

Rateless Coding for Non-Ergodic Channels with Decoder Channel State Information

Stark C. Draper, Brendan Frey, and Frank R. Kschischang

Abstract

Rateless coding has recently been the focus of much practical as well as theoretical research. In this paper we argue that rateless codes find a natural application in non-ergodic channels where the channel law varies unpredictably. Such unpredictability means that to guarantee reliability block-codes are limited by worst-case channel variations. However, the dynamic-decoding nature of rateless codes allows them to adapt opportunistically to the realized channel variations. If the channel state selector is not malicious, but also not predictable, decoding can occur earlier, producing a rate of communication that can be much higher than the worst case. Indeed, we argue that the application of fountain codes to the binary erasure channel can be understood as an example of these ideas – channel ergodicity is not required. Further, this sort of decoding can be usefully understood as an incremental form of erasure decoding. We show how to use ideas of erasure decoding to make significant increases in the reliability function of our scheme.

1 Introduction

Rateless codes are a type of incremental redundancy code¹. They do not operate at a pre-set rate. Rather, they use very long codewords and select their decoding time dynamically—to match the rate of communication to empirical channel behavior. When the channel can support a higher rate of communication decoding occurs sooner. When it is more noisy decoding occurs after a longer delay. The destination decides when to decode and uses a feedback channel to signal that it is time to terminate transmission. The termination signal is low-rate and delays and noise on the feedback link can be accommodated [9]. These characteristics make rateless strategies of great practical utility.

Rateless coding ideas have received much attention of late. In large measure this is due to the low-complexity encoding and decoding algorithms for LT and Raptor codes [17, 23]. These are rateless codes designed for binary erasure channels (BECs). Myriad applications from multicasting to layered video coding for wireless have been proposed based on LT and Raptor codes (see, e.g., [4]). The application of LT and Raptor codes to noisy channels, such as the binary-symmetric channel (BSC), the additive white Gaussian

S. C. Draper is with the Department of Electrical Engineering and Computer Science, University of California Berkeley, Berkeley, CA 94720 (E-mail: sdraper@eecs.berkeley.edu).

B. Frey is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4 (E-mail: frey@psi.toronto.edu).

F. R. Kschischang is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4 (E-mail: frank@comm.utoronto.edu).

¹Also known as incremental hybrid automatic repeat-request (HARQ) codes.

noise (AWGN) binary-input channel, and general symmetric channels is explored in [11, 19]. Through layered coding, the binary-input constraint for AWGN channels is avoided in [10]. The information theory of rateless coding for unknown but fixed discrete memoryless channels (DMCs), i.e., the compound channel, is explored in [24, 8, 27]. In source-coding parallel, rateless coding for various distributed source-coding scenarios with unknown joint statistics (e.g., Slepian-Wolf and Wyner-Ziv) is explored in [24, 7].

In this paper we extend the application of rateless coding ideas to non-ergodic channels. We develop rateless strategies for memoryless channels with an arbitrarily varying channel law where channel state information (CSI) is available at the receiver. At one level, our channel model is equivalent to an arbitrarily varying channel (AVC). The difference is that while we require communications to be reliable in the face of any channel variation (as for the AVC), we do not assume that the channel state selector necessarily tries to jam communications. We operate opportunistically: decoding sooner and realizing a higher rate of communication when the channel is not malicious. The rate of communication achieved can be designed to be at least as large as that which can be reliably achieved without feedback. The latter is given by the capacity of the AVC with receiver CSI [26]. Without feedback the block-length must be fixed and the rate picked small enough to make decoding reliable in the face of worst-case channel variations. In contrast, feedback enables the code to adapt its rate to actual channel variations.

While we do not put any probabilistic model on the channel variations, our results give insight into situations where the channel is ergodic but where there is a latency constraint on communication. A latency constraint may force the completion of communications before the system has seen average channel behavior. In this case, if the code rate is set to equal the ergodic channel capacity, there will be a probability of outage. If the coding rate is set equal to that which can be reliably supported under worst-case channel behavior, the communication rate will often be quite conservative. Rateless codes, while having an unpredictable decoding time, do not suffer outages and can often be decoded before worst-case block code designs. The use of incremental redundancy codes in wireless fading channels has been studied in, e.g., [25, 5, 21]. One message of this paper is that, from an information-theoretic perspective, rateless coding can be considered as an adaptive coding strategy for a general class of non-ergodic channels.

