{i \in [N]} \left\{ \frac{1}{K} \sum_{k=1}^{K} \varepsilon(i, U_k^{-1} \mathbf{a}_k) \geq t \right\} \right) \tag{5.26}$$

$$\leq \exp\left(-K(t \log \delta^{-1} - h_b(t) \log 2) + \log M + \log N \right) \ . \tag{5.27}$$

Results of Wyner and Lapidoth [157, 103] tell us that there exists a collection of $\exp(n(\rho + \epsilon))$ points on the $\sqrt{n\Lambda}$-sphere such that any point on the $\sqrt{n\Lambda}$-sphere is at most a distance η from one of the points, where η and ρ are related by $\rho = (1/2)\log(\Lambda/\eta^2)$. If we choose $M = \exp(n(\rho + \epsilon))$ and let $\{\mathbf{a}_m\}$ be the corresponding rate-distortion codebook, we get from (5.27):

$$\mathbb{P}\left(\bigcup_{m=1}^{M} \bigcup_{i \in [N]} \left\{ \frac{1}{K} \sum_{k=1}^{K} \varepsilon(i, U_k^{-1} \mathbf{a}_m) \geq t \right\} \right) \tag{5.28}$$

$$\leq \exp\left(-K(t \log \delta^{-1} - h_b(t) \log 2) + n(\rho + R + \epsilon) \right) \ . \tag{5.29}$$

If $K(n)/n \to \infty$ then the probability that the error is smaller than t for the M points $\{\mathbf{a}_m\}$ can be made arbitrarily close to 1 for any η. The next step is to argue that we can extend the bound from $\mathbf{s} \in \{\mathbf{a}_m\}$ to all $\mathbf{s}$.

Because $R < C_r(\Gamma, \Lambda)$, for a sufficiently small constant ν we have

$$R < \frac{1}{2} \log\left(1 + \frac{\Gamma}{(1+\nu)^2 \Lambda} \right) \ , \tag{5.30}$$

for some sufficiently small constant ν. That is, we can choose our code to have small error probability for noise of variance $(1+\nu)^2\Lambda$. From (5.29) we know that for each message i there is a set $\mathcal{K}_i$ of at least $(1-t)K$ keys for which

$$\left\|\mathbf{x}_i - \mathbf{x}_j + (1+\nu)U_k^{-1}\mathbf{a}_m\right\| > (1+\nu)\|\mathbf{a}_m\| \qquad \forall j \neq i \ . \tag{5.31}$$

Equivalently, we can write

$$2\left\langle \mathbf{x}_j - \mathbf{x}_i,\ U_k^{-1}\mathbf{a}_m \right\rangle < \frac{1}{1+\nu}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \ . \tag{5.32}$$

Now suppose the jammer's input is $\mathbf{s}$. For each $k \in \mathcal{K}_i$ s, the rate-distortion codebook property guarantees that $\mathbf{s}$ is only a distance $\eta\sqrt{n}$ from some point $U_k^{-1}\mathbf{a}_m$. We would like to prove a bound like (5.32) for all $\mathbf{s}$. To start:

$$\begin{aligned} 2\left\langle \mathbf{x}_j - \mathbf{x}_i,\ \mathbf{s}\right\rangle &= 2\left\langle \mathbf{x}_j - \mathbf{x}_i,\ \mathbf{s} - U_k^{-1}\mathbf{a}_m\right\rangle + 2\left\langle \mathbf{x}_j - \mathbf{x}_i,\ U_k^{-1}\mathbf{a}_m\right\rangle && (5.33) \\ &< 2 \cdot \|\mathbf{x}_i - \mathbf{x}_j\| \cdot \eta\sqrt{n} + \frac{1}{1+\nu}\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 && (5.34) \\ &< \left(2\frac{\eta}{\mu} + \frac{1}{1+\nu}\right)\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \ , && (5.35) \end{aligned}$$

where we used (5.32), the Cauchy-Schwartz inequality, the distortion bound for $\{\mathbf{a}_m\}$, and and (5.18). Now choose η sufficiently small so that

$$2\left\langle \mathbf{x}_j - \mathbf{x}_i,\ \mathbf{s}\right\rangle < \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \ . \tag{5.36}$$

This shows that the minimum distance decoding rule results in a small error probability for all $\mathbf{s}$ with $\|\mathbf{s}\|^2 = n\Lambda$.

The last thing we need is to show that the average error probability is monotonic in the length of the jamming vector for a given direction. Suppose that there was an error for $\mathbf{s}$ but now the jammer inputs $(1+b)\mathbf{s}$. Then there is an error if

$$\|\mathbf{x}_i - \mathbf{x}_k + (1+b)\mathbf{s}\| \leq (1+b)\left\|\mathbf{s}\right\| \ .$$

But we can easily bound this using the triangle inequality:

$$\begin{aligned} \|\mathbf{x}_i - \mathbf{x}_k + (1+b)\mathbf{s}\| &\le \|\mathbf{x}_i - \mathbf{x}_k + \mathbf{s}\| + \|b\mathbf{s}\| \\ &\le (1+b)\,\|\mathbf{s}\| \ . \end{aligned}$$

Thus the error probability can only become smaller for shorter jamming inputs $\mathbf{s}$.

We have shown that for any $t > 0$ there is an n sufficiently large such that with high probability, choosing a random set of K unitary matrices $\{\mathbf{U}_k\}$ results in a randomized code whose error can be made smaller than t. As in Theorem 13 we can choose the error bound t to satisfy (5.20). Therefore there such a randomized code exists. □

The extension to the noisy case follows the same argument. The equivalent channel is

$$\mathbf{y} = \mathbf{x_i} + U_k^{-1}(\mathbf{s} + \mathbf{w}) \ . \tag{5.37}$$

The effective noise $U_k^{-1}\mathbf{w}$ has the same distribution as $\mathbf{w}$, and we can follow the arguments previously by taking expectations over the noise $\mathbf{w}$.

Theorem 19. *Let $K(n)$ be chosen such that $K(n)/n \to \infty$ and $n^{-1}\log(K(n)/n) \to 0$. For input power constraint Γ, jammer power constraint Λ, and $\zeta > 0$ there is an n sufficiently large and an $(n, N, K(n))$ randomized code for the GAVC of rate $R < C_r(\Gamma, \Lambda)$, where*

$$C_r(\Gamma, \Lambda) = \frac{1}{2}\log\left(1 + \frac{\Gamma}{\Lambda}\right) \ , \tag{5.38}$$

whose error satisfies

$$\varepsilon(n) = \zeta \frac{n}{K(n)} \ . \tag{5.39}$$

That is, the randomized coding capacity $C_r(\Gamma, \Lambda)$ is achievable using codes whose key size is superlinear in the blocklength.

We can characterize different key-error tradeoffs using the proof of Theorem 19.

- Suppose $K(n) = n^{\alpha}$ for $\alpha > 1$. Then

$$\varepsilon(n) = O(n^{1-\alpha}) \ . \tag{5.40}$$

- Suppose $K(n) = \exp(n^{\gamma})$ for $\gamma \in (0, 1)$. In this case we have

$$\varepsilon(n) = O(\exp(-n^{\gamma})) \ . \tag{5.41}$$

The first example show that choosing $K(n)$ polynomial in n is possible, so $O(\log n)$ bits is a sufficient key size.

5.2.2 Gaussian Broadcast

We can apply Theorem 19 to the a degraded broadcast channel with a common jammer. Here the two users receive

$$\mathbf{y}_1 = \mathbf{x} + \mathbf{s} + \mathbf{w}_1 \tag{5.42}$$

$$\mathbf{y}_2 = \mathbf{x} + \mathbf{s} + \mathbf{w}_2 \ , \tag{5.43}$$

where $\mathbf{w}_1$ is iid Gaussian with variance σ_1^2, $\mathbf{w}_2$ is iid Gaussian with variance σ_2^2, and $\sigma_1^2 < \sigma_2^2$. The channel is shown in Figure 5.2. We call receiver 1 the strong user and receiver 2 the weak user.

An (n, N_1, N_2) **deterministic code** with power constraint Γ for this channel is a

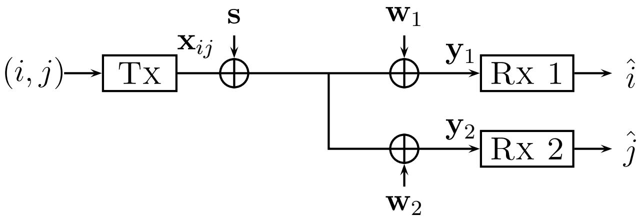


Figure 5.2: The arbitrarily varying degraded Gaussian broadcast channel. The jammer is shared between both receivers. Since we assume the noise $\mathbf{w}_2$ has higher variance that $\mathbf{w}_1$, we call receiver 1 the "strong" user and receiver 2 the "weak" user.

tuple of maps (ϕ, ψ_1, ψ_2), where

$$\phi : [N_1] \times [N_2] \to \mathbb{R}^n \tag{5.44}$$

$$\psi_1 : \mathbb{R}^n \to [N_1] \tag{5.45}$$

$$\psi_2 : \mathbb{R}^n \to [N_2] \ , \tag{5.46}$$

and $\|\phi(i,j)\|^2 \le n\Gamma$ for all (i,j). The map ϕ is the encoder and the maps ψ_1 and ψ_2 are the decoders for users 1 and 2. The average probability of error for the code under state constraint Λ is

$$\bar{\varepsilon} = \max_{\mathbf{s}:\|\mathbf{s}\|^2 \le n\Lambda} \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \mathbb{P}\left(\psi_1(\phi(i,j) + \mathbf{s} + \mathbf{w}_1) \ne i,\ \ \psi_2(\phi(i,j) + \mathbf{s} + \mathbf{w}_2) \ne j\right) \ . \tag{5.47}$$

The error is averaged over the messages to both users. We say the pair of rates (R_1, R_2) is achievable if there exists a sequence of $(n, \exp(nR_1), \exp(nR_2))$ deterministic codes whose average error goes to 0 as $n \to \infty$. The capacity region is the union of achievable rates.

The discrete arbitrarily varying broadcast channel without constraints was first studied by Jahn [89], who proved an achievable rate region for randomized coding and then applied the elimination technique [6] to derandomize the code. As we have seen, this approach does not work in general for constrained AVCs. Discrete constrained AVCs with degraded message sets were studied by Hof and Bross [82]. Their achievable strategy requires a number of non-symmetrizability conditions which are analogous to our result in Theorem 20. By using our earlier results on randomized coding, we obtain a much shorter proof.

We build a superposition code [40] based on our rotated codebook construction. The strong user can treat the message for the weak user as a random key in a randomized code. The codebook for user 2 is a deterministic code ("cloud centers") with power $\alpha\Gamma$ and the codebook for user 1 is a randomized code with power $(1-\alpha)\Gamma$, where the randomization is over the codewords of user 2. From Theorem 19 we can see that the randomization provided by user 2's message is sufficient for user 1 to achieve the randomized coding capacity. This scheme is limited by the result in [48] to those α for which $\alpha\Gamma > \Lambda$.

Theorem 20. *If $\Lambda \geq \Gamma$ then the deterministic coding capacity region of the arbitrarily varying degraded Gaussian broadcast channel is the empty set. If $\Lambda \leq \Gamma$ then for $\alpha \in (\Lambda/\Gamma, 1]$, the rates (R_1, R_2) satisfying the following inequalities are achievable with deterministic codes for the arbitrarily varying degraded Gaussian broadcast channel under average probability of error:*

$$R_1 < \frac{1}{2}\log\left(1 + \frac{(1-\alpha)\Gamma}{\Lambda + \sigma_1^2}\right) \tag{5.48}$$

$$R_2 < \frac{1}{2}\log\left(1 + \frac{\alpha\Gamma}{(1-\alpha)\Gamma + \Lambda + \sigma_2^2}\right) \tag{5.49}$$

$$R_1 + R_2 < \frac{1}{2}\log\left(1 + \frac{\Gamma - \Lambda}{\Lambda + \sigma_1^2}\right) + \frac{1}{2}\log\left(1 + \frac{\Lambda}{\Gamma + \sigma_2^2}\right) \ . \tag{5.50}$$

Proof. The converse follows from the converse for the standard AVC. Since we are limited to deterministic codes, if $\Lambda \geq \Gamma$ the jammer can choose a message pair (i', j')

and transmit $\phi(i', j')$ plus additional noise.

To show the achievable rate region, suppose that $\Lambda < \Gamma$ and generate a codebook using Theorem 19 containing $N_2 = \exp(nR_2)$ codewords $\{\mathbf{v}_j\}$ on the $\sqrt{n\alpha\Gamma}$-sphere. For each $\mathbf{u}_i$ generate $N_1 = \exp(nR_1)$ codewords $\{\mathbf{v}_{ij}\}$ uniformly on the $\sqrt{(n-1)(1-\alpha)\Gamma}$-sphere. Let the overall codebook be:

$$\mathbf{x}_{ij} = \mathbf{u}_i + A_i U_i \mathbf{v}_{ij} \ , \tag{5.51}$$

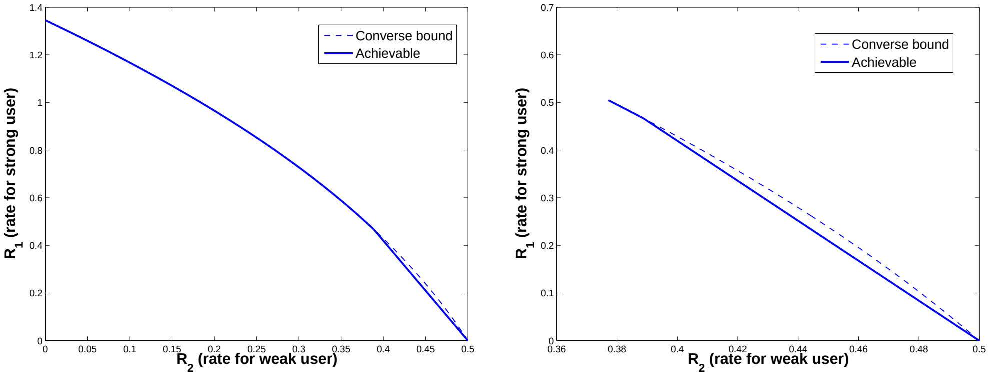
where A_i is an isometric mapping of $\mathbb{R}^{n-1}$ to the plane orthogonal to $\mathbf{u}_i$ and U_i is a random unitary transformation as in the construction of Theorem 19. The codebooks satisfy the the power constraint.

The weak decoder first decodes $\mathbf{u}_i$, treating the signal $A_i U_i \mathbf{v}_{ij}$ as additional noise. From the results of Csiszár and Narayan [48] we know that the average probability of error can be made small if $\alpha\Gamma > \Lambda$. This gives the first rate bound.

The strong decoder replicates the first step of the weak user. If message i was decoded correctly, it can subtract out $\mathbf{u}_i$ and the residual channel is identical to a GAVC with input power $(1-\alpha)\Gamma$ using the codebook of Theorem 19. This gives us the second rate bound.

To see the sum-rate bound (5.50), note that the weak user can give up any part of its message to the strong user, which means that rate splitting between the points where $\alpha = \Lambda/\Gamma$ and $\alpha = 0$ are also achievable. □

A plot of the achievable rate region is shown in Figure 5.3. This achievable region is tight for $\alpha > \Lambda/\Gamma$ because the jammer could just add Gaussian noise to make the channel a degraded Gaussian broadcast channel [23, 24, 25, 68]. The coding scheme above cannot be used in the regime where $\alpha \leq \Lambda/\Gamma$ because the jammer can symmetrize the $\{\mathbf{u}_j\}$ codebook to the stronger user. At present we do not know if new achievable strategies can achieve higher rates in this regime or if different converse



(a) Complete region

(b) Detail of gap

Figure 5.3: Achievable rates for the degraded broadcast Gaussian AVC with $\Gamma = 6$, $\Lambda = 1$, $\sigma_1^2 = 0.1$, and $\sigma_2^2 = 5$.

arguments can show that rate splitting is optimal.

5.3 Dirty paper coding

In this section we turn to a different AVC model in which there are two sources of interference, one of which is known to the transmitter. The benefits of channel state information at the transmitter have been investigated by researchers since Shannon [138]. In one version of the problem, a time-varying state sequence is known non-causally at the transmitter, and the encoder can base its codebook on this known sequence. The capacity for discrete channels with iid state sequences was found in the celebrated paper of Gel'fand and Pinsker [70]. Costa [39] showed an analogous result for the Gaussian case and showed that the capacity is equal to that of a channel with no interference at all. His strategy is called a "dirty paper code." These results have found applications to inter-symbol interference (ISI) channels [57], watermarking [38], multi-antenna broadcasting [155], and models for "cognitive radio" [50, 90]. The purpose of this section is to quantify the benefits of known interference in the context of arbitrarily varying channels.

For the GAVC, if sufficient common randomness is available to the encoder and decoder, the capacity is equal to that of an additive white Gaussian noise (AWGN) channel treating the jammer as independent noise. However, under deterministic coding, the capacity for these channels is zero when the jammer's power limit is greater than or equal to that of the encoder:

$$\overline{C}_d(\Gamma, \Lambda) = \begin{cases} 0 & \Gamma \le \Lambda \\ \frac{1}{2} \log\left(1 + \frac{\Gamma}{\Lambda + \sigma^2}\right) & \Gamma > \Lambda \ . \end{cases} \tag{5.52}$$

We will show that without common randomness, dirty paper coding can use a known interference signal to boost the effective power of the encoder and thus enlarge the region where the capacity is equal to the AWGN capacity.

We will consider channels with inputs and outputs in $\mathbb{R}^n$ of the form

$$\mathbf{Y} = \mathbf{X} + \mathbf{T} + \mathbf{S} + \mathbf{W} \ . \tag{5.53}$$

Here we take $\mathbf{W} \sim \mathcal{N}(0, \sigma_W^2 I)$, $\|\mathbf{S}\|^2 \le \Lambda n$, $\|\mathbf{X}\|^2 \le \Gamma n$, and $\mathbf{T} \sim \mathcal{N}(0, \sigma_t^2 I)$. The channel input created by the transmitter is $\mathbf{X}$, the vector $\mathbf{T}$ is interference known to the transmitter, $\mathbf{S}$ is jamming interference, and $\mathbf{W}$ is the independent noise at the receiver. If randomized coding is allowed, then Costa's result implies that the capacity is equal to the AWGN capacity without $\mathbf{T}$ and the jammer treated as additional noise.

In this section we will deal with deterministic coding. Again, if the jammer's power limit is sufficiently high it can simulate both the interference and the codeword, rendering the capacity 0 (Lemma 20). Our main result (Theorem 21) is an achievable rate region using a dirty-paper code in the spirit of [38] for this channel. For a range of the parameters $(\Gamma, \Lambda, \sigma_t^2)$ this scheme is capacity achieving (Corollary 5) – in this regime Costa's strategy cannot be defeated by adversarial interference. For

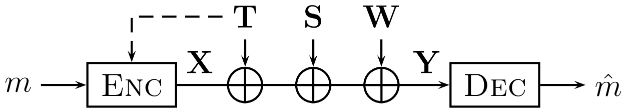


Figure 5.4: The Gaussian arbitrarily varying channel with a known interference signal at the encoder.

other sets of parameters we achieve positive rates but there is no matching converse. A discrete AVC model with the jammer input $\mathbf{S}$ known to the encoder was studied by Ahlswede [10], but his techniques do not apply directly to power-constrained continuous-alphabet channels.

One way of interpreting our results is as the deterministic coding analogue of the additive attack Gaussian watermarking game [38]. Under additive attacks of power Λ and watermark power Γ, the watermarking capacity under randomized coding is $(1/2)\log(1+\Gamma/\Lambda)$, regardless of the value of σ_t^2. The case $\sigma_t^2 = 0$ corresponds to randomized coding for the Gaussian AVC [86]. We can achieve the same rates without randomization as long as σ_t^2 is large enough. Another application of this result is to a particular model for spectrum-sharing systems in which a secondary system can code its message with knowledge of a primary system's message. For this cognitive radio model, a more robust assumption is that the interference seen by the secondary system is only partially known – the remaining interference may not be well-modeled by an iid noise process. In the AVC version of this channel we calculate achievable rates for the secondary system.