The outline of the paper is as follows. In Section 2 we define an information rate function to which we later give an operational meaning—it equals the rate of communication that can be reliably supported by our scheme. We compare this rate function to the communication rates possible in related communication contexts. We then build intuition behind our results and the stopping rule we use to determine when to terminate transmission. In Section 3 we give the basic definitions of our coding scheme and error events. In Section 4 we combine our stopping rule with a maximum likelihood (ML) decoding rule and show that the result is a reliable and efficient coding strategy that achieves the rate function defined in Section 2. The reliability function (error exponent) of ML decoding can be greatly increased by using the feedback link

to implement an erasure-type decoding rule. In Section 5 we extend our results to erasure decoding. We conclude in Section 6.

2 Information rate function and stopping rule

The channels we consider in this paper are defined by a finite family \mathcal{W} of discrete memoryless channels $W : \mathcal{X} \rightarrow \mathcal{Y}$ where \mathcal{X} and \mathcal{Y} are the finite input and output alphabets, respectively. Elements of the family $W \in \mathcal{W}$ are written as $W(y|x, s)$ where $y \in \mathcal{Y}$, $x \in \mathcal{X}$, and $s \in \mathcal{S}$ is the channel state identifying a particular channel in the family. The transition probabilities corresponding to a sequence of states $\mathbf{s} = s_1 \dots s_n$ are

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

The family of channels $W^n(\cdot|\cdot, \mathbf{s}) : \mathcal{X}^n \rightarrow \mathcal{Y}^n$, $\mathbf{s} \in \mathcal{S}^n$ is denoted as \mathcal{W}^n . Elements of this family are $W^n \in \mathcal{W}^n$. We put no probabilistic model or constraint on the state sequence.

We now define an information rate function that we later show has an operational meaning.

Definition 1 *Under any input distribution $p_x(x)$ on \mathcal{X} we define the information rate that can be supported by the channel W^n to be*

$$R(W^n, p_x) = \frac{1}{n} \sum_{i=1}^n I(x; y|s = s_i) = \sum_s \Lambda_n(s) I(x; y|s) = I(x; y|\mathbf{s}), \quad (1)$$

where $\Lambda_n(s)$ is the empirical distribution of the state sequence corresponding to W^n . That is, $\Lambda_n(s) = \sum_{i=1}^n \mathbf{1}\{s_i = s\}$ where $\mathbf{1}\{\cdot\}$ is the indicator function, taking on the value one if the argument is true, and zero otherwise. The mutual informations $I(x; y|s = s_i)$ and $I(x; y|\mathbf{s})$ are calculated with respect to the distributions $p_x(x)\mathbf{1}\{s = s_i\}W(y|x, s)$ and $p_x(x)\Lambda_n(s)W(y|x, s)$, respectively.

We will find all three equivalent forms of $R(W^n, p_x)$ given in (1) to be useful in the sequel.

The first thing to note about $R(W^n, p_x)$ is that because W^n is not known we cannot necessarily pick $p_x(x)$ to maximize $R(W^n, p_x)$. In general, while for a particular W^n one choice of $p_x(x)$ may maximize $R(W^n, p_x)$, for a different W^n the same $p_x(x)$ may not. We give an example of this in Section 2.2. In certain cases one choice will dominate. An example is when all the channels in \mathcal{W} are symmetric. For this example $p_x(x)$ should be chosen equal to the uniform distribution. A conservative choice is to pick $p_x(x)$ to maximize the worst-case $R(W^n, p_x)$ over all channels $W^n \in \mathcal{W}^n$. This choice connects to the AVC, and we discuss it further below.

2.1 Relation of $R(W^n, p_x)$ to other rate functions

We now connect (1) to rates that can be achieved in related communication contexts. We first compare it to the case when both transmitter and receiver have CSI. We then comment on the case when neither

transmitter nor receiver has CSI. Finally, we compare it to the capacity of the AVC with receiver CSI.

If both transmitter and receiver have CSI they can condition the code in use on the current value of the state. The system would use a state-dependent input distribution $p_{x|s}(x|s)$. The conditional distribution of x should be chosen to equal the capacity-achieving input for the channel indexed by state s . The information capacity of a channel $W^n \in \mathcal{W}^n$ with channel state information at the transmitter (CSIT—receiver channel state information is assumed) equals

$$C_{\text{CSIT}}(W^n) = \max_{p_{x|s}} I(x; y|s), \quad (2)$$

where now the mutual information is calculated with respect to $p_{x|s}(x|s)\Lambda_n(s)W(y|x, s)$. Each term $I(x; y|s = s)$ can be designed to be at least as large as each of the corresponding terms in (1). Therefore, a higher rate of communication can be achieved when state information is also known at the transmitter. The operational meaning of $C_{\text{CSIT}}(W^n)$ will parallel the operational meaning of $R(W^n, p_x)$ for channels with CSIT.

Next consider the case when neither transmitter nor receiver has channel CSI. In this case a decoder that is not tuned to the channel in use would have to be used. We are currently extending our results to this situation. By combining the universal decoding ideas for rateless coding over compound channels developed in [8] with decoding rules designed for AVCs we can build reliable and variable-length codes for this situation.

Finally, consider the AVC with decoder CSI. The capacity of this AVC is calculated by Stambler in [26]. As is discussed in [16], the basic insight into Stambler’s channel is that its capacity equals the capacity of the equivalent AVC $\tilde{W} : \mathcal{X} \times \mathcal{S} \rightarrow \mathcal{Y} \times \mathcal{S}$ defined by $\tilde{W}(y, s|x, s') = W(y|x, s)$ if $s' = s$, and $\tilde{W}(y, s|x, s') = 0$ if $s' \neq s$. Using this idea Stambler shows that the channel’s deterministic code capacity for average probability of error is

$$C_{\text{AVC}}(\mathcal{W}) = \max_{p_x} \min_{s \in \mathcal{S}} I(x; y|s = s), \quad (3)$$

i.e., it is equal to the capacity of the compound DMC over the family $\mathcal{W}(\cdot|s) : \mathcal{X} \rightarrow \mathcal{Y}, s \in \mathcal{S}$. The input distribution that maximizes (3) is denoted $p_{x, \text{AVC}}(x)$.

If we use the $p_{x, \text{AVC}}(x)$ that maximizes (3) in (1), the resulting $R(W^n, p_{x, \text{AVC}})$ will always be at least as great as (3):

$$R(W^n, p_{x, \text{AVC}}) = \sum_s \Lambda_n(s) I(x; y|s = s) \geq \min_{s \in \mathcal{S}} I(x; y|s = s). \quad (4)$$

Equality is achieved, e.g., by the distribution $\Lambda_n(s)$ that puts all its weight on the worst-case state. In other words, $\Lambda_n(s) = \mathbf{1}\{s = \arg \min_s I(x; y|s = s)\}$. The worst-case state is not necessarily unique. The example given in Section 2.2 illustrates this. If the channel state selector sometimes picks a non-worst-case state, the realized rate will be strictly greater than (3).

It is important to note that while the capacity of Stambler’s channel equals that of a particular compound DMC, operationally the channel is not a compound channel. If it were, a standard training-based scheme would work. We would use a known training sequence to estimate the channel at the decoder, feed back that

estimate (or some quantized version thereof), and follow up with a feedback-free block code to communicate the data. By using a long-enough training sequence, the unknown channel law could be learned to any desired degree of accuracy. Then, in a second phase, data could be communicated at a rate arbitrarily close to the capacity of the underlying channel. The cost of channel estimation (in channel uses) would be amortized over the subsequent (very long) data transmission phase. It would not, therefore, effect the achievable rate. It would, however, effect the exponent. This training plus block-coding scheme is variable-length and works for any compound DMC when feedback is available. However, it fails in the current context since the channel is not stationary. Therefore channel estimates that are derived during training do not predict future channel behavior. In the scheme’s follow-on block code, unless the transmission rate is below (3) we cannot guarantee reliability. In contrast, the coding strategy proposed in this paper adapts the rate of communication during data transmission to match the rate that the channel can support.