5.3.1 Channel and strategy

For simplicity, we will redefine inner products and norms to be normalized by the dimension, so $\|\mathbf{X}\|^2 = \dim(\mathbf{X})^{-1}\|\mathbf{X}\|_2^2$ and $\langle \mathbf{T},\ \mathbf{X}\rangle = \dim(\mathbf{X})^{-1}\mathbf{T}^T\mathbf{X}$.

An (n, N) code with power constraint Γ for this channel is a pair of functions (ϕ, ψ), where $\phi : [N] \times \mathbb{R}^n \to \mathbb{R}^n$ and $\psi : \mathbb{R}^n \to [N]$ and

$$\|\phi(i, \mathbf{T})\|^2 \leq \Gamma \qquad a.s. \ . \tag{5.54}$$

The average probability of error (over $\mathbf{W}$ and $\mathbf{T}$) for this code with jammer power Λ is given by

$$\overline{\varepsilon} = \max_{\mathbf{S}:\|\mathbf{S}\|^2 \leq \Lambda} \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}\left(\psi(\phi(i, \mathbf{T}) + \mathbf{T} + \mathbf{S} + \mathbf{W}) \neq i\right) \ . \tag{5.55}$$

A rate R is *achievable* if there exists a sequence of $(n, \lceil \exp(nR) \rceil)$ codes with $\overline{\varepsilon}_n \to 0$ as $n \to \infty$. The *capacity* $\overline{C}_d$ is defined to be the supremum of all achievable rates. For $\sigma_t^2 = 0$ this channel model reduces to the Gaussian AVC [48], whose capacity as we saw earlier exhibits the following dichotomy:

$$\overline{C}_d = \begin{cases} 0 & \Gamma \leq \Lambda \\ \frac{1}{2} \log\left(1 + \frac{\Gamma}{\Lambda + \sigma_W^2}\right) & \Gamma > \Lambda \ . \end{cases} \tag{5.56}$$

Our codebook construction uses two auxiliary rates R_U and R_{bin} and will depend on parameters α and ρ to be chosen later and positive constants ϵ_1 and ϵ_2 that can be made arbitrarily close to 0.

1. The encoder will generate an auxiliary codebook $\{\mathbf{U}_j\}$ of $\exp(n(R_U - \epsilon_1))$ vectors drawn uniformly from the n-sphere of power P_U, where

$$P_U = \Gamma + 2\rho\alpha\sqrt{\Gamma\sigma_t^2} + \alpha^2\sigma_t^2 \ . \tag{5.57}$$

2. These codewords are divided randomly into $\exp(n(R - 2\epsilon_1))$ bins $\{\mathcal{B}_m\}$ such that each bin has $\exp(n(R_{\text{bin}} + \epsilon_1))$ codewords. We denote the i-th codeword of

bin $\mathcal{B}_m$ by $\mathbf{U}(m,i)$.

3. Given a message m and an interference vector $\mathbf{T}$, the encoder chooses the vector $\mathbf{U}(m,i) \in \mathcal{B}_m$ that is closest to $\beta \mathbf{T}$, where

$$\beta^2 = \frac{P_U}{\sigma_t^2} \left(1 + \frac{(1-\rho^2)\Gamma}{P_U - (1-\rho^2)\Gamma} \right) \ . \tag{5.58}$$

If no such $\mathbf{U}(m,i)$ exists then we declare an encoder error. The encoder transmits

$$\mathbf{X} = \mathbf{U}(m,i) - \alpha \mathbf{T} \ . \tag{5.59}$$

We will show that for $\epsilon_2 > 0$, we can choose n sufficiently large so that $\|\mathbf{X}\|^2 \le \Gamma$ and

$$\langle \mathbf{U}(m,i) - \alpha \mathbf{T}, \ \mathbf{T} \rangle \ge \rho \sqrt{\Gamma \sigma_t^2} - \epsilon_2 \ . \tag{5.60}$$

4. The decoder first attempts to decode $\mathbf{U}(m,i)$ out of the overall codebook $\{\mathbf{U}_j\}$ and produces an estimate $\mathbf{U}(\hat{m},\hat{i})$. It then outputs the estimated message index $\hat{m}$.

Let $P_I = \Lambda + \sigma_W^2$ denote the expected power of the interference plus noise. Define

$$P_Y = \Gamma + 2\rho\sqrt{\Gamma \sigma_t^2} + \sigma_t^2 + \Lambda + \sigma_W^2 \ . \tag{5.61}$$

In Costa's original paper, choosing $\rho = 0$ and $\alpha = \alpha_0$, where

$$\alpha_0 = \frac{\Gamma}{\Gamma + \Lambda + \sigma_W^2} \tag{5.62}$$

gives an achievable rate of $(1/2)\log(1+\Gamma/(\Lambda+\sigma_W^2))$. We will analyze the performance

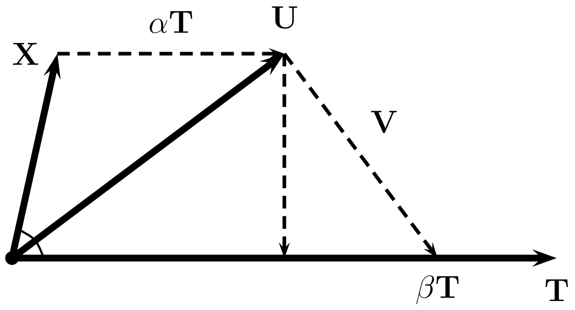


Figure 5.5: Geometric picture for dirty-paper encoding with general parameters.

of this coding strategy on an AVC with for general ρ and α.

5.3.2 Main result

Our main result is an achievable rate region for the Gaussian AVC with partial state information at the encoder that is achievable using this generalized dirty-paper code. For some parameter values the achievable rate is the capacity of the channel. One way of interpreting this result is that the presence of extra interference known to the transmitter boosts its effective power and therefore lowers the power threshold for the standard Gaussian AVC.

Our first lemma guarantees that the encoding is successful.

Lemma 19. *For any $\epsilon_1 > 0$ and $\epsilon_2 > 0$ we can choose sufficiently large n such that the probability that no $\mathbf{U}(m,i) \in \mathcal{B}_m$ exists satisfying equation (5.60) and $\|\mathbf{X}\|^2 \le \Gamma$ can be made as small as we like provided*

$$R_{\text{bin}} \ge \frac{1}{2}\log\left(\frac{P_U}{(1-\rho^2)\Gamma}\right) \ . \tag{5.63}$$

Proof. Consider the picture in Figure 5.5 and let

$$\beta^2 = \frac{P_U}{\sigma_t^2}\left(1 + \frac{(1-\rho^2)\Gamma}{P_U - (1-\rho^2)\Gamma}\right) . \tag{5.64}$$

We must show that a $\mathbf{U}(m,i) \in \mathcal{B}_m$ exists satisfying (5.60). By the rate-distortion theorem for Gaussian sources, for R_{bin} satisfying (5.63), the codebook $\mathcal{B}_m$ chosen uniformly on the sphere of power P_U can compress the source $\beta\mathbf{T}$ to distortion

$$D = \frac{P_U(1-\rho^2)\Gamma}{P_U - (1-\rho^2)\Gamma} . \tag{5.65}$$

To see this, consider the test channel $\beta\mathbf{T} = \mathbf{U} + \mathbf{V}$, where $\mathbf{V}$ is iid Gaussian with variance D. The mutual information of this test channel is

$$\frac{1}{n} I\left(\beta\mathbf{T} \ \wedge \ \mathbf{U}\right) = \frac{1}{2}\log\left(1 + \frac{P_U}{D}\right) \tag{5.66}$$

$$= \frac{1}{2}\log\left(\frac{P_U}{(1-\rho^2)\Gamma}\right) . \tag{5.67}$$

We can choose $\mathbf{U}(m,i)$ to be the codeword in $\mathcal{B}_m$ that corresponds to quantizing $\beta\mathbf{T}$. For any $\epsilon_1 > 0$ the codebook $\mathcal{B}_m$ has rate greater than the rate distortion function for the source $\beta\mathbf{T}$ with distortion D, so there exists an $\epsilon > 0$ such that with high probability,

$$\|\beta\mathbf{T} - \mathbf{U}(m,i)\|^2 \leq D - \epsilon . \tag{5.68}$$

For any $\delta > 0$ we can choose n sufficiently large such that with high probability we have

$$\langle \mathbf{U}, \ \mathbf{T} \rangle > \sqrt{\frac{P_U \sigma_t^2}{P_U + D - \epsilon}} - \delta . \tag{5.69}$$

If we choose δ small enough, we can find an $\eta > 0$ such that with high probability,

$$\langle \mathbf{U},\ \mathbf{T} \rangle > \sqrt{\frac{P_U \sigma_t^2}{P_U + D}} + \eta \tag{5.70}$$

$$= \sqrt{\sigma_t^2 (P_U - (1-\rho^2)\Gamma)} + \eta \tag{5.71}$$

$$= \rho\sqrt{\Gamma\sigma_t^2} + \alpha\sigma_t^2 + \eta\ . \tag{5.72}$$

Therefore with high probability we have

$$\langle \mathbf{U} - \alpha\mathbf{T},\ \mathbf{T} \rangle > \rho\sqrt{\Gamma\sigma_t^2} + \eta/2\ . \tag{5.73}$$

Therefore we can choose ϵ_2 to satisfy (5.60).

The last thing to check is that we can satisfy the encoder power constraint with high probability. Setting $\mathbf{X} = \mathbf{U} - \alpha\mathbf{T}$, we can see that for any $\delta' > 0$, with high probability

$$\|\mathbf{X}\|^2 \le P_U - 2\alpha \langle \mathbf{U},\ \mathbf{T} \rangle + \alpha^2\sigma_t^2 + \delta'\ . \tag{5.74}$$

Choosing δ' sufficiently small yields $\|\mathbf{X}\|^2 < \Gamma - \alpha\eta$, which proves the result. □

Theorem 21. *Let*

$$\mathcal{A}(\Lambda) = \left\{ (\alpha, \rho) : \frac{\left(\Gamma + (1+\alpha)\rho\sqrt{\Gamma\sigma_t^2} + \alpha\sigma_t^2\right)^2}{\Gamma + 2\rho\alpha\sqrt{\Gamma\sigma_t^2} + \alpha^2\sigma_t^2} > \Lambda \right\}\ . \tag{5.75}$$

The following rate is achievable:

$$R = \max_{(\alpha,\rho)\in\mathcal{A}(\Lambda)} \frac{1}{2} \log\left(\frac{(1-\rho^2)\Gamma P_Y}{(1-\alpha)^2(1-\rho^2)\Gamma\sigma_t^2 + P_I P_U} \right)\ . \tag{5.76}$$

Proof. We will choose the constants ϵ_1 and ϵ_2 according to Lemma 19. The decoder must decode $\mathbf{U}(m, i)$ from the received signal $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{U}(i) + (1-\alpha)\mathbf{T} + \mathbf{S} + \mathbf{W} \ . \tag{5.77}$$

The codebook $\{\mathbf{U}_k\}$ can be used to achieve any rate below the deterministic coding capacity of the GAVC with input $\mathbf{U}$, noise $\mathbf{W} + (1-\alpha)\mathbf{T}$, and jamming interference $\mathbf{S}$, provided $P_U > \Lambda$. We can therefore choose R_U to be equal to this capacity and for fixed α and ρ we calculate the capacity in what follows.

We first find the power of the component of $\mathbf{T}$ that is orthogonal to $\mathbf{U}$:

$$\mathbf{T} = \frac{\langle \mathbf{U},\ \mathbf{T} \rangle}{\|\mathbf{U}\|^2}\mathbf{U} + \left(\mathbf{T} - \frac{\langle \mathbf{U},\ \mathbf{T} \rangle}{\|\mathbf{U}\|^2}\mathbf{U} \right) \ . \tag{5.78}$$

From (5.60) we see that for any $\delta > 0$ we can choose n sufficiently large that

$$\mathbb{P}\left(\langle \mathbf{U},\ \mathbf{T} \rangle \geq \rho\sqrt{\Gamma\sigma_t^2} - \alpha\sigma_t^2 - 2\epsilon_2 \right) \geq 1 - \delta \ . \tag{5.79}$$

Let P_T be the expected power in the second term of (5.78). Then for sufficiently large n we also have

$$\mathbb{P}\left(P_T \leq \left(\sigma_t^2 - \frac{\left(\rho\sqrt{\Gamma\sigma_t^2} + \alpha\sigma_t^2 - 2\epsilon_2 \right)^2}{P_U} \right) \right) \geq 1 - \delta \ . \tag{5.80}$$

Some algebraic manipulation reveals that there is a constant c such that

$$\mathbb{P}\left(P_T - \frac{(1-\rho^2)\Gamma\sigma_t^2}{P_U} \leq c\epsilon_2 \right) \geq 1 - \delta \ . \tag{5.81}$$

In the GAVC (5.77) we define the equivalent noise variance as $P_I + (1-\alpha)^2 P_T$.

In order for $\mathbf{U}$ to be decodable, R_U must be smaller than the capacity of the corresponding AWGN channel:

$$R_U < \frac{1}{2} \log \left(\frac{P_U P_Y}{(1-\alpha)^2 (1-\rho^2) \Gamma \sigma_t^2 + P_I P_U} \right) . \tag{5.82}$$

Then $R_U - R_{\text{bin}}$ gives the term to be maximized in (5.76). Note that in the presence of a jammer with power constraint Λ, the $\mathbf{U}$ codebook is only capacity achieving if the received power in the $\mathbf{U}$ direction exceeds Λ. This received power is:

$$\gamma(\alpha, \rho) = \frac{\left(\Gamma + (1+\alpha)\rho\sqrt{\Gamma \sigma_t^2} + \alpha \sigma_t^2 \right)^2}{P_U} . \tag{5.83}$$

Thus for $(\alpha, \rho) \in \mathcal{A}(\Lambda)$ the the GAVC threshold for the $\mathbf{U}$ codebook can be met and $\mathbf{U}$ can be decoded. Lemma 19 shows that for large n the encoding will succeed, so the probability of error can be made as small as we like. □

For parameter values such that the point $(\alpha_0, 0) \in \mathcal{A}(\Lambda)$, the Costa rate is achievable. Since this rate corresponds to the jammer adding iid noise, the dirty-paper code is capacity achieving.

Corollary 5 (Capacity achieving parameters)**.** *If Γ, Λ and σ_t^2 are such that*

$$\Lambda < \frac{(\Gamma + \alpha_0 \sigma_t^2)^2}{\Gamma + \alpha_0^2 \sigma_t^2} , \tag{5.84}$$

then the capacity of the channel (5.53) under deterministic coding is

$$\overline{C}_d = \frac{1}{2} \log \left(1 + \frac{\Gamma}{\Lambda + \sigma_W^2} \right) , \tag{5.85}$$

and is achievable using the dirty paper code.

There is a threshold on Λ making the capacity equal to 0. If the jammer can simulate both the known interference and the transmitter's strategy, then it can symmetrize the channel.

Lemma 20 (Naïve outer bound)**.** *We have*

$$\overline{C}_d \leq \begin{cases} 0 & \Lambda > (\sigma_t + \sqrt{\Gamma})^2 \\ \frac{1}{2}\log\left(1 + \frac{\Gamma}{\Lambda + \sigma_W^2}\right) & \textit{otherwise} \ . \end{cases} \tag{5.86}$$

Proof. If $\Lambda > (\sigma_t + \sqrt{\Gamma})^2$, the jammer chooses a message m' uniformly in $[N]$ and creates a variable $\mathbf{T}'$ identically distributed to $\mathbf{T}$. It then mimics the encoder ϕ to create $\mathbf{X}' = \phi(m', \mathbf{T}')$ and sets $\mathbf{S} = \mathbf{T}' + \mathbf{X}'$. Since $\Lambda > (\sigma_t + \sqrt{\Gamma})^2$, this is a valid jamming strategy. The signal seen by the receiver is

$$\mathbf{Y} = (\phi(m, \mathbf{T}) + \mathbf{T}) + (\phi(m', \mathbf{T}') + \mathbf{T}') + \mathbf{W} \ . \tag{5.87}$$

Since the channel is symmetric, standard AVC arguments show that the capacity must be 0. □

Figure 5.6 shows an example of the achievable rate versus Γ. The two circles show the thresholds given by Corollary 5 and the threshold for the standard Gaussian AVC with deterministic coding and average error. The presence of the known interference $\mathbf{T}$ extends the capacity region relative to the standard AVC and achieves capacity for values of Γ that are smaller than the jammer constraint Λ. Thus far we have been unable to improve the converse for the region in which DPC does not achieve capacity; it may be that a different coding scheme utilizing the interference $\mathbf{T}$ can achieve higher rates in this regime.

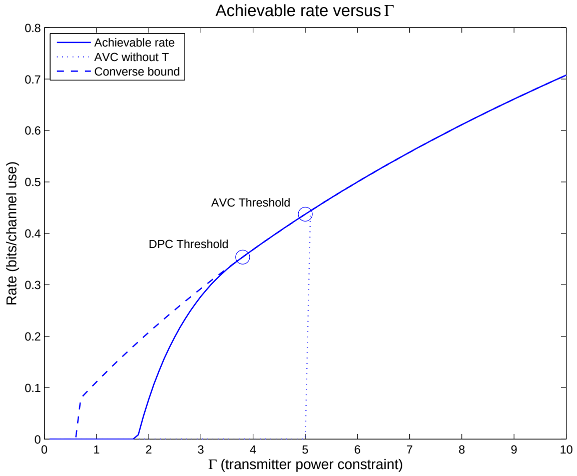


Figure 5.6: Rates versus Γ for $\Lambda = 5$, $\sigma_t^2 = 2$, and $\sigma_W^2 = 1$. The solid line is the achievable rate and the dashed line is the outer bound. The dotted line is the AVC capacity without the known interference signal. The threshold for the dotted line is at $\Gamma = \Lambda$ and the DPC threshold is given by (5.84).

5.3.3 Watermarking for additive attacks

One application in which DPC is useful is Gaussian watermarking [38]. In the watermarking problem, an encoder must encode a message m in a given *covertext* (e.g. an image) which is modeled by an iid Gaussian sequence $\mathbf{T}$. The encoder produces a *stegotext* $\mathbf{V} = \phi(m, \mathbf{T}) + \mathbf{T}$ that satisfies a distortion constraint $\|\mathbf{V} - \mathbf{T}\| = \|\phi(m, \mathbf{T})\|^2 \leq \Gamma$. In the standard watermarking game, $\mathbf{V}$ is then subjected to an attack that can depend on $\mathbf{V}$ to produce a compromised text $\mathbf{Y}$. The attack is also required to have limited distortion, so $\|\mathbf{Y} - \mathbf{V}\|^2 \leq \Lambda$. The decoder must recover the message m from $\mathbf{Y}$. The encoder and decoder are typically allowed to use (unlimited) common randomness. A modified dirty-paper code achieves the watermarking capacity, which is defined to be the maximum rate at which a watermark can be conveyed with vanishingly small probability of decoding error (averaged over messages).

A limited class of attacks are *additive attacks*, which take the form of an additive signal $\mathbf{S}$ that is independent of the stegotext $\mathbf{V}$. Under randomized coding, this kind of attacker can do no worse than add Gaussian noise to the stegotext, and the capacity is given by $(1/2)\log(1+\Gamma/\Lambda)$. We can use Theorem 21 with the noise $\mathbf{W}$ set to 0 to find achievable rates for this problem under deterministic coding; a decoder should be able to read the watermark without sharing a secret key with the encoder. Because the encoder does not want to distort the covertext by too much, an interesting regime for deterministic watermarking is when the power σ_t^2 of the covertext is much higher than the distortion limit Γ of the encoder. From (5.84) we can see that large σ_t^2 benefits the encoder by increasing the effective power of the auxiliary codebook to beat the jammer Λ.