2.2 Example: A family of three channels

In this section we compare $R(W^n, p_x)$, $C_{\text{CSIT}}(W^n)$ and $C_{\text{AVC}}(\mathcal{W})$ for an illustrative family \mathcal{W} consisting of three channels—two binary asymmetric channels (BACs) and a BSC. The first channel in the family is the (higher-capacity) BAC defined by $\Pr[y = 1|x = 0] = 0.45$ and $\Pr[y = 0|x = 1] = 0.001$. The capacity of this channel is 0.357 bits per symbol. The second channel is the (lower-capacity) BAC defined by $\Pr[y = 1|x = 0] = 0.03$ and $\Pr[y = 0|x = 1] = 0.45$. The capacity of this channel is 0.279 bits per symbol. The third channel is the BSC that has a cross-over probability equal to 0.2. The capacity of this channel is 0.278 bits per symbol.

In Figure 1 we plot the mutual informations $I(x; y|s = s_0)$ induced across each channel by the spectrum of possible choices of $p_x(x)$. The capacity-achieving input distributions are specified by $\Pr[x = 0] = 0.405$, $\Pr[x = 0] = 0.565$, and $\Pr[x = 0] = 0.5$, respectively. The input distribution $p_{x, \text{AVC}}(x)$ that achieves the AVC capacity, maximizing (3), is specified by $\Pr[x = 0] = 0.525$, giving $C_{\text{AVC}}(\mathcal{W}) = 0.2774$ bits per symbol. For this family $p_{x, \text{AVC}}(x)$ does not achieve the capacity of any of the channels individually. On the plot we label two capacity-achieving input distributions, $p_{x, \text{AVC}}(x)$ and $p_{x, \text{opt}}(x)$. The latter is an “optimistic” choice. It achieves the highest throughput that can supported under any state sequence if the channel selector always chooses the highest capacity channel (the BAC with capacity 0.357).

To illustrate our results we specify an empirical state distribution $\Lambda_n(s)$. We compare $R(W^n, p_x)$ for two choices of $p_x(x)$ to $C_{\text{CSIT}}(W^n)$ and $C_{\text{AVC}}(\mathcal{W})$. The empirical distribution of the state sequence is given by $\Lambda_n(1) = \lambda$, $\Lambda_n(2) = 0.5(1 - \lambda)$ and $\Lambda_n(3) = 0.5(1 - \lambda)$ where $0 \leq \lambda \leq 1$. The three states correspond to the higher-capacity BAC, the lower-capacity BAC, and the BSC, respectively. In Figure 2 we plot $C_{\text{CSIT}}(W^n)$ and $C_{\text{AVC}}(\mathcal{W})$ as a function of λ . We also plot $R(W^n, p_{x, \text{AVC}})$ and $R(W^n, p_{x, \text{opt}})$.

The example illustrates two main points. The first is that feedback increases the reliably attainable

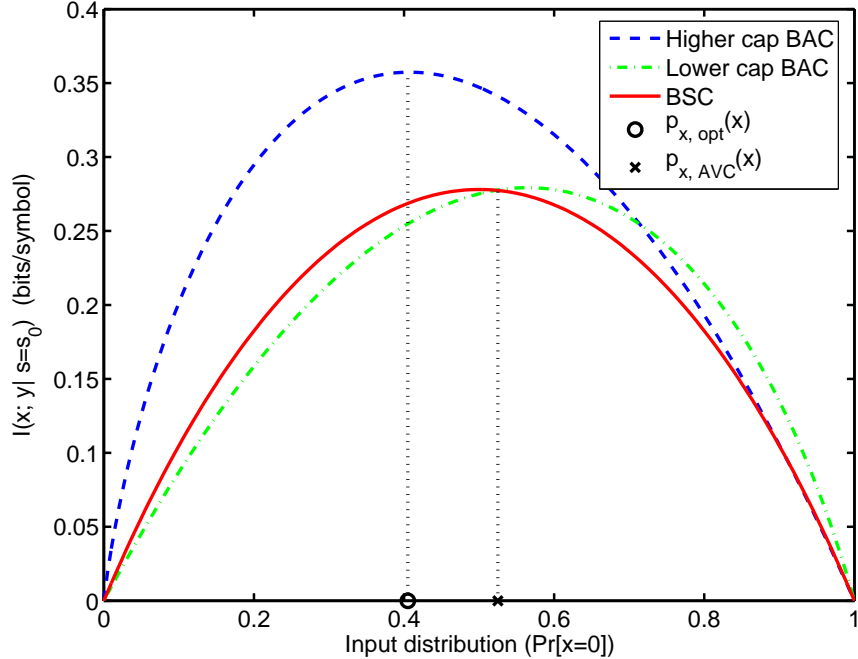


Figure 1: The mutual information induced across each channel in \mathcal{W} by different choices of $p_x(x)$. The capacity-achieving input distributions are indicated for the higher-capacity BAC, $p_{x,\text{opt}}(x)$, and for the AVC with decoder side information defined by the same family of channels, $p_{x,\text{AVC}}(x)$.

throughput of the channel. Additionally, it illustrates the fact that a single input distribution does not necessarily dominate for all choices of λ . To see the first point note that $R(W^n, p_{x,\text{AVC}}) \geq C_{\text{AVC}}(\mathcal{W})$, with equality only when $\lambda = 0$. When $\lambda = 0$ no symbols go through the higher-capacity BAC, half go through the lower-capacity BAC, and half through the BSC. Referring to Figure 1 we see that under $p_{x,\text{AVC}}(x)$ these latter two channels can both support the same rate, which equals $C_{\text{AVC}}(\mathcal{W})$. This means that for any $\lambda > 0$ the rate achieved is strictly larger than the worst case. This is true for any state distribution $\Lambda_n(s)$ such that $\Lambda_n(1) \neq 0$.

Next we see that one input distribution does not dominate. From Figure 1 we can infer that $p_x(x) = p_{x,\text{AVC}}(x)$ is the uniquely best worst-case choice. For all other choices of $p_x(x)$ there is a W^n that leads to a lower $R(W^n, p_x)$. However, for many channels W^n the rate function $R(W^n, p_{x,\text{AVC}})$ is exceeded by other choices of $p_x(x)$. As one example, in the figure we plot $R(W^n, p_{x,\text{opt}})$ which is larger than $R(W^n, p_{x,\text{AVC}})$ for $\lambda > 0.5$. Since $p_{x,\text{AVC}}(x)$ leads to the largest worst-case capacity, but the rate it can support is below that which can be supported by other choices of $p_x(x)$ for certain channels, no single input distribution dominates.

Finally we point out that for $\lambda = 0$, $C_{\text{CSIT}}(W^n) > C_{\text{AVC}}(\mathcal{W})$. This is because $p_{x,\text{AVC}}(x)$ is not capacity-achieving for either the low-capacity BAC nor for the BSC.