By setting equality in (5.84), $\rho = 0$, and $\alpha = \alpha_0$, we can solve for σ_t^2 to get

$$\sigma_t^2 = \frac{1}{2}\Lambda \left(5 + 4\frac{\Lambda}{\Gamma} \right)^{1/2} - \Gamma - \frac{1}{2}\Lambda \ . \tag{5.88}$$

Let us set $\beta = \Lambda/\Gamma$ be the ratio of the attack distortion to the watermark distortion. We can rewrite the achievable σ_t^2 threshold as a function of β:

$$\sigma_t^2 = \Gamma \left(\frac{1}{2}\beta(5+4\beta)^{1/2} - \frac{1}{2}\beta - 1 \right) \ . \tag{5.89}$$

As we can see, the required σ_t^2 grows like $\beta^{3/2}$. Therefore for a fixed watermark distortion, the cover text variance must increase like $\Lambda^{3/2}$ in order to communicate at the randomized watermarking capacity.

An additive attack is a weak model for watermarking. It is more interesting to consider attackers which can directly manipulate the stegotext, as in the nosy noise error model. With deterministic coding, the jammer can read the watermark and then base its attack on that. This corresponds to an attack on the $\mathbf{U}$ codebook in the dirty-paper code, and appears to be a difficult sphere-packing problem. An easier

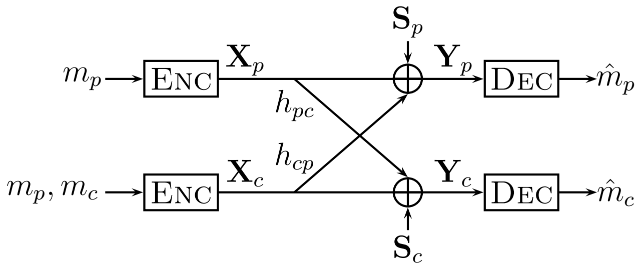


Figure 5.7: An AVC model for a cognitive radio system. The cognitive encoder can use the message m_p for the primary user to encode its own codeword.

problem may be to develop codes that use limited common randomness, as in Chapter 3, but we leave this for future work.

5.3.4 An AVC perspective on cognitive radio

A cognitive radio is a communication system that acts as a secondary user of licensed spectrum by agreeing to not impair the performance of the system used by a primary licensee. An interesting challenge posed by these new communications systems is to find theoretical models whose analysis reveals fundamental design tradeoffs. One model that has been proposed by Devroye, Mitran, and Tarokh [50] is the "cognitive interference channel" as shown in Figure 5.7. In this model, the primary and secondary (or cognitive) users can collaborate to communicate their messages to their respective decoders. More specifically, the cognitive transmitter is given non-causal knowledge of the primary user's message.

Let m_p and m_c denote the messages for the primary and cognitive systems. The primary encoder maps m_p into a codeword $\mathbf{X}_p$ of power Γ_p and the cognitive encoder maps m_c and m_p into a codeword $\mathbf{X}_c$ of power Γ_c. The received signals at the decoders are

$$\mathbf{Y}_p = \mathbf{X}_p + h_{cp}\mathbf{X}_c + \mathbf{S}_p \tag{5.90}$$

$$\mathbf{Y}_c = h_{pc}\mathbf{X}_p + \mathbf{X}_c + \mathbf{S}_c \ . \tag{5.91}$$

The gains h_{cp} and h_{pc} are assumed to be known to all parties. Devroye et al. model the noise signals $\mathbf{S}_p$ and $\mathbf{S}_c$ as Gaussian noise of variance Λ_p and Λ_c. The requirements are that the primary system should be agnostic to the actions of the cognitive system. This means that the rate

$$R_p = \frac{1}{2} \log \left(\frac{\Gamma_p + \Lambda_p}{\Lambda_p} \right) \tag{5.92}$$

should be achievable with the same primary encoder/decoder, regardless of the presence of the cognitive system.

The capacity under this model was found by Jovičić and Viswanath [90] when $\mathbf{X}_p$ is generated from a Gaussian codebook. The achievable scheme is based on spending part of the cognitive system's power to boost the primary's signal and using the remaining power in a dirty-paper code treating the primary's signal and boosting as known interference. For $\Lambda_c = 1$ the cognitive link achieves a rate

$$R_c = \frac{1}{2} \log \left(1 + (1-\mu)\Gamma_c\right) \ , \tag{5.93}$$

where μ is a constant that depends on the other parameters. More precisely, the cognitive encoder transmits $\mathbf{X}_c = \mathbf{V} + \sqrt{\mu\Gamma_c/\Gamma_p}\mathbf{X}_p$, where $\mathbf{V}$ is encoded using a dirty paper code treating $(h_{pc} + \sqrt{\mu\Gamma_c/\Gamma_p})\mathbf{X}_p$ as known interference.

We will focus on the cognitive system and instead treat $\mathbf{S}_c$ as arbitrarily varying interference of power Λ_c while leaving $\mathbf{S}_p$ as Gaussian noise. This corresponds to a model where the interference seen by the cognitive system can be broken into one component from a known interferer (the primary) and another from unknown interferers (possibly other cognitive systems). We can evaluate the performance of generalized dirty-paper coding for this cognitive radio model by modifying our main result. We choose the cognitive system's codeword to be

$$\mathbf{X}_c = \mathbf{V} + \eta \mathbf{X}_p \ , \tag{5.94}$$

where $\mathbf{V}$ is encoded using our the AVC dirty paper code. From the perspective of the cognitive system, we define $\mathbf{T} = (\eta + h_{pc})\mathbf{X}_p$ to get the channel

$$\mathbf{Y}_c = \mathbf{V} + \mathbf{T} + \mathbf{S}_c \ . \tag{5.95}$$

Here $\sigma_t^2 = (\eta + h_{pc})^2 \Gamma_p$. The parameters (η, α, ρ) must be chosen to satisfy the overall transmit power constraint $\|\mathbf{X}_c\|^2 \leq \Gamma_c$, the AVC threshold constraint, and the coexistence condition (5.92).

Before we can continue with the analysis, we must address one difference in this channel and the channel analyzed in Theorem 21. In Theorem 21 we could bound the average error of the scheme by noting that $\{\mathbf{U}_j\}$ is a capacity-achieving codebook for the GAVC under average error. The distribution on $\{\mathbf{U}_j\}$ induced by Gaussian interference $\mathbf{T}$ is uniform in expectation, so bounding the average error of $\{\mathbf{U}_j\}$ gives a bound on the error of the scheme. In the cognitive radio channel, the known interference $\mathbf{X}_p$ is not Gaussian. However, we will assume that $\mathbf{X}_p$ is drawn from a Gaussian codebook. Therefore the induced distribution on $\{\mathbf{U}_j\}$ converges to the uniform distribution and therefore the results will still hold.

First fix (η, α, ρ). Then the expected transmit power is

$$\Gamma_c = \|\mathbf{X}_c\|^2 = \|\mathbf{V}\|^2 + \eta^2 \Gamma_p + \frac{2\eta}{\eta + h_{pc}} \langle \mathbf{V}, \ \mathbf{T} \rangle \ . \tag{5.96}$$

Let $\Gamma_v = \|\mathbf{V}\|^2$. Then, using the definition of ρ in (5.60), we obtain the following quadratic equation in $\sqrt{\Gamma_v}$:

$$0 = \Gamma_v + 2\eta\rho\sqrt{\Gamma_p}\sqrt{\Gamma_v} + (\eta^2\Gamma_p - \Gamma_c) \ . \tag{5.97}$$

Solving, we find

$$\Gamma_v = \left(\left(\Gamma_c - \eta^2(1-\rho^2)\Gamma_p\right)^{1/2} - \eta\rho\sqrt{\Gamma_p} \right)^2 \ . \tag{5.98}$$

Thus we can use power Γ_v in the dirty paper code.

With Γ_v and σ_t^2 thus defined, the expected received power P_{Y_c} and auxiliary codebook power P_U are given by

$$P_{Y_c} = \Gamma_v + 2\rho\sqrt{\Gamma_v\sigma_t^2} + \sigma_t^2 + \Lambda_c \tag{5.99}$$

$$P_U = \Gamma_v + 2\alpha\rho\sqrt{\Gamma_v\sigma_t^2} + \alpha^2\sigma_t^2 \ . \tag{5.100}$$

Theorem 21 gives a condition for which a choice of (η, α, ρ) will be sufficient to overcome the AVC threshold $\Lambda_c = 1$:

$$\Lambda < \frac{(\Gamma_v + (1+\alpha)\rho\sqrt{\Gamma_v\sigma_t^2} + \alpha\sigma_t^2)^2}{P_U \ .} \ . \tag{5.101}$$

If $\rho \neq 0$, then the transmitted $\mathbf{V}$ contributes power to the primary received signal in the $\mathbf{X}_p$ direction. This affects how the coexistence condition (5.92) is satisfied. The received signal at the primary can be rewritten as

$$\mathbf{Y}_p = (1 + h_{cp}\eta)\mathbf{X}_p + h_{cp}\mathbf{V} + \mathbf{S}_p \ . \tag{5.102}$$

The projection of $\mathbf{V}$ in the direction of $\mathbf{X}_p$ has power $\rho^2\Gamma_v$, so the overall noise power with respect to $\mathbf{X}_p$ is $(1-\rho^2)h_{cp}^2\Gamma_v + \Lambda_p$. The total received power $P_{Y_p} = \|\mathbf{Y}_p\|^2$ is

$$\begin{aligned} P_{Y_p} = (1 + h_{cp}\eta)^2\Gamma_p + 2(1+h_{cp})h_{cp}\rho\sqrt{\Gamma_p\Gamma_v} \\ + h_{cp}^2\Gamma_v + \Lambda_p \ . \end{aligned} \tag{5.103}$$

Thus the coexistence constraint can be expressed as:

$$\frac{\Gamma_p + \Lambda_p}{\Lambda_p} = \frac{P_{Y_p}}{(1-\rho^2)h_{cp}^2\Gamma_v + \Lambda_p} \ . \tag{5.104}$$

We can express this achievable rate region in the following corollary:

Corollary 6. *Let Γ_v be given by (5.98), $\sigma_t^2 = (\eta + h_{pc})^2\Gamma_p$, and*

$$\mathcal{A} = \{(\eta, \alpha, \rho) : (5.101), (5.104) \ \textit{satisfied}\} \ . \tag{5.105}$$

Then the following rate is achievable for the cognitive system:

$$R_c = \max_{(\eta,\alpha,\rho)\in\mathcal{A}} \frac{1}{2}\log\left(\frac{(1-\rho^2)\Gamma_v P_{Y_c}}{(1-\alpha)^2(1-\rho^2)\Gamma\sigma_t^2 + \Lambda_c P_U}\right) \ . \tag{5.106}$$

Figure 5.8 and 5.9 give two plots of the achievable rates as a function of the gain h_{pc} from the primary transmitter to the cognitive receiver. Figure 5.8 corresponds to small h_{cp}, which means the cognitive transmitter is far from the primary receiver. In this case, the effect of spending power to aid the primary transmission hampers the cognitive system and it requires larger h_{pc} to overcome the jammer. Figure 5.9 is for larger h_{cp}, which means the cognitive transmitter is closer to the primary receiver. Here the cognitive system achieves lower rates but reaches capacity faster as a function of h_{pc}. That is, stronger known interference from the primary system can helps weak cognitive transmitter reach capacity in the presence of jamming interference.

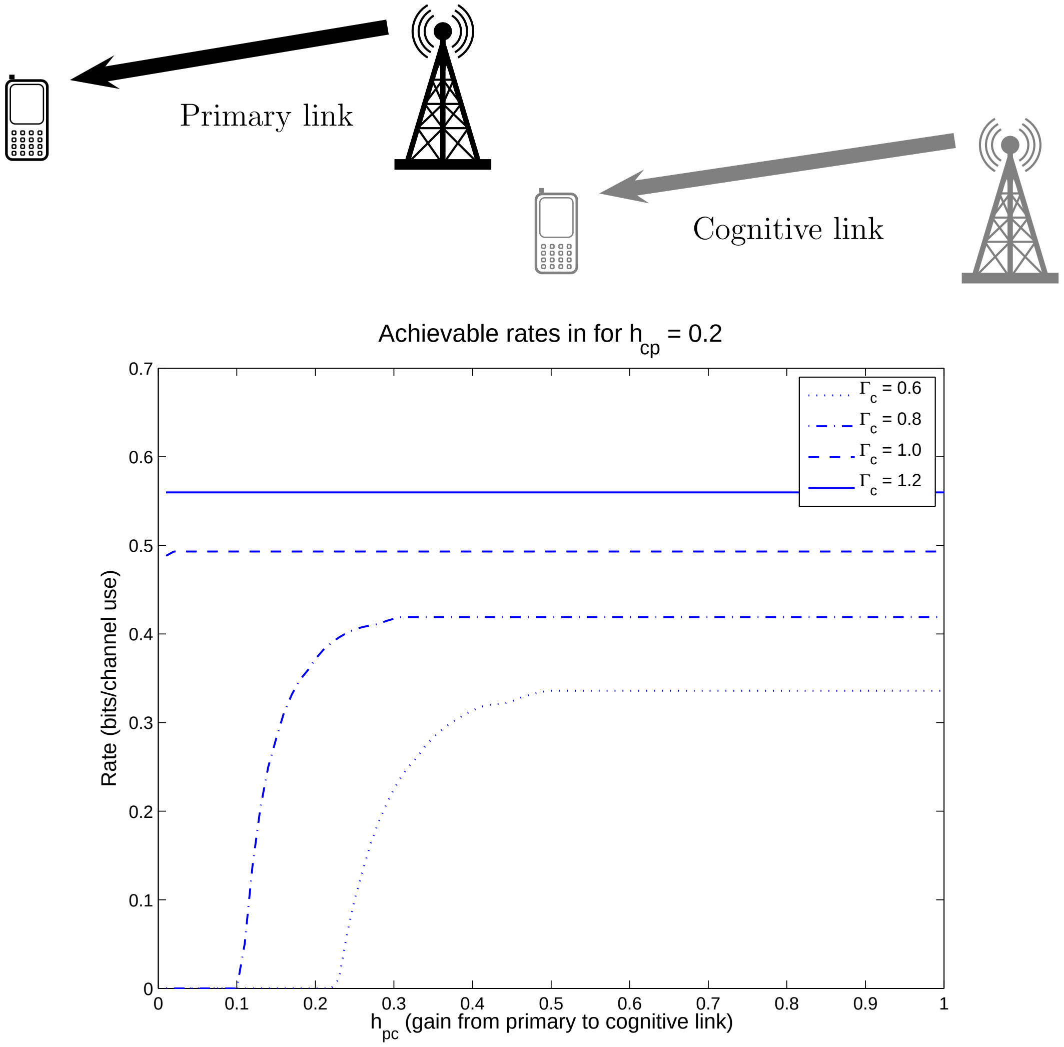


Figure 5.8: Plot of achievable rates versus varying gain from the primary transmitter to the cognitive receiver when the primary receiver is far from the cognitive transmitter (small h_{cp}). The flat region corresponds to the capacity assuming $\mathbf{S}_c$ is noise.

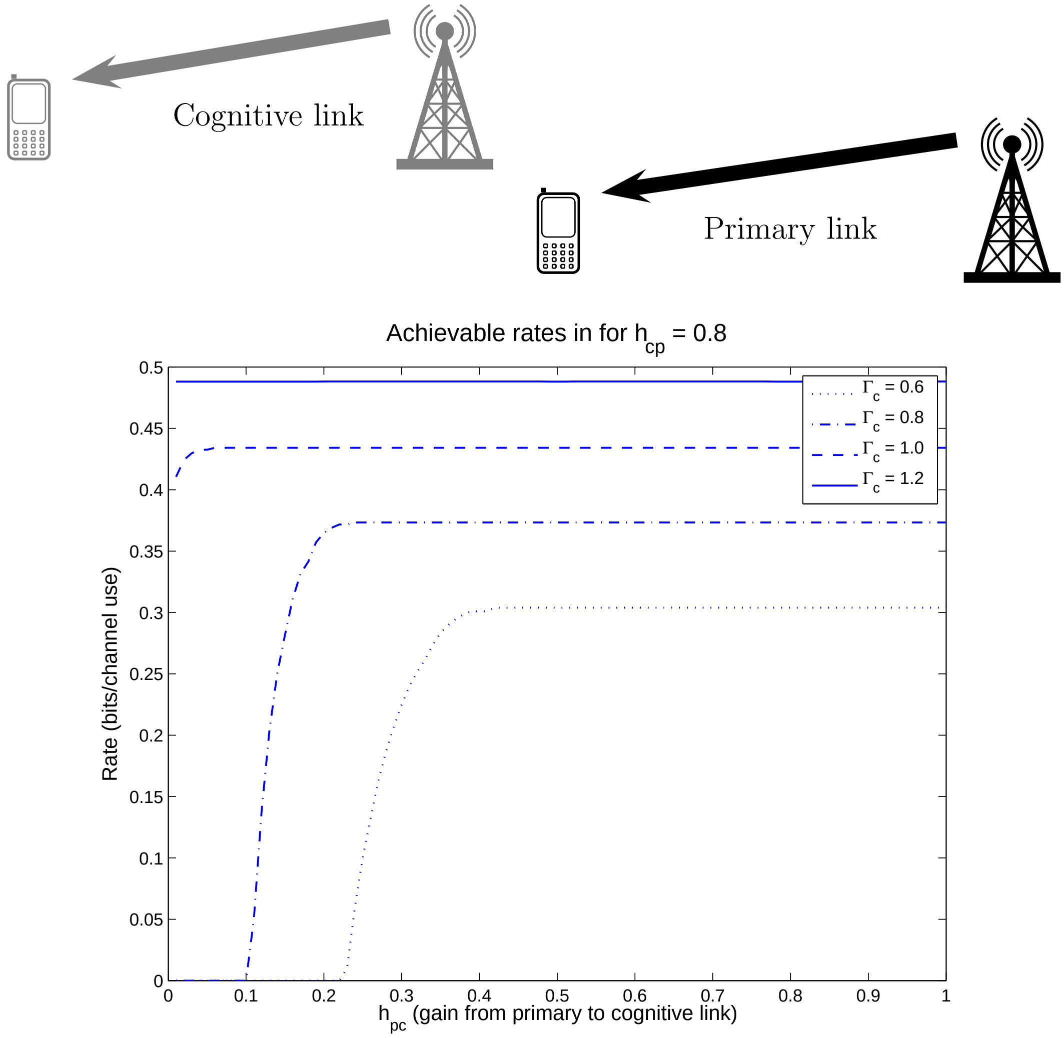


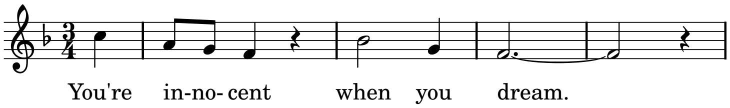
Figure 5.9: Plot of achievable rates versus varying gain from the primary transmitter to the cognitive receiver when the primary receiver is close to the cognitive transmitter (large h_{cp}). The flat region corresponds to the capacity assuming $\mathbf{S}_c$ is noise.

5.4 Conclusions

In this chapter we showed that a modest amount of randomization is sufficient to achieve the randomized coding capacity of the GAVC and saw an example of how to use the randomization for broadcast channels. If there exists a source of common randomness [16,17] R' shared between the encoder and decoder, they can generate a key rate of R', which is sufficient for an exponential error decay. However, our results show that a more modest key size is sufficient to achieve capacity.

Because the required key size is small, we could instead consider a case where the encoder and decoder must generate it from a correlated source. Suppose instead that a sequence of random variables $\{U_i : i = 1, 2, \ldots n\}$ and $\{V_i : i = 1, 2, \ldots, n\}$ are given to the encoder and decoder respectively. For discrete AVCs, it was shown by Ahlswede and Cai [14] that even if the common randomness $CR(U, V) = 0$, as long as the mutual information $I\left(U \ \wedge \ V\right) > 0$ the deterministic coding capacity under average error is always equal to the randomized coding capacity. An open question is whether this remains true for the Gaussian AVC.