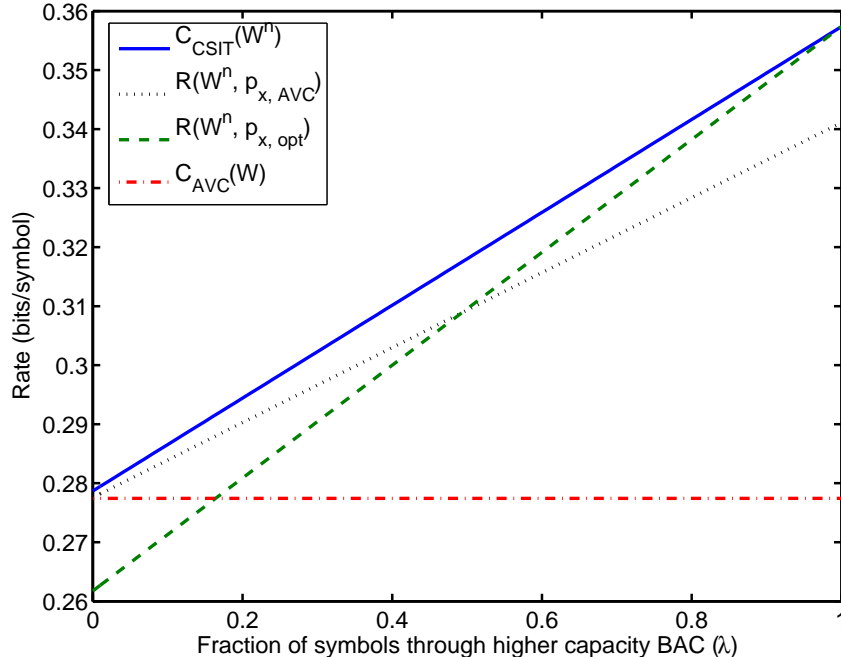


Figure 2: The communication rates of different coding scenarios as a function of the distribution of the state sequence. A fraction λ of the symbols go through the higher capacity BAC, a fraction $0.5(1-\lambda)$ through the lower-capacity BAC, and the remaining fraction $0.5(1-\lambda)$ through the BSC. The first communication scenario is when both encoder and decoder have channel state information, $C_{\text{CSIT}}(W^n)$. The second is when only the decoder has channel state information (plotted both for optimistic, $p_{x,\text{opt}}(x)$, and conservative, $p_{x,\text{AVC}}(x)$, choices of input distribution). The third is the capacity of the arbitrarily varying channel, $C_{\text{AVC}}(W)$.

2.3 Stopping rule for the BEC and beyond

The coding strategies we develop are most naturally understood by considering rateless coding over the binary erasure channel. In this case the encoding and decoding rules have simple interpretations. Say that each parity bit transmitted is a random linear combination of the information bits. Then, each parity that is not erased define one equation in the $\log M$ information bits, where M is the (fixed) number of messages. Roughly, when the number of parities received exceeds the number of information bits, this set of equations is invertible and decoding is possible. If the channel erasure probability is not known a priori, then how long one must wait for decoding, i.e., the effective block length, is also not known. With high likelihood the set of equations will be invertible, and so transmission can stop, at first time n that

$$\log M \leq (1 - \alpha) \sum_{i=1}^n [1 - \mathbf{1}\{y_i = \epsilon\}], \quad (5)$$

where y_i is the observation at time i and $y_i = \epsilon$ indicates that the i th bit has been erased. The constant $0 < \alpha < 1$ makes sure that decoding doesn't occur until the number of parities received unerased equals the number of information bits, and then allows for a few extra so that with high probability the set of equations is invertible.

Now consider a non-ergodic erasure channel where the erasure pattern does not follow any stationary probabilistic model. If parities are generated independently and according to the same rule (e.g., random linear combinations), each is equally likely to be useful to the decoder. Thus, the rate at which the decoder accumulates parities is not important. Rather, it is simply the total number of parities received that should determine whether the set of equations is invertible. Therefore the stopping rule (5) is still a good one for non-ergodic erasure channels. It was this observation that motivated our work in this paper that extends this setting to more general families of non-ergodic channels.

When transmitting over the BEC the receiver causally acquires channel state information. Parities received erased have gone through a channel with erasure probability one, and those received unerased have gone through a noiseless channel. This perspective leads us to rewrite (5) as

$$\log M \leq (1 - \alpha) \sum_{i=1}^n I(x; y | \mathbf{s} = s_i), \quad (6)$$

where the i th state s_i indexes the channel in use at time i and $I(x; y | \mathbf{s} = s_i)$ is the mutual information induced across the channel in use at time i by the chosen input distribution $p_x(x)$. As discussed above, since the transmitter does not know the state of the channel, the input is not a function of the state so $p_{x|s}(x|s) = p_x(x)$. For the BEC the input distribution $p_x(x)$ should be uniform. If the BEC is in the erasing state at time i then $I(x; y | \mathbf{s} = s_i) = 0$, while it equals one if the channel is noiseless. Substituting these values into (6) leads to the equivalent expression (5).

The expression (6) is not specific to erasure channels. We show that for any family of discrete memoryless channel laws a good choice of stopping time n (i.e., one that achieves a communication rate approximately equal to $R(W^n, p_x)$) is the first n such that (6) is satisfied. This n is well defined for any set of channels because mutual information is positive—the right-hand-side of (6) is monotonically increasing in n and therefore such an n exists.

If we group the time indices in (6) that correspond to a single state we get the equivalent stopping rule

$$\log M \leq (1 - \alpha) \sum_s n \Lambda_n(s) I(x; y | \mathbf{s} = s). \quad (7)$$

As before, $\Lambda_n(s)$ is the empirical distribution (type) of \mathbf{s} . Therefore $n\Lambda_n(s)$ is the number of symbols that have passed through the channel indexed by $\mathbf{s} = s$. Since decoding did not occur at time $n - 1$ the stopping rule also implies that

$$\log M > (1 - \alpha) \sum_s (n - 1) \Lambda_{n-1}(s) I(x; y | \mathbf{s} = s).$$

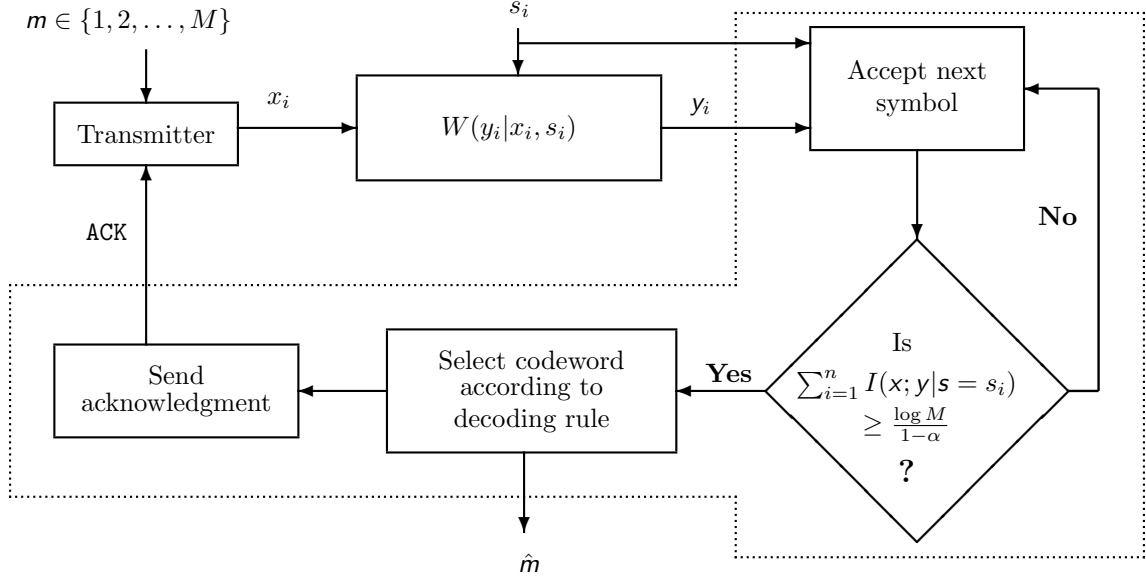


Figure 3: Block diagram of communication system, and decoding algorithm. The operations carried out by the decoder are enclosed within the dotted region.

Incorporating the empirical distributions $\Lambda_n(s)$ and $\Lambda_{n-1}(s)$ into the conditional mutual information terms gives

$$(1 - \alpha) \frac{n-1}{n} R(W^{n-1}, p_x) = \frac{n-1}{n} (1 - \alpha) I(\tilde{x}; \tilde{y} | \tilde{s}) < \frac{\log M}{n} \leq (1 - \alpha) I(x; y | s) = (1 - \alpha) R(W^n, p_x), \quad (8)$$

where the mutual informations are calculated with respect to the joint distributions $p_{\tilde{x}, \tilde{y}, \tilde{s}}(x, y, s) = p_x(x) \Lambda_{n-1}(s) W(y|x, s)$ and $p_{x, y, s}(x, y, s) = p_x(x) \Lambda_n(s) W(y|x, s)$, respectively.