We also investigated how a known interference signal at the transmitter can help a transmitter overcome arbitrarily varying interference. By using a dirty paper code, the transmitter uses some of the interference power to augment its own effective power. This leads into an improved threshold for the capacity region of Gaussian AVCs. Our theorem shows that deterministic watermarking codes can achieve the randomized coding capacity if the covertext variance is sufficiently large. We also found achievable rates for an AVC model for cognitive radios in which a cognitive system may cooperate with a primary system but is subject to unknown interference from other signals.



– *Innocent When You Dream*, Tom Waits

Chapter 6

Looking Ahead

As architectures for communication systems transition from centralized to decentralized designs, new challenges will arise in the design of robust and adaptive protocols. Platforms such as sensor networks, ad-hoc networks, and cognitive radio require communication schemes to work in a variety of channel conditions that may not be modeled by the stationary channel models most often studied in information theory. The arbitrarily varying channel is an alternative model that makes very little assumptions on the channel dynamics.

Assuming an adversarial model for the channel state selection may not be philosophically or phenomenologically satisfying for "real-world" communication systems outside of military communications. The AVC model can be relaxed in several different ways that make the jammer seem less adversarial. Allowing randomization, list decoding, feedback, or moving to the average error criterion are examples of these relaxations. In this thesis we saw how extra resources can help make point-to-point communication over AVCs more robust and adaptive.

For the capacity of the AVC under nosy noise, we saw that list decoding plays an important role in the achievable scheme using randomization. The recent results on list decoding for Reed-Solomon codes and variants [79, 78, 119] mean that it may be possible to design practical randomized codes that achieve the capacity for some AVC models. It would be extremely interesting to find list decoding algorithms for other ensembles of codes, such as low-density parity check (LDPC) codes.

The bulk of this thesis focused on discrete AVCs, which had been studied fairly extensively in the research literature. The Gaussian AVC has been less studied and there are still many interesting aspects to consider in that setting in terms of robustness and adaptivity. For example, multiple antenna (MIMO) systems are an important first step in understanding distributed collaboration schemes. In the Gaussian AVC context we may model the jammer as another system with a different number of antennas than the transmitter and receiver. Preliminary work on this problem suggests that the behavior may be quite interesting [131].

There has been a spate of recent work in the information theory literature on cooperative communication protocols (see for example the special issue of the *IEEE Transactions on Information Theory* [98], the recent monograph by Kramer, Marić, and Yates [99], and references therein). This research deals with quantifying the potential gains from user cooperation in a communication system and how how to realize these gains. In contrast, there is an ever-growing literature in networking on congestion control and resource allocation via market mechanisms such as pricing [93, 94], auctions [83], and game theory. These models assume users do not cooperate and in fact are competing for the same resources in the network.

One difference between these two views is granularity – the users in the network context may be groups of cooperating users in the physical layer context. In the multiple access case, this is similar to the model studied by La and Anantharam [100]. Another model may be via coalitional game theory [110]. Arbitrarily varying channels may be one way to merge these two perspectives – a group of cooperating users can share joint randomization and other resources to compete against other users. The channels seen by each group are AVCs treating the other groups as jammers. Coding strategies for more general AVCs can therefore be seen as a way of insulating groups of cooperating users from the vagaries of other groups. As wireless systems become more complex, this isolation may be an important tool for reducing the complexity of system design.

Appendix A

Notation

Typeface conventions

Style	Example	Description
calligraphic	$\mathcal{X}$, $\mathcal{S}$, etc.	alphabets for random variables
boldface	$\mathbf{x}$, $\mathbf{s}$, etc.	vector $(x_1, x_2, \ldots, x_n)$

Sets, types

Symbol	Description	Reference
$[M]$	the set $\{1, 2, \ldots, M\}$	Page 7
$\mathcal{P}(\mathcal{X})$	probability distributions on $\mathcal{X}$	Page 7
$\mathcal{P}(\mathcal{Y}\|\mathcal{X})$	conditional distributions on $\mathcal{Y}$ given $\mathcal{X}$	Page 7
$N(x\|\mathbf{x})$	the number of times x appears in $\mathbf{x}$	Page 7
$T_{\mathbf{x}}$	the type of $\mathbf{x}$	(1.1)
$\mathcal{T}_n(P)$	set of all length-n sequences of type P	(1.2)
T_P^ϵ	ϵ-typical P-sequences	(1.8)
$T_V^\epsilon(\mathbf{x})$	(V, ϵ)-shell around $\mathbf{x}$	(1.9)

Information measures and metrics

Symbol	Description	Reference
$d_{\max}(P, Q)$	ℓ_∞ distance between distributions P and Q	(1.7)
$H(X)$	entropy of a random variable X	[41, 44]
$H(P)$	entropy of the distribution P	[41, 44]
$H(Y\|X)$	conditional entropy of Y given X	[41, 44]
$H(V\|P)$	$H(Y\|X)$ under the distribution $V(y\|x)P(x)$	[41, 44]
$I(X \wedge Y)$	mutual information between X and Y	[41, 44]
$I(P, V)$	mutual information under the distribution $V(y\|x)P(x)$	(1.26)
$h_b(\cdot)$	binary entropy function	Page 22

Arbitrarily varying channel parameters

Symbol	Description	Reference
$\mathcal{W}$	an AVC $\{W(\cdot\|\cdot, s) : s \in \mathcal{S}\}$	Page 9
$g(\cdot)$	input cost function	Page 9
γ^*	upper bound on $g(\cdot)$	Page 9
Γ	input cost bound	(1.13)
$l(\cdot)$	state cost function	Page 9
λ^*	upper bound on $l(\cdot)$	Page 9
Λ	state cost bound	(1.14)

Sets and functions for the AVC

Symbol	Description	Reference
$\mathcal{S}^n(\Lambda)$	set of sequences with average cost $\leq \Lambda$	(1.15)
$\mathcal{I}(\Gamma)$	input distributions satisfying cost constraint	(1.27)
$\mathcal{Q}(\Lambda)$	state distributions satisfying cost constraint	(1.28)
$\mathcal{U}(P, \Lambda)$	nosy noise jammer channels satisfying cost constraint	(1.29)
$\mathcal{W}_{std}(\Lambda)$	convex closure of the AVC	(1.30)
$\mathcal{W}_{dep}(P, \Lambda)$	row-convex closure of the AVC with input P	(1.31)
$\mathcal{U}_{\mathrm{sym}}(m)$	set of channels $\mathcal{X}^m \to \mathcal{S}$ that symmetrize the AVC	(1.37)
$C_{\mathrm{std}}(\Gamma, \Lambda)$	max-min under iid state selection	(1.38)
$C_{\mathrm{dep}}(\Gamma, \Lambda)$	max-min under input-dependent state selection	(1.39)
$E_r(R, P, \Lambda)$	error exponent for AVCs under maximal error	(1.75)

Coding and errors

Symbol	Description	Reference
(ϕ, ψ)	deterministic encoder and decoder	(1.20)
(Φ, Ψ)	randomized encoder and decoder	Page 13
D_i	set of outputs which decode to message i	Page 12
$\varepsilon(i, \mathbf{s})$	probability of error for message i given state $\mathbf{s}$	(1.21)
ε	maximal error	(1.22)
$\bar{\varepsilon}$	average error	(1.23)
$\hat{\varepsilon}$	error under nosy noise	(1.25)
ε_L	error for list decoding under maximal error	(1.35)
$\bar{\varepsilon}_L$	error for list decoding under average error	(1.36)

Capacities

Symbol	Description
$\hat{C}_d$	capacity for deterministic coding and nosy noise error
C_d	capacity for deterministic coding and maximal error
$\overline{C}_d$	capacity for deterministic coding and average error
$\hat{C}_r$	capacity for randomized coding and nosy noise error
C_r	capacity for randomized coding and maximal error
$\overline{C}_r$	capacity for randomized coding and average error
C_L	capacity under list decoding with list size L and maximal error
$\overline{C}_L$	capacity under list decoding with list size L and average error

List decoding

Symbol	Description	Reference
L_{sym}	unconstrained symmetrizability	Page 19
$\lambda_m(P)$	strong symmetrizing cost	(2.51)
$L_{\mathrm{sym}}(P, \Lambda)$	strong symmetrizability	(2.52)
$\tilde{\lambda}_m(P)$	weak symmetrizing cost	(2.53)
$\tilde{L}_{\mathrm{sym}}(P, \Lambda)$	weak symmetrizability	(2.54)
$\mathcal{G}_\eta(\Lambda)$	allowable joint types for average-error decoding	(2.115)

Rateless codes

Symbol	Description	Reference
$I\left(P,\mathcal{V}\right)$	minimum mutual information $\min_{V\in\mathcal{V}} I\left(P,V\right)$	(4.5)
c	chunk size for rateless code	Page 128
$\mathcal{V}_m$	state information set after m chunks	Page 128
$\mathbf{V}(c)$	possible values for $\mathcal{V}_m$	Page 128
λ_M	average state cost over chunk m	(4.16)
$\hat{\lambda}_m$	side information about cost at chunk m	(4.17)
Λ_M	average state cost up to chunk m	(4.22)
$\hat{\Lambda}_M$	average cost estimate up to chunk M	(4.23)
$\tilde{\Lambda}_M$	cost used by codebook after chunk M	(4.32)
v	polynomial exponent for size of $\mathbf{V}(c)$	Page 130
$\mathbf{x}^{(mc)}$, $\mathbf{y}^{(mc)}$, $\mathbf{s}^{(mc)}$	m-th chunk of vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{s}$	Page 128.
$\mathbf{x}_1^r$, $\mathbf{y}_1^r$, $\mathbf{s}_1^r$	first r elements of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{s}$	Page 128
Φ_m	rateless code encoder for m-th chunk	(4.6)
τ_m	rateless code feedback function for m-th chunk	(4.7)
Ψ_m	rateless code decoder for m-th chunk	(4.8)
$\mathbf{M}$	decoding time for rateless code	(4.9)
$\mathcal{M}$	possible values of the decoding time $\mathbf{M}$	Page 129
M_*	minimum value of $\mathbf{M}$	Page 129
M^*	maximum value of $\mathbf{M}$	Page 129
$\varepsilon(M,\mathbf{s})$	rateless coding maximal error	(4.11)
$\hat{\varepsilon}(M,J)$	rateless coding error under nosy noise	(4.12)

Appendix B

Some simple inequalities

B.1 Continuity bounds for entropies

We need a short technical lemma about concave functions.

Lemma 21. *Let f be a concave increasing function on $[a,b]$. Then if $a \leq x < x+\epsilon \leq b$, we have*

$$f(x+\epsilon) - f(x) \leq f(a+\epsilon) - f(a) \ . \tag{B.1}$$

Proof. Without loss of generality we can take $a = 0$, $b = 1$, and $f(a) = 0$. Now consider

$$\begin{aligned} f(x) = f\left(\frac{x}{x+\epsilon} \cdot (x+\epsilon) + \frac{\epsilon}{x+\epsilon} \cdot 0\right) &\geq \frac{x}{x+\epsilon} f(x+\epsilon) + \frac{\epsilon}{x+\epsilon} f(0) \\ &= \frac{x}{x+\epsilon} f(x+\epsilon) \\ f(\epsilon) = f\left(\frac{x}{x+\epsilon} \cdot 0 + \frac{\epsilon}{x+\epsilon} \cdot (x+\epsilon)\right) &\geq \frac{x}{x+\epsilon} f(0) + \frac{\epsilon}{x+\epsilon} f(x+\epsilon) \\ &= \frac{\epsilon}{x+\epsilon} f(x+\epsilon) \ . \end{aligned}$$

Therefore

$$f(x) + f(\epsilon) \geq f(x + \epsilon) \ , \tag{B.2}$$

as desired. □

Using the preceding lemma, we can show that a bound on the total variational distance between two distributions gives a bound on the entropy between those two distributions.

Lemma 22. *Let P and Q be two distributions on a finite set $\mathcal{S}$ with $|\mathcal{S}| \geq 2$. If*

$$|P(s) - Q(s)| \leq \epsilon \qquad \forall s \in \mathcal{S} \ , \tag{B.3}$$

then

$$|H(P) - H(Q)| \leq (|\mathcal{S}| - 1) \cdot h_b(\epsilon) + (|\mathcal{S}| - 1) \log(|\mathcal{S}| - 1) \cdot \epsilon \ , \tag{B.4}$$

where $h_b(\cdot)$ is the binary entropy function.

Proof. Let $\mathcal{S} = \{s_1, s_2, \ldots\}$. We proceed by induction on $|\mathcal{S}|$. Suppose $|\mathcal{S}| = 2$, and let $p = P(s_1)$ and $q = Q(s_1)$. The entropy function $h_b(x)$ is concave, increasing on $[0, 1/2]$ and decreasing on $[1/2, 1]$. Applying Lemma 21 to each interval, we obtain the bound:

$$|h_b(x + \epsilon) - h_b(x)| \leq h_b(\epsilon) \ . \tag{B.5}$$

Since $H(P) = h_b(p)$ and $H(Q) = h_b(q)$, this proves our result.

Now suppose that the lemma holds for $|\mathcal{S}| \leq m-1$, and consider the case $|\mathcal{S}| = m$. Without loss of generality, let $P(s_m) > 0$ and $Q(s_m) > 0$. Let $\lambda = (1-P(s_m))$ and $\mu =$

$(1-Q(s_m))$ and note that $|\lambda-\mu|<\epsilon$ by assumption. Define the $(m-1)$ dimensional distributions $P' = \lambda^{-1}(P(s_1),\ldots,P(s_{m-1}))$ and $Q' = \lambda^{-1}(Q(s_1),\ldots,Q(s_{m-1}))$, so that

$$\begin{aligned} P &= (\lambda P', (1-\lambda)) \\ Q &= (\mu Q', (1-\mu)) \ . \end{aligned}$$

Therefore,

$$\begin{aligned} H(P) &= h_b(\lambda) + \lambda H(P') \\ H(Q) &= h_b(\mu) + \mu H(Q') \ . \end{aligned}$$

Now we we can expand the difference of the entropies, using the fact that $\lambda < 1$, the induction hypothesis on $|H(P')-H(Q')|$ and $|h_b(\lambda)-h_b(\mu)|$, and the cardinality bound on the entropy $H(Q')$ to obtain

$$\begin{aligned} |H(P)-H(Q)| &= |\lambda H(P') - \mu H(Q') + h_b(\lambda) - h_b(\mu)| \\ &\leq \lambda |H(P') - H(Q')| + |\lambda - \mu| H(Q') + |h_b(\lambda) - h_b(\mu)| \\ &\leq (m-2)\cdot h_b(\epsilon) + (m-2)\log(m-2)\cdot\epsilon + \log(m-1)\cdot\epsilon + h_b(\epsilon) \\ &\leq (m-1)\cdot h_b(\epsilon) + (m-1)\log(m-1)\cdot\epsilon \ . \end{aligned}$$

□

Lemma 23. *Let $W(y|x)$ and $V(y|x)$ be two channels with finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$. If*

$$|W(y|x) - V(y|x)| \leq \epsilon \qquad \forall (x,y) \in \mathcal{X}\times\mathcal{Y} \ , \tag{B.6}$$

then for any input distribution P on $\mathcal{X}$ we have

$$|I(P,W) - I(P,V)| \leq 2(|\mathcal{Y}|-1) \cdot h_b(\epsilon) + 2(|\mathcal{Y}|-1)\log(|\mathcal{Y}|-1) \cdot \epsilon \ , \tag{B.7}$$

where $h_b(\cdot)$ is the binary entropy function.

Proof. We simply apply Lemma 22 twice. Let Q_W and Q_V be the marginal distributions on $\mathcal{Y}$ under channels W and V respectively. Then

$$|Q_W(y) - Q_V(y)| \leq \sum_x P(x)|W(y|x) - V(y|x)| \leq \epsilon \ .$$

Now we can break apart the mutual information and use Lemma 22 on each term:

$$\begin{aligned} |I(P,W) - I(P,V)| &\leq |H(Q_W) - H(Q_V)| \\ &\qquad + \sum_x P(x)|H(W(Y|X=x)) - H(V(Y|X=x))| \\ &\leq 2(|\mathcal{Y}|-1) \cdot h_b(\epsilon) + 2(|\mathcal{Y}|-1)\log(|\mathcal{Y}|-1) \cdot \epsilon \ . \end{aligned}$$

□

B.2 Properties of concatenated fixed composition sets

Let $\tau(\mathbf{x})$ be the type of $\mathbf{x}$. Let $\mathbf{T}_n(P) = \{\mathbf{x} \in \mathcal{X}^n : \tau(\mathbf{x}) = P\}$ be the set of of all length-n vectors of type P. For a vector $\mathbf{x}$, let $\mathbf{x}_1^m$ be the first m elements of $\mathbf{x}$.

Lemma 24. *For all finite sets $\mathcal{X}$, and all types P with $p_0 = \min_{x \in \mathcal{X}} P(x) > 0$, there exists $\eta = \eta(P) < \infty$ such that for all integers $M, n > 0$,*

$$\frac{|\mathcal{T}_n(P)|^M}{|\mathcal{T}_{mn}(P)|} \geq \exp(-\eta M \log(n+1)) \ . \tag{B.8}$$

Proof. We begin by expanding the ratio:

$$\frac{|\mathcal{T}_n(P)|^M}{|\mathcal{T}_{mn}(P)|} = \frac{\binom{n}{p_1 n,\ p_2 n,\ \ldots,\ p_{|\mathcal{X}|} n}^M}{\binom{Mn}{p_1 Mn,\ p_2 Mn,\ \ldots,\ p_{|\mathcal{X}|} Mn}} \ .$$

We can bound the multinomial coefficient using Stirling's approximation [65, pp. 50–53] :

$$\begin{aligned}
&\binom{n}{p_1 n,\ p_2 n,\ \ldots,\ p_{|\mathcal{X}|} n} \\
&\qquad = \frac{n!}{(p_1 n)! \cdot (p_2 n)! \cdots (p_{|\mathcal{X}|} n)!} \\
&\qquad \geq (\sqrt{2\pi})^{-|\mathcal{X}|+1} \cdot \frac{n^n \sqrt{n}}{\prod_{x=1}^{|\mathcal{X}|} (p_x n)^{p_x n} \sqrt{p_x n}} \cdot \exp\left(\frac{1}{12n+1} - \sum_{x=1}^{|\mathcal{X}|} \frac{1}{12 p_x n} \right) ,
\end{aligned}$$

and

$$\begin{aligned}
&\binom{Mn}{p_1 Mn,\ p_2 Mn,\ \ldots,\ p_{|\mathcal{X}|} Mn} \\
&\quad = \frac{(Mn)!}{(p_1 Mn)! \cdot (p_2 Mn)! \cdots (p_{|\mathcal{X}|} Mn)!} \\
&\quad \leq (\sqrt{2\pi})^{-|\mathcal{X}|+1} \cdot \frac{(Mn)^{Mn} \sqrt{Mn}}{\prod_{x=1}^{|\mathcal{X}|} (p_x Mn)^{p_x Mn} \sqrt{p_x Mn}} \cdot \exp\left(-\frac{1}{12Mn+1} + \sum_{x=1}^{|\mathcal{X}|} \frac{1}{12 p_x Mn} \right) \ .
\end{aligned}$$

Now we can cancel some terms to get a further lower bound for some $0 < \nu(P) < \infty$:

$$\begin{aligned}
&\frac{|\mathcal{T}_n(P)|^M}{|\mathcal{T}_{mn}(P)|} \\
&\quad \geq (\sqrt{2\pi})^{-(M-1)(|\mathcal{X}|-1)} \cdot \frac{n^{Mn}}{(Mn)^{Mn}} \cdot \left(\prod_{x=1}^{|\mathcal{X}|} \frac{(p_x Mn)^{p_x Mn}}{(p_x n)^{p_x Mn}} \right) \\
&\qquad\qquad \cdot \left(\frac{(Mn)^{(|\mathcal{X}|-1)}}{n^{M(|\mathcal{X}|-1)}} \cdot \prod_{x=1}^{|\mathcal{X}|} p_x^{-(M-1)} \right)^{1/2} \\
&\qquad\qquad \cdot \exp\left(\frac{M}{12n+1} - \sum_{x=1}^{|\mathcal{X}|} \frac{M}{12 p_x n} + \frac{1}{12Mn+1} - \sum_{x=1}^{|\mathcal{X}|} \frac{1}{12 p_x Mn} \right) \\
&\quad \geq \exp(-M|\mathcal{X}| \log \sqrt{2\pi}) \cdot \exp\left(\frac{1}{2}(|\mathcal{X}|-1)(\log Mn - M \log n) \right) \cdot \exp\left(-\nu(P)M/n\right) \\
&\quad \geq \exp\left(-M \left(|\mathcal{X}| \log \sqrt{2\pi} + \frac{1}{2}(|\mathcal{X}|-1) \log n - \frac{(|\mathcal{X}|-1)\log Mn}{2M} + \frac{\nu(P)}{n} \right) \right) \\
&\quad \geq \exp(-\eta M \log(n+1)) \ ,
\end{aligned}$$

where $\eta = \eta(P) < \infty$. □

Appendix C

Computations for examples

This appendix contains some of the computations used in the examples provided in the thesis.

C.1 $C_{\text{dep}}(\Gamma, \Lambda)$ for binary real additive channels

In Example 1.5 on page 22, we claimed the that $C_{\text{dep}}(\Gamma, \Lambda)$ of the binary-input, binary-state real additive channel was given by

$$C_{\text{dep}}(\Gamma, \Lambda) = \begin{cases} h_b\left(\frac{1-\Lambda}{2}\right) - \frac{1+\Lambda}{2} h_b\left(\frac{2\Lambda}{1+\Lambda}\right) & \Gamma \geq \frac{1-\Lambda}{2} \\ h_b(\Gamma) - (\Lambda + \Gamma) h_b\left(\frac{\Lambda}{\Lambda+\Gamma}\right) & \Gamma < \frac{1-\Lambda}{2} \end{cases} \tag{C.1}$$

To show this we must compute the max-min expression

$$C_{\text{dep}}(\Gamma, \Lambda) = \max_{P \in \mathcal{I}(\Gamma)} \min_{V \in \mathcal{W}_{dep}(P,\Lambda)} I\left(P, V\right) \ . \tag{C.2}$$

Intuitively we expect that the optimal jamming strategy is to set $s = 0$ when $x = 1$ and $s = 1$ when $x = 0$ in order to make the output equal 1 as often as possible. We can verify that this is indeed the optimal strategy.

Fix an input distribution $P = (1-p, p)$ so that $P(X=1) = p$. We can define a channel $U(s|x) \in \mathcal{U}(P, \Lambda)$ via a matrix

$$U \begin{pmatrix} U_{00} & 1 - U_{00} \\ 1 - U_{11} & U_{11} \end{pmatrix} , \tag{C.3}$$

where $U_{ij} = U(j|i)$. Under U the averaged channel $V = \sum_s W(y|x,s)U(s|x)$ can be written as

$$V = \begin{pmatrix} U_{00} & 1 - U_{00} & 0 \\ 0 & 1 - U_{11} & U_{11} \end{pmatrix} , \tag{C.4}$$

For fixed p want to minimize the mutual information $I\left(P, V\right)$ over U_{00} and U_{11} subject to the cost constraint

$$\mathbb{E}_{PU}[l(s)] = (1-p)(1-U_{00}) + pU_{11} \le \Lambda \ . \tag{C.5}$$

We form the Lagrangian

$$\begin{aligned} J(U_{00}, U_{11}, \mu) = (1-p)U_{00} \log \frac{1}{1-p} &+ pU_{11} \log \frac{1}{1-p} \\ &+ (1-p)(1-U_{00}) \log \frac{1-U_{00}}{(1-p)(1-U_{00}) + p(1-U_{11})} \\ &+ p(1-U_{11}) \log \frac{1-U_{11}}{(1-p)(1-U_{00}) + p(1-U_{11})} \\ &+ \mu((1-p)(1-U_{00}) + pU_{11}) \ . \end{aligned} \tag{C.6}$$

Differentiating with respect to U_{00} and U_{11} and canceling terms we obtain:

$$\begin{aligned}
\frac{\partial J}{\partial U_{00}} &= (1-p)\log\frac{1}{1-p} - (1-p)\log\frac{1-U_{00}}{(1-p)(1-U_{00})+p(1-U_{11})} - (1-p)\mu \\
&= -(1-p)\left(\log\frac{(1-p)(1-U_{00})}{(1-p)(1-U_{00})+p(1-U_{11})} + \mu\right) \qquad \text{(C.7)} \\
\frac{\partial J}{\partial U_{11}} &= p\log\frac{1}{p} - p\log\frac{1-U_{11}}{(1-p)(1-U_{00})+p(1-U_{11})} + p\mu \\
&= -p\left(\log\frac{p(1-U_{11})}{(1-p)(1-U_{00})+p(1-U_{11})} - \mu\right)\ . \qquad \text{(C.8)}
\end{aligned}$$

If the constraint is inactive, then the multiplier $\mu = 0$ and we can see that the the partial derivatives with respect to U_{00} and U_{11} are both positive. Thus the minimizing point is at $U_{00} = 0$ and $U_{11} = 0$. The resulting mutual information is 0 because the output is always equal to 1. Substituting 0 for U_{00} and U_{11} in the cost formula shows that the constraint is inactive when

$$\Lambda \geq 1 - p\ . \qquad \text{(C.9)}$$

Turning now to the case where the constraint is active, we have $\mu > 0$. Again, the derivative with respect to U_{11} is always positive, so the optimal $U_{11} = 0$. Since the constraint is active we can see from the cost expression that

$$1 - U_{00} = \frac{\Lambda}{1-p}\ . \qquad \text{(C.10)}$$

Substituting this into the mutual information expression we obtain

$$
\begin{aligned}
f(p) &= (1-p-\Lambda)\log\frac{1}{1-p} + \Lambda\log\frac{\Lambda/(1-p)}{\Lambda+p} + p\log\frac{1}{\Lambda+p} && \text{(C.11)} \\
&= h_b(p) + \Lambda\log\frac{\Lambda}{\Lambda+p} + p\log\frac{p}{\Lambda+p} && \text{(C.12)} \\
&= h_b(p) - (\Lambda+p)h_b\left(\frac{\Lambda}{\Lambda+p}\right) \ . && \text{(C.13)}
\end{aligned}
$$

We now have to maximize $f(p)$ subject to $p \le \min(\Gamma, 1-\Lambda)$. Taking the derivative of $f(p)$ we get

$$
\begin{aligned}
\frac{df}{dp} &= \log\frac{1-p}{p} + \frac{\Lambda}{\Lambda+p}\log\frac{p}{\Lambda} - h_b\left(\frac{\Lambda}{\Lambda+p}\right) && \text{(C.14)} \\
&= \log\frac{1-p}{\Lambda+p} \ . && \text{(C.15)}
\end{aligned}
$$

We have $df/dp = 0$ at $p = (1-\Lambda)/2$. Therefore if $\Gamma \ge (1-\Lambda)/2$ then we can choose $p = (1-\Lambda)/2$. In the range $(0, (1-\Lambda)/2)$ we have $df/dp > 0$, so if $\Gamma < (1-\Lambda)/2$ the optimal choice is $p = \Gamma$. To summarize:

$$
C_{\mathrm{dep}}(\Gamma, \Lambda) = \begin{cases} h_b\left(\frac{1-\Lambda}{2}\right) - \frac{1+\Lambda}{2}h_b\left(\frac{2\Lambda}{1+\Lambda}\right) & \Gamma \ge \frac{1-\Lambda}{2} \\ h_b(\Gamma) - (\Lambda+\Gamma)h_b\left(\frac{\Lambda}{\Lambda+\Gamma}\right) & \Gamma < \frac{1-\Lambda}{2} \end{cases} \qquad \text{(C.16)}
$$

C.2 Not all AVCs are distortion-constrained channels

In this section we will look at an example of an AVC that cannot be thought of as a distortion channel in the sense of [1]. Because the capacity formulae under both models involves the minimization over a constrained set of channels, we will look at the generic minimization problem over a set of binary-input binary-output channels.

We want to investigate the structure of a minimization

$$\min_{\mathbf{z}} I(P, W(\mathbf{z})) \tag{C.17}$$

$$s.t. \qquad f(P, \mathbf{z}) \geq 0 \tag{C.18}$$

Where $P = (1-p, p)$ is the input distribution of a binary input, binary output channel, and

$$W = \begin{pmatrix} W_{00}(\mathbf{z}) & W_{01}(\mathbf{z}) \\ W_{10}(\mathbf{z}) & W_{11}(\mathbf{z}) \end{pmatrix} . \tag{C.19}$$

is the channel transition matrix that is a function of some parameter $\mathbf{z}$.

We form the Lagrangian

$$J(\mathbf{z}) = I(P, W(\mathbf{z})) + \lambda f(P, \mathbf{z}) . \tag{C.20}$$

Let z be some element of $\mathbf{z}$. Then

$$\frac{\partial}{\partial z}\left((1-p)W_{00}\log\left(\frac{W_{00}}{(1-p)W_{00}+pW_{10}}\right)\right) \tag{C.21}$$

$$= (1-p)\frac{\partial W_{00}}{\partial z}\log\left(\frac{W_{00}}{(1-p)W_{00}+pW_{10}}\right) + (1-p)\left((1-p)W_{00}+pW_{10}\right)\frac{\partial}{\partial z}\left(\frac{W_{00}}{(1-p)W_{00}+pW_{10}}\right) \tag{C.22}$$

$$= (1-p)\frac{\partial W_{00}}{\partial z}\log\left(\frac{W_{00}}{(1-p)W_{00}+pW_{10}}\right) + (1-p)p\left(\frac{W_{10}\frac{\partial W_{00}}{\partial z} - W_{00}\frac{\partial W_{10}}{\partial z}}{(1-p)W_{00}+pW_{10}}\right) . \tag{C.23}$$

Similarly:

$$\frac{\partial}{\partial z}\left(pW_{10}\log\left(\frac{W_{10}}{(1-p)W_{00}+pW_{10}}\right)\right) \tag{C.24}$$

$$= p\frac{\partial W_{10}}{\partial z}\log\left(\frac{W_{10}}{(1-p)W_{00}+pW_{10}}\right)+(1-p)p\left(\frac{W_{00}\frac{\partial W_{10}}{\partial z}-W_{10}\frac{\partial W_{00}}{\partial z}}{(1-p)W_{00}+pW_{10}}\right)\ . \tag{C.25}$$

Therefore the second terms cancel and we get

$$\begin{aligned}\frac{\partial I(P,W)}{\partial z} &= (1-p)\frac{\partial W_{00}}{\partial z}\log\left(\frac{W_{00}}{(1-p)W_{00}+pW_{10}}\right)\\ &\quad + (1-p)\frac{\partial W_{01}}{\partial z}\log\left(\frac{W_{01}}{(1-p)W_{01}+pW_{11}}\right)\\ &\quad + p\frac{\partial W_{10}}{\partial z}\log\left(\frac{W_{10}}{(1-p)W_{00}+pW_{10}}\right)\\ &\quad + p\frac{\partial W_{11}}{\partial z}\log\left(\frac{W_{11}}{(1-p)W_{01}+pW_{11}}\right)\ .\end{aligned} \tag{C.26}$$

Now we can use the fact that $W_{00} = 1 - W_{01}$ and $W_{11} = 1 - W_{10}$ to get

$$\begin{aligned}\frac{\partial I(P,W)}{\partial z} &= (1-p)\frac{\partial W_{01}}{\partial z}\log\left(\frac{W_{01}}{1-W_{01}}\cdot\frac{(1-p)(1-W_{01})+pW_{10}}{(1-p)W_{01}+p(1-W_{10})}\right)\\ &\quad + p\frac{\partial W_{10}}{\partial z}\log\left(\frac{W_{10}}{1-W_{10}}\cdot\frac{(1-p)W_{01}+p(1-W_{10})}{(1-p)(1-W_{01})+pW_{10}}\right)\ .\end{aligned} \tag{C.27}$$

C.2.1 Cost constrained AVCs with nosy noise

Suppose we have a state set $\mathcal{S}$ and cost function $l : \mathcal{S} \to \mathbb{R}^+$. The channel model we consider is an AVC with nosy noise. There is a channel $V(y|x,s)$ where the state s can depend on the transmitted symbol x. We define the class of memoryless randomized strategies for choosing the state by channels $U : \mathcal{X} \to \mathcal{S}$. We can write these as a channel matrix

$$U = \begin{pmatrix} U_{0,1} & U_{0,2} & \cdots & U_{0,|\mathcal{S}|} \\ U_{1,1} & U_{1,2} & \cdots & U_{1,|\mathcal{S}|} \end{pmatrix} , \tag{C.28}$$

Where $U_{0,1} = 1 - \sum_{j>1} U_{0,j}$ and $U_{1,1} = 1 - \sum_{j>1} U_{1,j}$. The channel induced by such a U is given by:

$$W_{00} = \sum_s V(0|0, s)U_{0,s} \tag{C.29}$$

$$W_{01} = \sum_s V(1|0, s)U_{0,s} \tag{C.30}$$

$$W_{10} = \sum_s V(0|1, s)U_{1,s} \tag{C.31}$$

$$W_{11} = \sum_s V(1|1, s)U_{1,s} \ . \tag{C.32}$$

For a cost constraint Λ, we restrict our attention to U such that

$$(1 - p) \sum_s U_{0,s} l(s) + p \sum_s U_{1,s} l(s) \leq \Lambda \ . \tag{C.33}$$

The U takes the place of our variable $\mathbf{z}$ in the previous section, and the constraint f is given by the cost constraint C.33.

We can compute partial derivatives:

$$\frac{\partial W_{01}}{\partial U_{0,s}} = V(1|0,s) - V(1|0,1) \qquad s > 1 \tag{C.34}$$

$$\frac{\partial W_{10}}{\partial U_{1,s}} = V(0|1,s) - V(0|1,1) \qquad s > 1 \tag{C.35}$$

$$\frac{\partial f(P,U)}{\partial U_{0,s}} = (1-p)(l(s) - l(1)) \tag{C.36}$$

$$\frac{\partial f(P,U)}{\partial U_{1,s}} = p(l(s) - l(1)) \ . \tag{C.37}$$

The other partial derivatives are 0.

C.2.2 The rate distortion problem

Suppose we fix an input distribution $P = (1-p,p)$ on $\mathcal{X}$ and then ask for the W that minimizes the mutual information $I(P,W)$ subject to a distortion constraint:

$$\sum_{x,y} P(x)W(y|x)d(x,y) \leq D \ . \tag{C.38}$$

We will write d_{xy} for the number $d(x,y)$. We can carry out the Lagrangian optimization directly to get the conditions

$$\frac{\partial J}{\partial W_{01}} = (1-p)\log\left(\frac{W_{01}}{1-W_{01}} \cdot \frac{(1-p)(1-W_{01}) + pW_{10}}{(1-p)W_{01} + p(1-W_{10})}\right) + (1-p)\lambda(d_{01} - d_{00}) \tag{C.39}$$

$$\frac{\partial J}{\partial W_{01}} = p\log\left(\frac{W_{10}}{1-W_{10}} \cdot \frac{(1-p)W_{01} + p(1-W_{10})}{(1-p)(1-W_{01}) + pW_{10}}\right) + p\lambda(d_{10} - d_{11}) \ . \tag{C.40}$$

So the conditions in the end are:

$$0 = \log\left(\frac{W_{01}}{1-W_{01}} \cdot \frac{(1-p)(1-W_{01}) + pW_{10}}{(1-p)W_{01} + p(1-W_{10})}\right) + \lambda(d_{01} - d_{00}) \tag{C.41}$$

$$0 = \log\left(\frac{W_{10}}{1-W_{10}} \cdot \frac{(1-p)W_{01} + p(1-W_{10})}{(1-p)(1-W_{01}) + pW_{10}}\right) + \lambda(d_{10} - d_{11}) \tag{C.42}$$

$$D = (1-p)W_{01}d_{01} + (1-p)(1-W_{01})d_{00} + pW_{10}d_{10} + p(1-W_{10})d_{11} \ . \tag{C.43}$$

C.2.3 The BSC / Z switching channel

Consider the AVC with binary inputs and outputs and ternary state $\mathcal{S} = \{0, 1, 2\}$. Under state 0 the channel is a BSC with crossover probability a, under 1 it is a Z-channel with parameter b, and under 2 it is an S-channel with parameter c:

$$V(y|x,0) = \begin{pmatrix} 1-a & a \\ a & 1-a \end{pmatrix} \tag{C.44}$$

$$V(y|x,1) = \begin{pmatrix} 1 & 0 \\ b & 1-b \end{pmatrix} \tag{C.45}$$

$$V(y|x,2) = \begin{pmatrix} 1-c & c \\ 0 & 1 \end{pmatrix} \ . \tag{C.46}$$

Then for a randomized jammer strategy:

$$U = \begin{pmatrix} 1 - U_{01} - U_{02} & U_{01} & U_{02} \\ 1 - U_{11} - U_{12} & U_{11} & U_{12} \end{pmatrix} \ . \tag{C.47}$$

We can compute W as:

$$W = \begin{pmatrix} (1-U_{01}-U_{02})(1-a)+U_{01}+U_{02}(1-c) & (1-U_{01}-U_{02})a+U_{02}c \\ (1-U_{11}-U_{12})a+U_{11}b & (1-U_{11}-U_{12})(1-a)+U_{11}(1-b)+U_{12} \end{pmatrix} . \tag{C.48}$$

Now,

$$\frac{\partial W_{01}}{\partial U_{01}} = -a \tag{C.49}$$

$$\frac{\partial W_{01}}{\partial U_{02}} = -a+c \tag{C.50}$$

$$\frac{\partial W_{10}}{\partial U_{11}} = -a+b \tag{C.51}$$

$$\frac{\partial W_{10}}{\partial U_{12}} = -a \ . \tag{C.52}$$

So we get the following equations from the optimization:

$$0 = -a\log\left(\frac{W_{01}}{1-W_{01}}\cdot\frac{(1-p)(1-W_{01})+pW_{10}}{(1-p)W_{01}+p(1-W_{10})}\right)+\lambda'(l(1)-l(0)) \tag{C.53}$$

$$0 = -(a-c)\log\left(\frac{W_{01}}{1-W_{01}}\cdot\frac{(1-p)(1-W_{01})+pW_{10}}{(1-p)W_{01}+p(1-W_{10})}\right)+\lambda'(l(2)-l(0)) \tag{C.54}$$

$$0 = -(a-b)\log\left(\frac{W_{10}}{1-W_{10}}\cdot\frac{(1-p)W_{01}+p(1-W_{10})}{(1-p)(1-W_{01})+pW_{10}}\right)+\lambda'(l(1)-l(0)) \tag{C.55}$$

$$0 = -a\log\left(\frac{W_{10}}{1-W_{10}}\cdot\frac{(1-p)W_{01}+p(1-W_{10})}{(1-p)(1-W_{01})+pW_{10}}\right)+\lambda'(l(2)-l(0)) \ . \tag{C.56}$$

Or:

$$0 = -a \log \left(\frac{W_{01}}{1 - W_{01}} \cdot \frac{(1-p)(1-W_{01}) + pW_{10}}{(1-p)W_{01} + p(1-W_{10})} \right) - \lambda' \frac{l(1) - l(0)}{a} \tag{C.57}$$

$$0 = -(a-c) \log \left(\frac{W_{01}}{1 - W_{01}} \cdot \frac{(1-p)(1-W_{01}) + pW_{10}}{(1-p)W_{01} + p(1-W_{10})} \right) - \lambda' \frac{l(2) - l(0)}{a-c} \tag{C.58}$$

$$0 = \log \left(\frac{W_{10}}{1 - W_{10}} \cdot \frac{(1-p)W_{01} + p(1-W_{10})}{(1-p)(1-W_{01}) + pW_{10}} \right) - \lambda' \frac{l(1) - l(0)}{a-b} \tag{C.59}$$

$$0 = \log \left(\frac{W_{10}}{1 - W_{10}} \cdot \frac{(1-p)W_{01} + p(1-W_{10})}{(1-p)(1-W_{01}) + pW_{10}} \right) - \lambda' \frac{l(2) - l(0)}{a} \,. \tag{C.60}$$

C.2.4 The final pieces

Consider the AVC above with total cost constraint Λ and let $P = (1-p, p)$ be the capacity-achieving input distribution for this AVC. For this P, let W be the minimizer of $I(P, W)$ subject to the constraints. Suppose that there exists a distortion measure $d(x, y)$ and total distortion D such that the Agarwal-Sahai-Mitter problem with input P, measure d and bound D yields W as the optimal test channel. Then an examination of the conditions in (C.41)–(C.42) and (C.57)–(C.60) yield the following:

$$-\lambda' \frac{l(2) - l(0)}{a-c} = \lambda(d_{01} - d_{00}) = -\lambda' \frac{l(1) - l(0)}{a} \,. \tag{C.61}$$

This implies that

$$\frac{l(2) - l(0)}{a-c} = \frac{l(1) - l(0)}{a} \,. \tag{C.62}$$

But this need not hold. Therefore in some cases we cannot find a distortion measure d that yields the same mutual information minimizing channel.

Appendix D

Proofs for list decoding under average error

D.1 Proof of Lemma 4

Proof of Lemma 4. Fix $\epsilon > 0$ and P. We consider two cases depending on whether $\min_{x \in \mathcal{X}} P(x) = 0$ or not.

Case 1. First suppose $\min_{x \in \mathcal{X}} P(x) = \beta > 0$. Consider a set of distributions $\{P_i : i \in [L]\}$ satisfying (2.64) and let $\overline{P}(x_1^L)$ be a joint distribution satisfying (2.65). We treat probability distributions as vectors in $\mathbb{R}^{|\mathcal{X}|^L}$. We can construct a distribution $\hat{P}$ satisfying (2.66) and (2.67) in two steps: first we project $\overline{P}$ onto the set of all vectors whose entries sum to 1 and satisfy (2.66), and then we find a $\hat{P}$ close to this projection which is a proper probability distribution.

Let $\mathcal{B}$ be the subspace of $\mathbb{R}^{|\mathcal{X}|^L}$ of all vectors P' satisfying the marginal constraints (2.66) as well as the sum probability constraint

$$\sum_{x_1^L} P'(x_1^L) = 1 \ . \tag{D.1}$$

We can summarize these linear constraints in the matrix form

$$AP' = b' \ , \tag{D.2}$$

where A contains the coefficients on the left-hand sides of the constraints (2.66) and (D.1) and b' has the right-hand sides. We can assume A has full row-rank by removing linearly dependent constraints. Note that the distribution $\overline{P}$ satisfies

$$A\overline{P} = \overline{b} \ , \tag{D.3}$$

where $\overline{b}$ has the right-hand sides of (2.65) instead of (2.66).

Now let $\tilde{P}$ be the Euclidean projection of $\overline{P}$ onto the subspace $\mathcal{B}$:

$$\tilde{P} = \overline{P} + A^T(AA^T)^{-1}(b' - A\overline{P}) \ . \tag{D.4}$$

The error in the projection is

$$\overline{P} - \tilde{P} = A^T(AA^T)^{-1}(A\overline{P} - b') \tag{D.5}$$

$$= A^T(AA^T)^{-1}(\overline{b} - b') \ . \tag{D.6}$$

From (2.64) we can see that all elements of $(\overline{b} - b')$ are in $(-\delta, \delta)$. Since the rows of A are linearly independent, the singular values of A are strictly positive and a function of $|\mathcal{X}|$ and L only. Therefore there is a function $\mu_1(|\mathcal{X}|, L)$ such that

$$\left\| A^T(AA^T)^{-1}(\overline{b} - b') \right\|_2 < \mu_1(|\mathcal{X}|, L) \cdot \delta \ . \tag{D.7}$$

Since $|\mathcal{X}|$ is finite there is a function $\mu_2(|\mathcal{X}|, L)$ such that

$$d_{\max}\left(\tilde{P}(x_1^L), \overline{P}(x_1^L)\right) < \mu_2(|\mathcal{X}|, L) \cdot \delta \ . \tag{D.8}$$

If the resulting $\tilde{P}$ from this first projection has all nonnegative entries, then we set $\hat{P} = \tilde{P}$ and choose δ sufficiently small so that $\mu_2(|\mathcal{X}|, L) \cdot \delta < \epsilon$.

If $\tilde{P}$ has entries that are not in $[0, 1]$ then it is not a valid probability distribution. However, since $\overline{P}$ is a probability distribution, we know that

$$\min_{x_1^L} \tilde{P}(x_1^L) > -\mu_2(|\mathcal{X}|, L) \cdot \delta \ . \tag{D.9}$$

Let P^L be the joint distribution on $\mathcal{X}^L$ with independent marginals P:

$$P^L(x_1, \ldots, x_L) = P(x_1) \cdots P(x_L) \ . \tag{D.10}$$

Since $\min_x P(x) > \beta$ we have $P^L(x_1^L) > \beta^L$ for all L. Let

$$\alpha = \frac{\mu_2(|\mathcal{X}|, L) \cdot \delta}{\beta^L} \ , \tag{D.11}$$

and set

$$\hat{P} = (1 - \alpha)\tilde{P} + \alpha P^L \ . \tag{D.12}$$

Then $\hat{P}(x_1^L) > 0$ for all x_1^L and by the triangle inequality:

$$d_{\max}\left(\overline{P}, \hat{P}\right) \leq d_{\max}\left(\overline{P}, \tilde{P}\right) + d_{\max}\left(\tilde{P}, \hat{P}\right) \tag{D.13}$$

$$< \mu_2(|\mathcal{X}|, L) \cdot \delta + \alpha d_{\max}\left(\tilde{P}, P^L\right) \tag{D.14}$$

$$< \left(1 + \frac{1}{\beta^L}\right) \mu_2(|\mathcal{X}|, L) \cdot \delta \ . \tag{D.15}$$

Therefore for δ sufficiently small, we can choose a $\hat{P}$ such that $d_{\max}\left(\overline{P}, \hat{P}\right) < \epsilon$ for any $\epsilon > 0$.

Case 2. We turn now to the second case. Suppose that $\min_{x \in \mathcal{X}} P(x) = 0$. Let $\mathcal{X}_0 = \{x \in \mathcal{X} : P(x) = 0\}$ and $\mathcal{Z} = \mathcal{X} \setminus \mathcal{X}_0$. Let $Q \in \mathcal{P}(\mathcal{Z})$ be the restriction of P to $\mathcal{Z}$. Then Q is a probability distribution on $\mathcal{Z}$. First suppose that $|\mathcal{Z}| = 1$. Then $P(x) = 1$ for some $x \in \mathcal{X}$. Let

$$\hat{P}(x_1^L) = P(x_1) \cdots P(x_L) \ . \tag{D.16}$$

Since all the marginal distributions P_i of $\overline{P}$ satisfy $d_{\max}(P, P_i) < \delta$ we know that $d_{\max}\left(\overline{P}, \hat{P}\right) < \delta$.

Now suppose $|\mathcal{Z}| \geq 2$. We can construct $\hat{P}$ by first finding a a joint distribution $\overline{Q}$ that is close to $\overline{P}$ and then invoking the first case of this proof on $\overline{Q}$. From (2.64) we know that for some $c > 0$ we have

$$\sum_{x_1^L \notin \mathcal{Z}^L} \overline{P}(x_1, x_2, \ldots, x_L) \triangleq c\delta \tag{D.17}$$

$$< |\mathcal{X}|^L \delta \ . \tag{D.18}$$

Define $\overline{Q}$ by

$$\overline{Q}(x_1^L) = \begin{cases} \overline{P}(x_1^L) + |\mathcal{Z}|^{-L} c\delta & x_1^L \in \mathcal{Z}^L \\ 0 & x_1^L \notin \mathcal{Z}^L \end{cases} \tag{D.19}$$

Since $\overline{Q}$ has support only on $\mathcal{Z}^L$ we can think of it either as a distribution on $\mathcal{X}^L$ or on $\mathcal{Z}^L$. Note that

$$d_{\max}\left(\overline{P}, \overline{Q}\right) < c\delta \ . \tag{D.20}$$

Let $\{Q_i : i \in [L]\}$ be the i-th marginal distributions of $\overline{Q}$:

$$Q_i(x_i) = \sum_{x_j : j \neq i} \overline{Q}(x_1, x_2, \ldots, x_L) = Q_i(x_i) \qquad \forall i,\ x_i \in \mathcal{Z} \ . \tag{D.21}$$

Then we have for some $c' > 0$

$$d_{\max}\left(Q, Q_i\right) < c'\delta \ . \tag{D.22}$$

Now we can apply Case 1 of this proof using the set $\mathcal{Z}$ and distributions Q, $\{Q_i\}$, and $\overline{Q}$. For any $\epsilon_1 > 0$ we can find a $\delta_1 > 0$ such that if $\{Q_i\}$ satisfy

$$d_{\max}\left(Q, Q_i\right) < \delta_1 \ , \tag{D.23}$$

then there exists a $\hat{Q}$ with marginals equal to Q such that

$$d_{\max}\left(\overline{Q}, \hat{Q}\right) < \epsilon_1 \ . \tag{D.24}$$

Let $\hat{P}$ be the extension of $\hat{Q}$ to a distribution on $\mathcal{X}^L$ by setting $\hat{P}(x_1^L) = \hat{Q}(x_1^L)$ for

$x_1^L \in \mathcal{Z}^L$ and 0 elsewhere. By the triangle inequality we have

$$d_{\max}\left(\overline{P}, \hat{Q}\right) \leq d_{\max}\left(\overline{P}, \overline{Q}\right) + d_{\max}\left(\overline{Q}, \hat{Q}\right) \tag{D.25}$$

$$< c\delta + \epsilon_1 \ . \tag{D.26}$$

We can choose δ sufficiently small so that δ_1 and ϵ_1 are sufficiently small to guarantee that this distance is less than ϵ. □

D.2 Proof of Lemma 11

Proof of Lemma 11. Fix $\epsilon_3 > 0$, $\beta > 0$, and distribution $P \in \mathcal{P}(\mathcal{X})$ with $I(P) > 0$ and $\min_x P(x) \geq \beta$. Let $M = \tilde{L}_{\text{sym}}(P, \Lambda) + 1$. Choose R such that $N = M\exp(nR)$ is an integer and

$$I(P, \Lambda) - \epsilon_3 < R < I(P, \Lambda) - 2\epsilon_3/3 \ . \tag{D.27}$$

For any $\epsilon > 0$ and blocklength n sufficiently large we can choose N codewords of type P with the properties guaranteed by Lemma 10. Denote these codewords by $\{\mathbf{x}_i : i \in [N]\}$. We will use these codewords together with the decoding rule in Definition 1 using a parameter $\eta > 0$ to form a list-decodable code of list size M. We first show how to choose η.

We can show that the list produced by the decoding rule does not have more than M codewords provided that η is sufficiently small. We have two conditions on η. First, suppose $P_{YXS} \in \mathcal{G}_\eta(\Lambda)$. Then from Pinsker's inequality [44, p. 58, Problem

17], we have:

$$\eta \geq D\left(P_{XSY} \ \| \ P_X \times P_S \times W\right) \tag{D.28}$$

$$\geq \frac{1}{2 \ln 2} \left(\sum_{y,x,s} |P_{YXS}(y,x,s) - W(y|x,s) P_X(x) P_S(s)| \right)^2 \ . \tag{D.29}$$

By choosing η sufficiently small, we can make the variational distance between $P_{YXS} \in \mathcal{G}_\eta(\Lambda)$ and $P_X \times P_S \times W$ as small as we like. Because the mutual information is uniformly continuous in P_{XY}, for any $\epsilon_3/3 > 0$ we can choose η sufficiently small such that for all $P_{YXS} \in \mathcal{G}_\eta(\Lambda)$ we have

$$I\left(X \ \wedge \ Y\right) \geq I(P_X) - \epsilon_3/3 \ . \tag{D.30}$$

Now, suppose that there exists a $\mathbf{y} \in \mathcal{Y}^n$ such that $M+1$ codewords $\{\mathbf{x}_{i_j} : j \in [M+1]\}$ exist satisfying the conditions of Definition 1. The decoding rule implies that there are state sequences $\{\mathbf{s}_{i_j} : j \in [M+1]\}$ for each codeword. Let the tuple of random variables $(\{X_j, S_j\}, Y)$ have joint distribution according to the type of $(\{\mathbf{x}_{i_j}, \mathbf{s}_{i_j}\}, \mathbf{y})$:

$$P_{\{X_j,S_j\},Y} = T_{\{\mathbf{x}_{i_j},\mathbf{s}_{i_j} : j \in [M+1]\},\mathbf{y}} \ . \tag{D.31}$$

Because $\{\mathbf{x}_{i_j} : j \in [M+1]\}$ satisfy the decoding rule, we have:

$$\min_x P(x) \geq \beta \tag{D.32}$$

$$P_{X_i} = P \tag{D.33}$$

$$P_{YX_iS_i} \in \mathcal{G}_\eta(\Lambda) \tag{D.34}$$

$$I\left(YX_i \ \wedge \ X_{-\{i\}}^{M+1} \middle| S_i\right) \leq \eta \ . \tag{D.35}$$

From Lemma 9 we can choose an η sufficiently small such that no tuple of $M+1$ random variables can satisfy these conditions. Therefore no such $\mathbf{y}$ can exist and the decoding rule always outputs a list of size M or smaller.

The remainder of the proof is to show (2.146), which says that the probability of error averaged over the messages can be made small for every $\mathbf{s} \in \mathcal{S}^n(\Lambda)$. We will bound the average error for a given $\mathbf{s}$:

$$\varepsilon(\mathbf{s}) = \frac{1}{N} \sum_{i=1}^{N} \varepsilon(i, \mathbf{s}) \ . \tag{D.36}$$

We have an error when transmitting message i under state sequence $\mathbf{s}$ if the codeword $\mathbf{x}_i$ does not satisfy the conditions of the decoder in Definition 1.

Fix $\mathbf{s} \in \mathcal{S}^n(\Lambda)$. We divide the set of messages $[N]$ into two sets based on the joint type $T_{\mathbf{x}_i \mathbf{s}}$

$$F(\mathbf{s}) = \left\{ i : I\left(X \ \wedge \ S\right) < \epsilon, \ P_{XS} = T_{\mathbf{x}_i \mathbf{s}} \right\} \ . \tag{D.37}$$

For a joint type P_{XS} we can bound $|F^c(\mathbf{s})|$ using part 1 of Lemma 10:

$$\begin{aligned} \frac{1}{N} \sum_{i \in F^c(\mathbf{s})} \varepsilon(i, \mathbf{s}) &\leq \frac{|F^c(\mathbf{s})|}{N} \\ &\leq \sum_{P_{XS} \in \mathcal{P}_n(\mathcal{X}, \mathcal{S})} \exp(-n\epsilon/2) \\ &\leq (n+1)^{|\mathcal{X}||\mathcal{S}|} \exp(-n\epsilon/2) \ . \end{aligned} \tag{D.38}$$

Therefore for n sufficiently large we can upper bound the error:

$$\varepsilon(\mathbf{s}) = \frac{1}{N} \sum_{i \in F^c(\mathbf{s})} \varepsilon(i, \mathbf{s}) + \frac{1}{N} \sum_{i \in F(\mathbf{s})} \varepsilon(i, \mathbf{s}) \tag{D.39}$$

$$\leq \exp(-n\epsilon/3) + \frac{1}{N} \sum_{i \in F(\mathbf{s})} \varepsilon(i, \mathbf{s}) \ . \tag{D.40}$$

We now turn to bounding the error $\varepsilon(i, \mathbf{s})$ for $i \in F(\mathbf{s})$. A decoding error occurs if the true codeword does not satisfy the first or second condition of the decoding rule. For each message i we first consider those $\mathbf{y} \in \mathcal{Y}^n$ such that the first part of the decoding rule fails. Define the set

$$A_i(\mathbf{s}) = \{\mathbf{y} : T_{\mathbf{x}_i \mathbf{s} \mathbf{y}} \in \mathcal{G}_\eta(\Lambda)\} \ . \tag{D.41}$$

An output $\mathbf{y} \in A_i^c(\mathbf{s})$ fails the first part of the decoding rule for message i. Since $i \in F(\mathbf{s})$ we know that for random variables X and S with joint distribution $P_{XS} = T_{\mathbf{x}_i \mathbf{s}}$:

$$D\left(P_{XS} \ \| \ P_X \times P_S\right) = I\left(X \ \wedge \ S\right) < \epsilon \ . \tag{D.42}$$

For $\mathbf{y} \in A_i^c(\mathbf{s})$, let $P_{XSY} = T_{\mathbf{x}_i \mathbf{s} \mathbf{y}}$. We have the following equality:

$$D\left(P_{XSY} \ \| \ P_{XS} \times W\right) + I\left(X \ \wedge \ S\right) = D\left(P_{XSY} \ \| \ P_X \times P_S \times W\right) \ . \tag{D.43}$$

Since $\mathbf{y} \in A_i^c(\mathbf{s})$, we know $T_{\mathbf{x}_i \mathbf{s} \mathbf{y}} \notin \mathcal{G}_\eta(\Lambda)$, which implies by (2.115) that

$$D\left(P_{XSY} \ \| \ P_X \times P_S \times W\right) \leq \eta \ . \tag{D.44}$$

Therefore:

$$D\left(P_{XSY} \ \| \ P_{XS} \times W\right) > \eta - \epsilon \ . \tag{D.45}$$

Thus we can bound using (D.45) and (2.9):

$$\begin{aligned}\sum_{i \in F(\mathbf{s})} W^n(A_i^c(\mathbf{s})|\mathbf{x}_i, \mathbf{s}) &\leq \sum_{P_{XSY} \notin \mathcal{G}_\eta(\Lambda)} W^n\left(\{\mathbf{y} : P_{XSY} = T_{\mathbf{x}_i\mathbf{s}\mathbf{y}}\}|\mathbf{x}_i\mathbf{s}\right) \\ &\leq (n+1)^{|\mathcal{X}||\mathcal{S}||\mathcal{Y}|} \exp(-nD\left(P_{XSY} \ \| \ P_{XS} \times W\right)) \\ &\leq (n+1)^{|\mathcal{X}||\mathcal{S}||\mathcal{Y}|} \exp(-n(\eta - \epsilon)) \ . \end{aligned} \tag{D.46}$$

Thus for ϵ sufficiently small and the blocklength n sufficiently large we have:

$$\frac{1}{N} \sum_{i \in F(\mathbf{s})} W^n(A_i^c(\mathbf{s})|\mathbf{x}_i, \mathbf{s}) \leq \exp(-n(\eta - \epsilon)/2) \ . \tag{D.47}$$

Now we turn to bounding the error for messages $i \in F(\mathbf{s})$ over those $\mathbf{y}$ such that $T_{\mathbf{x}_i\mathbf{s}\mathbf{y}} \in \mathcal{G}_\eta(\Lambda)$. In this case there is an error if the second part of the decoding rule fails. For a subset $J \subset [N]$, let $\mathbf{x}_J = \{\mathbf{x}_j : j \in J\}$. An output $\mathbf{y} \in A_i(\mathbf{s})$ results in an error if there exists a set $J \subset [N] \setminus \{i\}$ with $|J| = M$ such that for each $j \in J$ the first part of the decoding rule is satisfied and

$$I\left(YX \ \wedge \ X^M \middle| S\right) \leq \eta \ , \tag{D.48}$$

where the mutual information is over the joint distribution $P_{YXX^MS} = T_{\mathbf{y},\mathbf{x}_i,\mathbf{x}_J,\mathbf{s}}$. Define the set of joint distributions which violate the second part of the decoding rule:

$$\begin{aligned}\mathcal{H}_\eta = \Big\{ P_{YXX^MS} \in \mathcal{P}_n(\mathcal{Y} \times \mathcal{X}^{M+1} \times \mathcal{S}) : &\\ & P_{YXS} \in \mathcal{G}_\eta(\Lambda) \\ & \exists S_j \ s.t. \ P_{YX_jS_j} \in \mathcal{G}_\eta(\Lambda), \ j \in [M] \\ & I\left(YX \ \wedge \ X^M \middle| S\right) > \eta \Big\} \ . \end{aligned} \tag{D.49}$$

For each type P_{YXX^MS}, message i and state sequence $\mathbf{s}$ we can define the set of outputs $\mathbf{y}$ for which a set J violating the second part of the decoding rule exists:

$$E(P_{YXX^MS}, i, \mathbf{s}) = \{\mathbf{y} : \exists J \subset [N] \setminus \{i\}, |J| = M, \ s.t. \ P_{YXX^MS} = T_{\mathbf{y}\mathbf{x}_i\mathbf{x}_J\mathbf{s}}\} \ . \tag{D.50}$$

Therefore we can write the remaining error term as:

$$\frac{1}{N} \sum_{i \in F(\mathbf{s})} \sum_{P_{YXX^MS} \in \mathcal{H}_\eta} W^n \left(E(P_{YXX^MS}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s} \right) \ . \tag{D.51}$$

In order to bound this error expression, we look at two different cases for each joint distribution P_{YXX^MS}. In the first case, suppose that

$$R < \min_j I\left(X_j \ \wedge \ S\right) \ . \tag{D.52}$$

We consider two sub-cases. If $I\left(X \ \wedge \ X^M S\right) \geq \epsilon$ then part 5 of Lemma 10 shows that

$$\frac{1}{N} \left| \{ i : (\mathbf{x}_i, \mathbf{x}_J, \mathbf{s}) \in T_{XX^LS} \text{ for some } J \subset [N] \setminus \{i\}, \ |J| = M \} \right| \leq \exp(-n\epsilon/2) \ . \tag{D.53}$$

Therefore for P_{YXX^MS} such that (D.52) holds and $I\left(X \;\wedge\; X^M S\right) \geq \epsilon$ we have

$$\frac{1}{N} \sum_{i \in F(\mathbf{s})} W^n \left(E(P_{YXX^MS}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s}\right) \leq \exp(-n\epsilon/2) \;. \tag{D.54}$$

The second sub-case is for $I\left(X \;\wedge\; X^M S\right) < \epsilon$. Let

$$\mathcal{J}(P_{YXX^MS}) = \{J \subset [N] \setminus \{i\} : T_{\mathbf{x}_i, \mathbf{x}_J, \mathbf{s}} = P_{YXX^MS}\} \;. \tag{D.55}$$

Then

$$W^n \left(E(P_{YXX^MS}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s}\right) \leq \sum_{J \in \mathcal{J}(P_{YXX^MS})} W^n \left(\{\mathbf{y} : P_{YXX^MS} = T_{\mathbf{y}, \mathbf{x}_i, \mathbf{x}_J, \mathbf{s}}\} | \mathbf{x}_i, \mathbf{s}\right) \;. \tag{D.56}$$

From (2.10) we know that each summand can be bounded:

$$W^n \left(\{\mathbf{y} : P_{YXX^MS} = T_{\mathbf{y}, \mathbf{x}_i, \mathbf{x}_J, \mathbf{s}}\} | \mathbf{x}_i, \mathbf{s}\right) \leq \exp\left(-nI\left(Y \;\wedge\; X^M \middle| XS\right)\right) \;. \tag{D.57}$$

Now (2.143) in Part 4 of Lemma 10 shows that $|\mathcal{J}(P_{YXX^MS})| \leq \exp(n\epsilon)$, so

$$W^n \left(E(P_{YXX^MS}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s}\right) \leq \exp\left(-n(I\left(Y \;\wedge\; X^M \middle| XS\right) - \epsilon)\right) \;. \tag{D.58}$$

Since $I\left(X \;\wedge\; X^M S\right) < \epsilon$ and $I\left(YX \;\wedge\; X^M \middle| S\right) > \eta$ for types $P_{YXX^MS} \in \mathcal{H}_\eta$ we can write:

$$\begin{aligned} I\left(Y \;\wedge\; X^M \middle| XS\right) &= I\left(YX \;\wedge\; X^M \middle| S\right) - I\left(X \;\wedge\; X^M \middle| S\right) && \text{(D.59)} \\ &\geq I\left(YX \;\wedge\; X^M \middle| S\right) - I\left(X \;\wedge\; X^M S\right) && \text{(D.60)} \\ &> \eta - \epsilon \;. && \text{(D.61)} \end{aligned}$$

Therefore we get the bound:

$$W^n \left(E(P_{YXX^M S}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s} \right) \leq \exp\left(-n(\eta - 2\epsilon) \right) \ . \tag{D.62}$$

For the second case, suppose the joint distribution $P_{YXX^M S}$. satisfies

$$R \geq \min_j I\left(X_j \ \wedge \ S \right) \ . \tag{D.63}$$

Pick some j for which $R \geq I\left(X_j \ \wedge \ S \right)$. We can first upperbound each summand of (D.51):

$$W^n \left(E(P_{YXX^M S}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s} \right) \leq W^n \left(\left\{ \mathbf{y} : \exists j \neq i \ s.t. \ T_{\mathbf{yxx_js}} = P_{YXX_jS} \right\} | \mathbf{x}_i, \mathbf{s} \right) \ . \tag{D.64}$$

We again consider two sub-cases. Suppose first that

$$I(X \wedge X_j S) \geq |R - I(X_j \wedge S)|^+ + \epsilon \ . \tag{D.65}$$

Then Part 2 of Lemma 10 says that

$$\frac{1}{N} \left| \left\{ i : (\mathbf{x}_i, \mathbf{x}_j, \mathbf{s}) \in T_{XX_jS} \text{ for some } j \neq i \right\} \right| \leq \exp(-n\epsilon/2) \ . \tag{D.66}$$

Therefore

$$\frac{1}{N} \sum_{i \in F(\mathbf{s})} W^n \left(E(P_{YXX^M S}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s} \right) \leq \exp(-n\epsilon/2) \ . \tag{D.67}$$

For the second sub-case, suppose

$$I(X \wedge X_j S) < |R - I(X_j \wedge S)|^+ + \epsilon \ . \tag{D.68}$$

We know $P_{X_j} = P$, so we can further upperbound (D.64):

$$W^n \left(E(P_{YXX^MS}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s} \right) \leq \sum_{j \neq i : T_{\mathbf{x}_i, \mathbf{x}_j, \mathbf{s}} = P_{XX_jS}} W^n \left(\left\{ \mathbf{y} : T_{\mathbf{yxx_js}} = P_{YXX_jS} \right\} | \mathbf{x}_i, \mathbf{s} \right) \ . \tag{D.69}$$

Now using (2.10) we can bound each summand:

$$W^n \left(\left\{ \mathbf{y} : T_{\mathbf{yxx_js}} = P_{YXX_jS} \right\} | \mathbf{x}_i, \mathbf{s} \right) \leq \exp \left(-nI \left(Y \ \wedge \ X_j | \, XS \right) \right) \ . \tag{D.70}$$

Part 3 of Lemma 10 shows that

$$\left| \left\{ j : T_{\mathbf{x}_i, \mathbf{x}_j, \mathbf{s}} = P_{XX_jS} \right\} \right| \leq \exp \left(n \left(|R - I \left(X_j \ \wedge \ XS \right) |^+ + \epsilon \right) \right) \ , \tag{D.71}$$

so we can write a new upperbound:

$$\begin{aligned} &W^n \left(E(P_{YXX^MS}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s} \right) \\ &\qquad \leq \exp \left(-n \left(I \left(Y \ \wedge \ X_j | \, XS \right) - |R - I \left(X_j \ \wedge \ XS \right) |^+ - \epsilon \right) \right) \ . \end{aligned} \tag{D.72}$$

Now we use the fact that $R \geq I \left(X_j \ \wedge \ S \right)$ with (D.68) to get

$$\begin{aligned} R &> I \left(X \ \wedge \ X_j S \right) + I \left(X_j \ \wedge \ S \right) - \epsilon && \text{(D.73)} \\ &\geq I \left(X_j \ \wedge \ X | \, S \right) + I \left(X_j \ \wedge \ S \right) - \epsilon && \text{(D.74)} \\ &= I \left(X_j \ \wedge \ XS \right) - \epsilon \ . && \text{(D.75)} \end{aligned}$$

So

$$I\left(Y \ \wedge \ X_j | \, XS\right) - |R - I\left(X_j \ \wedge \ XS\right)|^{+} - \epsilon \geq I\left(X_j \ \wedge \ YXS\right) - R - 2\epsilon \tag{D.76}$$

$$\geq I\left(X_j \ \wedge \ Y\right) - R - 2\epsilon \ . \tag{D.77}$$

From the definition of $\mathcal{H}_\eta$ we know there exists a random variable S_j for which $P_{YX_jS_j} \in \mathcal{G}_\eta(\Lambda)$ and see that the joint distribution $P_{YX_jS_j}$ yields a mutual information $I\left(X_j \ \wedge \ Y\right) = I\left(X \ \wedge \ Y\right)$, so by (D.27) we have

$$I\left(X_j \ \wedge \ Y\right) - R \geq I(P) - R - \epsilon_3/3 \tag{D.78}$$

$$> \epsilon_3/3 \ . \tag{D.79}$$

Therefore, we can bound:

$$W^n\left(E(P_{YXX^MS}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s}\right) \leq \exp\left(-n(\epsilon_3/3 - 2\epsilon)\right) \ . \tag{D.80}$$

We can finally put the bounds together. From the two cases earlier, we can take the bounds (D.54), (D.62), (D.67) and (D.80) on the sets $E(P_{YXX^MS}, i, \mathbf{s})$ together with a union bound over all types in $\mathcal{H}_\eta$ to get:

$$\frac{1}{N} \sum_{i \in F(\mathbf{s})} \sum_{P_{YXX^MS} \in \mathcal{H}_\eta} W^n\left(E(P_{YXX^MS}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s}\right)$$

$$\leq |\mathcal{H}_\eta| \exp\left(-n \min\left\{\epsilon/2, \eta - 2\epsilon, \epsilon_3/3 - 2\epsilon\right\}\right) \ . \tag{D.81}$$

Since ϵ can be made arbitrarily small, for ϵ sufficiently small and n sufficiently large,

$$\frac{1}{N} \sum_{i \in F(\mathbf{s})} \sum_{P_{YXX^M S} \in \mathcal{H}_\eta} W^n\left(E(P_{YXX^M S}, i, \mathbf{s}) | \mathbf{x}_i, \mathbf{s} \right) \le |\mathcal{H}_\eta| \exp(-n\epsilon/2) \tag{D.82}$$

$$\le \exp(-n\epsilon/3) \ . \tag{D.83}$$

Then (D.38) and (D.47) give us

$$\varepsilon(\mathbf{s}) = \frac{1}{N} \sum_{i=1}^{N} \varepsilon(i, \mathbf{s}) < \exp(-n\epsilon/3) \ . \tag{D.84}$$

Then for $\delta = \epsilon/3$ we have (2.146). Note that ϵ and the blocklength n depend on β, ϵ_3, and $\mathcal{W}$. The bound holds for all $\mathbf{s} \in \mathcal{S}^n(\Lambda)$ so we are done. □

Bibliography

[1] Agarwal, M., Sahai, A., and Mitter, S. Coding into a source: a direct inverse rate-distortion theorem. In *45th Annual Allerton Conference on Communication, Control and Computation* (2006).

[2] Ahlswede, R. A note on the existence of the weak capacity for channels with arbitrarily varying channel probability functions and its relation to Shannon's zero-error capacity. *Annals of Mathematical Statistics 41*, 3 (1970), 1027–1033.

[3] Ahlswede, R. Group codes do not achieve Shannon's channel capacity for the general discrete memoryless channel. *Annals of Mathematical Statistics 42*, 1 (1971), 224–240.

[4] Ahlswede, R. Channel capacities for list codes. *Journal of Applied Probability 10*, 4 (1973), 824–836.

[5] Ahlswede, R. Channels with arbitrarily varying channel probability functions in the presence of noiseless feedback. *Zeitschrift für Wahrscheinlichkeit und verwandte Gebiete 25*, 3 (1973), 239–252.

[6] Ahlswede, R. Elimination of correlation in random codes for arbitrarily varying channels. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 44*, 2 (1978), 159–175.

[7] Ahlswede, R. Coloring hypergraphs : A new approach fo multi-user source coding – I. *Journal of Combinatorics, Information, and System Sciences 4*, 1 (1979), 76–115.

[8] Ahlswede, R. Coloring hypergraphs : A new approach fo multi-user source coding – II. *Journal of Combinatorics, Information and System Sciences 5*, 3 (1980), 220–268.

[9] Ahlswede, R. A method of coding and an application to arbitrarily varying channels. *Journal of Combinatorics, Information and System Sciences 5*, 1 (1980), 10–35.

[10] Ahlswede, R. Arbitrarily varying channels with states sequence known to the sender. *IEEE Transactions on Information Theory 32*, 5 (1986), 621–629.

[11] Ahlswede, R. The maximal error capacity of arbitrarily varying channels for constant list sizes. *IEEE Transactions on Information Theory 39*, 4 (1993), 1416–1417.

[12] Ahlswede, R., Bassalygo, L., and Pinsker, M. Localized random and arbitrary errors in light of arbitrarily varying channel theory. *IEEE Transactions on Information Theory 41*, 1 (1995), 14–25.

[13] Ahlswede, R., and Cai, N. Two proofs of Pinsker's conjecture concerning arbitrarily varying channels. *IEEE Transactions on Information Theory 37*, 6 (1991), 1647–1649.

[14] Ahlswede, R., and Cai, N. Correlated sources help transmission over an arbitrarily varying channel. *IEEE Transactions on Information Theory 43*, 4 (1997), 1254–1255.

[15] Ahlswede, R., and Cai, N. Arbitrarily varying multiple-access channels part I – Ericson's symmetrizability is adequate, Gubner's conjecture is true. *IEEE Transactions on Information Theory 45*, 2 (March 1999), 742–749.

[16] Ahlswede, R., and Csiszár, I. Common randomness in information theory and cryptography – Part I : Secret sharing. *IEEE Transactions on Information Theory 39*, 4 (1993), 1121–1132.

[17] Ahlswede, R., and Csiszár, I. Common randomness in information theory and cryptography – Part II : CR capacity. *IEEE Transactions on Information Theory 44*, 1 (1998), 225–240.

[18] Ahlswede, R., and Simonyi, G. Reusable memories in the light of the old arbitrarily varying and a new outputwise varying channel theory. *IEEE Transactions on Information Theory 37*, 4 (July 1991), 1143–1150.

[19] Ahlswede, R., and Wolfowitz, J. Correlated decoding for channels with arbitrarily varying channel probability functions. *Information and Control 14* (1969), 457–473.

[20] Ahlswede, R., and Wolfowitz, J. The capacity of a channel with arbitrarily varying channel probability functions and binary output alphabet. *Zeitschrift für Wahrscheinlichkeit und verwandte Gebiete 15*, 3 (1970), 186–194.

[21] Başar, T., and Başar, T. *Minimax Causal Transmission of Gaussian Stochastic Processes over Channels Subject to Correlated Jamming*, vol. 129 of *Lecture Notes in Information Sciences and Systems.* Springer-Verlag, 1989, pp. 39–49.

[22] Baker, C., and Chao, I. Information capacity of channels with partially unknown noise. I. finite-dimensional channels. *SIAM Journal of Applied Mathematics 56*, 3 (June 1996), 946–963.

[23] Bergmans, P. Random coding theorem for broadcast channels with degraded components. *IEEE Transactions on Information Theory 19*, 2 (March 1973), 197–207.

[24] Bergmans, P. A simple converse for broadcast channels with additive white Gaussian noise. *IEEE Transactions on Information Theory 20*, 2 (March 1974), 279–280.

[25] Bergmans, P., and Cover, T. Cooperative broadcasting. *IEEE Transactions on Information Theory 20*, 2 (March 1974), 317–324.

[26] Blachman, N. Communication as a game. In *IEE Wescon 1957 Conference Record* (1957), vol. II, pp. 61–66.

[27] Blackwell, D., Breiman, L., and Thomasian, A. The capacity of a class of channels. *Annals of Mathematical Statistics 30*, 4 (1959), 1229–1241.

[28] Blackwell, D., Breiman, L., and Thomasian, A. The capacities of certain channel classes under random coding. *Annals of Mathematical Statistics 31*, 3 (1960), 558–567.

[29] Blinovsky, V., Narayan, P., and Pinsker, M. Capacity of the arbitrarily varying channel under list decoding. *Problems of Information Transmission 31*, 2 (1995), 99–113.

[30] Blinovsky, V., and Pinsker, M. Estimation of the size of the list when decoding over an arbitrarily varying channel. In *Proceedings of 1st French-Israeli Workshop on Algebraic Coding* (Berlin, July 1993), G. Cohen, S. Litsyn, A. Lobstein, and G. Zémor, Eds., no. 781 in Lecture Notes in Computer Science, Springer-Verlag.

[31] Borden, J., Mason, D., and McEliece, R. Some information-theoretic saddlepoints. *SIAM Journal on Control and Optimization 23* (1985), 129–143.

[32] Boyd, S., and Vandenberghe, L. *Convex Optimization.* Cambridge University Press, Cambridge, UK, 2004.

[33] Breiman, L., Blackwell, D., and Thomasian, A. Proof of Shannon's transmission theorem for finite-state indecomposable channels. *Annals of Mathematical Statistics 29*, 4 (1958), 1209–2220.

[34] Brooks, R., Ramanathan, P., and Sayeed, A. Distributed target classification and tracking in sensor networks. *Proceedings of the IEEE 91*, 8 (August 2003), 1163–1171.

[35] Bross, S., and (Shitz), S. S. Capacity and decoding rules for the Poisson arbitrarily varying channel. *IEEE Transactions on Information Theory 49*, 11 (2003), 3076–93.

[36] Chen, J., and Berger, T. The capacity of finite-state Markov channels with feedback. *IEEE Transactions on Information Theory 51*, 3 (March 2005), 780–798.

[37] Chong, C.-Y., and Kumar, S. Sensor networks: evolution, opportunities, and challenges. *Proceedings of the IEEE 91*, 8 (August 2003), 1247–1256.

[38] Cohen, A., and Lapidoth, A. The Gaussian watermarking game. *IEEE Transactions on Information Theory 48*, 6 (2002), 1639–1667.

[39] Costa, M. Writing on dirty paper. *IEEE Transactions on Information Theory IT-29*, 3 (May 1983), 439–441.

[40] Cover, T. Broadcast channels. *IEEE Transactions on Information Theory 18*, 1 (1972), 2–14.

[41] Cover, T., and Thomas, J. *Elements of Information Theory.* Wiley, New York, 1991.

[42] Csiszár, I. Arbitrarily varying channels with general alphabets and states. *IEEE Transactions on Information Theory 38*, 6 (1992), 1725–1742.

[43] Csiszár, I., and Körner, J. On the capacity of the arbitrarily varying channel for maximum probability of error. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57*, 1 (1981), 87–101.

[44] Csiszár, I., and Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Akadémi Kiadó, Budapest, 1982.

[45] Csiszár, I., and Narayan, P. Arbitrarily varying channels with constrained inputs and states. *IEEE Transactions on Information Theory 34*, 1 (1988), 27–34.

[46] Csiszár, I., and Narayan, P. The capacity of the arbitrarily varying channel revisited : Positivity, constraints. *IEEE Transactions on Information Theory 34*, 2 (1988), 181–193.

[47] Csiszár, I., and Narayan, P. Capacity and decoding rules for classes of arbitrarily varying channels. *IEEE Transactions on Information Theory 35*, 4 (1989), 752–769.

[48] Csiszár, I., and Narayan, P. Capacity of the Gaussian arbitrarily varying channel. *IEEE Transactions on Information Theory 37*, 1 (1991), 18–26.

[49] Dembo, A., and Zeitouni, O. *Large Deviations Techniques and Applications.* Springer, New York, 1998.

[50] Devroye, N., Mitran, P., and Tarokh, V. Achievable rates in cognitive radio channels. *IEEE Transactions on Information Theory 52*, 5 (May 2006), 1813–1827.

[51] Dobrushin, R. Optimum information transmission through a channel with unknown parameters. *Radiotekhnika i Electronika 4*, 1951–1956 (1959).

[52] Dobrushin, R., and Stambler, S. Coding theorems for classes of arbitrarily varying discrete memoryless channels. *Problems of Information Transmission 11*, 2 (1975), 97–112.

[53] Draper, S., Frey, B., and Kschischang, F. Rateless coding for non-ergodic channels with decoder channel state information. Submitted to IEEE Transactions of Information Theory.

[54] Draper, S., Frey, B., and Kschischang, F. Efficient variable length channel coding for unknown DMCs. In *Proceedings of the 2004 International Symposium on Information Theory* (Chicago, USA, 2004).

[55] Elias, P. List decoding for noisy channels. In *Wescon Convention Record, Part 2* (1957), Institute of Radio Engineers (now IEEE), pp. 94–104.

[56] Erdös, P., Frankl, P., and Füredi, Z. Families of finite sets in which no set is covered by the union of r others. *Israel Journal of Mathematics 51*, 1–2 (1985), 79–89.

[57] Erez, U., Shamai (Shitz), S., and Zamir, R. Capacity and lattice strategies for canceling known interference. *IEEE Transactions on Information Theory 51*, 11 (November 2005), 3820–3833.

[58] Erez, U., Trott, M., and Wornell, G. Rateless coding for Gaussian channels. arXiv:0708.2575v1 [cs.IT], August 2007.

[59] Ericson, T. Exponential error bounds for random codes on the arbitrarily varying channel. *IEEE Transactions on Information Theory 31*, 1 (1985), 42–48.

[60] Estrin, D., Culler, D., Pister, K., and Sukhatme, G. Connecting the physical world with pervasive networks. *IEEE Pervasive Computing 1*, 1 (January–March 2002), 59–69.

[61] Eswaran, K., Sarwate, A., Sahai, A., and Gastpar, M. Binary additive channels with individual noise sequences and limited active feedback. In *Proceedings of the 2007 IEEE International Symposium on Information Theory* (Nice, France, 2007).

[62] Eswaran, K., Sarwate, A., Sahai, A., and Gastpar, M. Limited feedback achieves the empirical capacity. Submitted to IEEE Transactions of Information Theory, November 2007.

[63] Federal Communications Commission. *Spectrum Policy Task Force Report.* ET Docket No. 02-155, November 2, 2002.

[64] Federal Communications Commission. *Auction of 700 MHz Band Licenses Scheduled for January 24, 2008; Notice and Filing Requirements, Minimum Opening Bids, Reserve Prices, Upfront Payments, and Other Procedures for Auctions 73 and 76.* Public Notice (DA 07-4171), October 5, 2007.

[65] Feller, W. *An Introduction to Probability Theory and Its Applications.* John Wiley and Sons, Inc., New York, 1968.

[66] Forney, G. Exponential error bounds for erasures, list, and decision feedback schemes. *IEEE Transactions on Information Theory 14*, 2 (March 1968), 206–220.

[67] Gallager, R. *Information Theory and Reliable Communication.* John Wiley and Sons, New York, 1968.

[68] Gallager, R. Capacity and coding for degraded broadcast channels. *Problems of Information Transmission*, 185–193 (July-September 1974).

[69] Gastpar, M. On capacity under receive and spatial spectrum-sharing constraints. *IEEE Transactions on Information Theory 53*, 2 (February 2007), 471–487.

[70] Gel'fand, S., and Pinsker, M. Coding for channel with random parameters. *Problems of Control and Information Theory 9*, 1 (1980), 19–31.

[71] Gharavi, H. Multichannel mobile ad hoc links for multimedia communications. *Proceedings of the IEEE 96*, 1 (January 2008), 77–96.

[72] Gubner, J. On the deterministic-code capacity of the multiple-access arbitrarily varying channel. *IEEE Transactions on Information Theory 36*, 2 (1990), 262–275.

[73] Gubner, J. State constraints for the multiple-access arbitrarily varying channel. *IEEE Transactions on Information Theory 37*, 1 (1991), 27–31.

[74] Gubner, J. On the capacity region of the discrete additive multiple-access arbitrarily varying channel. *IEEE Transactions on Information Theory 38*, 4 (1992), 1344–1347.

[75] Gubner, J., and Hughes, B. Nonconvexity of the capacity region of the multiple-access arbitrarily varying channel subject to constraints. *IEEE Transactions on Information Theory 41*, 1 (1995), 3–13.

[76] Guruswami, V. List decoding from erasures: Bounds and code constructions. *IEEE Transactions on Information Theory 49*, 11 (2003), 2826–2833.

[77] Guruswami, V., Håstad, J., Sudan, M., and Zuckerman, D. Combinatorial bounds for list decoding. *IEEE Transactions on Information Theory 48*, 5 (2002), 1021–1034.

[78] Guruswami, V., and Rudra, A. Explicit capacity-achieving list-decodable codes. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing (STOC)* (2006), pp. 1–10.

[79] Guruswami, V., and Rudra, A. Concatenated codes can achieve list decoding capacity. In *Proceedings of the 2008 Symposium on Discrete Algorithms (SODA)* (2008), pp. 258–267.

[80] Guruswami, V., and Sudan, M. Improved decoding of Reed-Solomon and algebraic-geometry codes. *IEEE Transactions on Information Theory 45*, 6 (1999), 1757–1767.

[81] Hegde, M., Stark, W., and Teneketzis, D. On the capacity of channels with unknown interference. *IEEE Transactions on Information Theory 35*, 4 (1989), 770–783.

[82] Hof, E., and Bross, S. On the deterministic-code capacity of the two-user discrete memoryless arbitrarily varying general broadcast channel with degraded message sets. *IEEE Transactions on Information Theory 52*, 11 (November 2006), 5023–5044.

[83] Huang, J., Berry, R., and Honig, M. Auction-based spectrum sharing. *ACM/Springer Journal of Mobile Networks and Applications (MONET) 11*, 3 (June 2006), 405–418.

[84] Hughes, B. Interleaving and the arbitrarily varying channel. *IEEE Transactions on Information Theory 37*, 2 (1991), 413–420.

[85] Hughes, B. The smallest list for the arbitrarily varying channel. *IEEE Transactions on Information Theory 43*, 3 (1997), 803–815.

[86] Hughes, B., and Narayan, P. Gaussian arbitrarily varying channels. *IEEE Transactions on Information Theory 33*, 2 (1987), 267–284.

[87] Hughes, B., and Narayan, P. The capacity of a vector Gaussian arbitrarily varying channel. *IEEE Transactions on Information Theory 34*, 5 (1988), 995–1003.

[88] Hughes, B. L., and Thomas, T. G. On error exponents for arbitrarily varying channels. *IEEE Transactions on Information Theory 42*, 1 (1996), 87–98.

[89] Jahn, J. Coding of arbitrarily varying multiuser channels. *IEEE Transactions on Information Theory 27*, 2 (1981), 212–226.

[90] Jovičić, A., and Viswanath, P. Cognitive radio: An information-theoretic perspective. Submitted to the IEEE Transactions on Information Theory, April 2006.

[91] Kambo, N., and Singh, S. The capacities of certain special channels with arbitrarily varying channel probability functions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 29*, 4 (1974), 331–344.

[92] Kashyap, A., Başar, T., and Srikant, R. Correlated jamming on MIMO Gaussian fading channels. *IEEE Transactions on Information Theory 50*, 9 (2–4), 2119–2123.

[93] Kelly, F. Charging and rate control for elastic traffic. *European Transactions on Telecommunication 8* (1997), 33–37.

[94] Kelly, F., Maulloo, A., and Tan, D. Rate control in communcation networks : shadow prices, proportional fairness and stability. *Journal of the Operations Research Society 49* (1998), 237–252.

[95] Kiefer, J., and Wolfowitz, J. Channels with arbitrarily varying channel probability functions. *Information and Control 5*, 1 (1962), 44–54.

[96] Koetter, R., and Vardy, A. Algebraic soft-decision decoding of Reed-Solomon codes. *IEEE Transactions on Information Theory 49*, 11 (2003), 2809–2825.

[97] Körner, J., and Orlitsky, A. Zero-error information theory. *IEEE Transactions on Information Theory 44*, 10 (October 1998).

[98] Kramer, G., Berry, R., El Gamal, A., El Gamal, H., Franceschetti, M., Gastpar, M., and Laneman, J. Introduction to the special issue on models, theory, and codes for relaying and cooperation in communication networks. *IEEE Transactions on Information Theory 53*, 10 (October 2007).

[99] Kramer, G., Marić, I., and Yates, R. Cooperative communications. *Foundations and Trends in Networking 1*, 3-4 (August 2008).

[100] La, R., and Anantharam, V. A game-theoretic look at the Gaussian multiaccess channel. In *DIMACS Workshop on Network Information Theory*, P. Gupta, G. Kramer, and A. van Winjgaarden, Eds., vol. 66 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. AMS DIMACS, Piscataway, NJ, March 17–19 2003, pp. 87–106.

[101] Langberg, M. Private codes or succinct random codes that are (almost) perfect. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2004)* (Rome, Italy, 2004).

[102] Lapidoth, A. Nearest neighbor decoding for additive non-Gaussian noise channels. *IEEE Transactions on Information Theory 42*, 5 (1996), 1520–1529.

[103] Lapidoth, A. On the role of mismatch in rate distortion theory. *IEEE Transactions on Information Theory 43*, 1 (January 1997), 38–47.

[104] Lapidoth, A., and Narayan, P. Reliable communication under channel uncertainty. *IEEE Transactions on Information Theory 44*, 10 (1998), 2148–2177.

[105] Leonard, N., Paley, D., Lekien, F., Sepuchre, R., Fratantoni, D., and Davis, R. Collective motion, sensor collective motion, sensor networks, and ocean sampling. *Proceedings of the IEEE 95*, 1 (January 2007), 48–74.

[106] Lovász, L. On the Shannon capacity of a graph. *IEEE Transactions on Information Theory 1* (1979), 1–7.

[107] Luby, M. LT codes. In *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science* (2002), p. 271.

[108] Mallik, R., Scholtz, R., and Papavassilopolous, G. Analysis of an on-off jamming situation as a dynamic game. *IEEE Transactions on Communications 48*, 8 (2000), 1360–1373.

[109] Marton, K. A coding theorem for the discrete memoryless broadcast channel. *IEEE Transactions on Information Theory 25*, 3 (May 1979), 306–311.

[110] Mathur, S., Shankar, L., and Mandayam, N. Coalitions in cooperative wireless networks. *IEEE Journal on Selected Areas in Communications (To appear)* (2008).

[111] McEliece, R. *Secure Digital Communications*. Springer-Verlag, New York, 1983, ch. Communications in the presence of jamming – An information-theoretic approach, pp. 127–166.

[112] McEliece, R., and Stark, W. An information theoretic study of communication in the presence of jamming. In *Proceedings of the 1981 IEEE International Conference on Communications* (1981), pp. 45.3.1–45.3.5.

[113] Médard, M. Capacity of correlated jamming channels. In *Proceedings of the 1997 Allerton Conference on Communications, Computing and Control* (University of Illinois, Oct. 1997).

[114] Mitola, J., and Maguire, G. Cognitive radio: Making software radios more personal. *IEEE Personal Communications 6*, 4 (August 1999), 13–18.

[115] Mitola III, J. *Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio*. PhD thesis, Royal Institute of Technology (KTH), Stockholm, Sweden, 2000.

[116] Moulin, P., and Wang, Y. Capacity and random-coding exponents for channel coding with side information. *IEEE Transactions on Information Theory 53*, 4 (April 2007), 1326–1347.

[117] Nishimura, S. The strong converse theorem in the decoding scheme of list size *L*. *Kōdai Mathematical Seminar Reports 21*, 4 (1969), 418–425.

[118] Oh, S., Schenato, L., Chen, P., and Sastry, S. Tracking and coordination of multiple agents using sensor networks: system design, algorithms and experiments. *Proceedings of the IEEE 95*, 1 (January 2007), 234–254.

[119] Parvaresh, F., and Vardy, A. Correcting errors beyond the Guruswami-Sudan radius in polynomial time. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (2005), pp. 285–294.

[120] Permuter, H., Weissman, T., and Goldsmith, A. Finite state channels with time-invariant deterministic feedback. submitted to IEEE Transactions of Information Theory, arXiv:cs/0608070v1 [cs.IT], August 2006.

[121] Royer, E., and Toh, C. A review of current routing protocols for ad hoc mobile wireless networks. *IEEE Personal Communications 6*, 2 (April 1999), 46–55.

[122] Sahai, A. Balancing forward and feedback error correction for erasure channels with unreliable feedback. Submitted to IEEE Transactions of Information Theory.

[123] Sahai, A. Why block-length and delay behave differently if feedback is present. *IEEE Transactions on Information Theory 54*, 5 (May 2008). to appear.

[124] Sahai, A., and Draper, S. Beating the Burnashev bound using noisy feedback. In *Proceedings of the Allerton Conference on Communication, Control, and Computing* (Monticello, IL, Sept. 2006).

[125] Sarwate, A., and Gastpar, M. Randomization Bounds on Gaussian arbitrarily varying channels. In *Proceedings of the 2006 International Symposium on Information Theory* (Seattle, WA, 2006).

[126] Sarwate, A., and Gastpar, M. Randomization for robust communication in networks, or "Brother, can you spare a bit?". In *Proceedings of the 44th Annual Allerton Conference on Commununication, Control and Computation* (Monticello, IL, September 2006).

[127] Sarwate, A., and Gastpar, M. Channels with nosy "noise". In *Proceedings of the 2007 IEEE International Symposium on Information Theory* (Nice, France, June 2007).

[128] Sarwate, A., and Gastpar, M. Deterministic list codes for state-constrained arbitrarily varying channels. Submitted to the IEEE Transactions on Information Theory, arXiv:cs/0701146v2 [cs.IT], September 2007.

[129] Sarwate, A., and Gastpar, M. Rateless coding with partial CSI at the decoder. In *Proceedings of the 2007 Information Theory Workshop* (Lake Tahoe, CA, September 2007).

[130] Sarwate, A., and Gastpar, M. Rateless coding with partial state information at the decoder. Submitted to IEEE Transactions of Information Theory, arXiv:0711.3926v1 [cs.IT], November 2007.

[131] Sarwate, A., and Gastpar, M. Adversarial interference models for multi-antenna cooperative systems. In *Proceedings of the 42nd Annual Conference on Information Sciences and Systems (CISS 2008)* (Princeton, NJ, March 19–21 2008).

[132] Sarwate, A., and Gastpar, M. Arbitrarily dirty paper coding and applications. In *Proceedings of the 2008 IEEE International Symposium on Information Theory* (Toronto, Canada, 2008).

[133] Shafiee, S., and Ulukus, S. Mutual Information Games in Multi-user Channels with Correlated Jamming. Submitted to IEEE Transactions on Information Theory, Oct. 2005.

[134] Shamai (Shitz), S., and Verdú, S. Worst-case power-constrained noise binary-input channels. *IEEE Transactions on Information Theory 38*, 5 (1992), 1494–1511.

[135] Shannon, C. A mathematical theory of communication. *Bell System Technical Journal 27* (1948), 379–423, 623–656.

[136] Shannon, C. Communication theory of secret systems. *Bell System Technical Journal 28*, 4 (1949), 656–715.

[137] Shannon, C. The zero error capacity of a noisy channel. *IRE Transactions on Information Theory IT-2*, 6 (1956), 8–19.

[138] Shannon, C. Channels with side information at the transmitter. *IBM Journal of Research Developments 2* (October 1958), 289–293.

[139] Shannon, C. Probability of error for optimal codes in a Gaussian channel. *Bell System Technical Journal 38* (1959), 611–656.

[140] Shayevitz, O., and Feder, M. Achieving the empirical capacity using feedback part I: Memoryless additive models. Submitted to IEEE Transactions on Information Theory.

[141] Shokrollahi, A. Fountain codes. In *Proceedings of the 41st Allerton Conference on Communication, Control, and Computing* (October 2003), pp. 1290–1297.

[142] Shulman, N. *Communication over an Unknown Channel via Common Broadcasting*. PhD thesis, Tel Aviv University, 2003.

[143] Smith, A. Scrambling adversarial errors using few random bits, optimal information reconciliation, and better private codes. In *Proceedings of the 2007 ACM-SIAM Symposium on Discrete Algorithms (SODA 2007)* (2007).

[144] Soljanin, E., Liu, R., and Spasojevic, P. Hybrid ARQ with random transmission assignments. In *DIMACS Workshop on Network Information Theory*, P. Gupta, G. Kramer, and A. van Winjgaarden, Eds., vol. 66 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. AMS DIMACS, Piscataway, NJ, March 17–19 2003, pp. 321–334.

[145] Stambler, S. Shannon's theorems for a complete class of discrete channels whose state is known at the output. *Problems of Information Transmission 11*, 4 (1975), 263–270.

[146] Stiglitz, I. Coding for a class of unknown channels. *IEEE Transactions on Information Theory 12*, 2 (April 1966), 189–195.

[147] Sudan, M. Decoding of Reed-Solomon codes beyond the error-correction bound. *Journal of Complexity 13*, 1 (March 1997), 180–193.

[148] Tandra, R., and Sahai, A. SNR walls for signal detection. *IEEE Journal on Selected Topics in Signal Processing 2*, 1 (February 2008), 4–17.

[149] Tatikonda, S., Yang, S., and Kavčić, A. Feedback capacity of finite-state machine channels. *IEEE Transactions on Information Theory 51*, 3 (March 2005), 799–810.

[150] Tchamkerten, A., and Telatar, I. E. Variable length coding over an unknown channel. *IEEE Transactions on Information Theory 52*, 5 (May 2006), 2126–2145.

[151] Thomas, T. G., and Hughes, B. Exponential error bounds for random codes on Gaussian arbitrarily varying channels. *IEEE Transactions on Information Theory 37*, 3 (1991), 643–649.

[152] Tse, D., and Viswanath, P. *Fundamentals of Wireless Communication.* Cambridge University Press, Cambridge, UK, 2005.

[153] Tsfasman, M. A., and Vlăduţ, S. G. Geometric approach to higher weights. *IEEE Transactions on Information Theory 41*, 6 (1995), 1564–1588.

[154] Wei, V. K. Generalized Hamming weights for linear codes. *IEEE Transactions on Information Theory 37*, 5 (1991), 1412–1418.

[155] Weingarten, H., Steinberg, Y., and Shamai (Shitz), S. The capacity region of the Gaussian multiple-input multiple output broadcast channel. *IEEE Transactions on Information Theory 52*, 9 (September 2006), 3936–3964.

[156] Wozencraft, J. List decoding. Quarterly progress report, Research Laboratory of Electronics, MIT, 1958.

[157] Wyner, A. Random packing and converings of the unit n-sphere. *Bell System Technical Journal 46*, 9 (November 1967), 2111–2118.

[158] Yang, J. Spatial channel characterization for cognitive radios. Master's thesis, University of California, Berkeley, 2004.