As a simple example of the decoding rule (7) consider a family \mathcal{W} of BSCs, each defined by a cross-over probability p_s where $s \in \mathcal{S}$. For this family $p_x(x)$ uniform is optimal as it is capacity-achieving for each channel in the family. The decoding rule (7) tells us to decode at the first n such that

$$\frac{\log M}{n} \leq (1 - \alpha) \left[1 - \sum_s \Lambda_n(s) H_B(p_s) \right], \quad (9)$$

where $H_B(\cdot)$ is the binary entropy function. By selecting $\alpha > 0$ small enough, the rate of communication can be made arbitrarily close to the average capacity of the family of channels.

2.4 Discussion

The rate adaptive strategies we propose offer no benefit in the case of an AVC. In that setting the state selector is malicious (it is a jammer) and it would pick the state sequence to give an equality in (4). This is why block coding sufficed in previous papers that considered arbitrarily varying channel laws with feedback

(see, e.g., [1] and [3]). However, in situations where the channel varies arbitrarily, but not necessarily maliciously, our approach will dominate. In the following sections we show that codes exist that are robust to worst-case variations and adapt in an opportunistic manner—decoding in less time when the channel is good, and taking more time to decode when the channel is bad.

The rate-adaptive coding philosophy we advocate has been proposed as a means of dealing with the challenge of fluctuating channel conditions in wireless applications. In that context a number of researchers have considered the use of incremental redundancy coding in wireless environments. Those papers have mainly focused on the Gaussian channel with fading. They consider both the information-theoretic performance of hybrid-ARQ in the Gaussian collision channel [5], and the analysis of incremental transmission performance of various families of codes. As a small sampling of the latter, in [18, 20] punctured turbo codes are considered, while LDPC ensembles are considered in [21] and [25] where the latter also considers Raptor codes. Another work related in spirit is [22] where the authors generalize a scheme of Horstein’s [15] to specify a feedback-dependent coding strategy for a binary channel with arbitrary noise sequence. The rate of the scheme approaches $1 - H_B(p_{emp})$ where p_{emp} is the empirical distribution of the noise sequence. Comparison with (9) shows that when specialized to an analogous case, the gain from having channel state information at the decoder is through the concavity of entropy since $1 - \sum_s \Lambda_n(s) H_B(p_s) \geq 1 - H_B(\sum_s \Lambda_n(s) p_s)$.

3 Code definitions, feedback, and error events

In this section we formally define rateless codes, the feedback channel, and the error events. In terms of the feedback link, we assume the availability of a noiseless feedback channel with input and output alphabets \mathcal{Z} . Rateless codes use the feedback channel only for very simple signaling, effectively just to inform the transmitter when to stop sending. For simplicity of exposition we assume the feedback channel is available after each channel use. However, in application it needs be available for use only much more rarely (see Section 4.4).

The class of variable-length codes considered in this paper are specified by a finite message set $\mathcal{M} = \{1, 2, \dots, M\}$, a sequence of encoding functions:

$$f = \{f_n : \mathcal{M} \times \mathcal{Z}^{n-1} \rightarrow \mathcal{X}\}_{1 \leq n \leq N}, \quad (10)$$

and a sequence of decoding functions:

$$g = \{g_n : \mathcal{Y}^n \times \mathcal{S}^n \times \mathcal{Z}^{n-1} \rightarrow \mathcal{M} \cup \{0\}\}_{1 \leq n \leq N}. \quad (11)$$

We use codes with a maximum block length N . Since most applications will have a maximum delay constraint, a maximum block length is a natural constraint. From a theoretical viewpoint a maximum block

length makes it easy to show a deterministic code exists. The definition of the decoding functions (11) indicates that the channel state is causally observed at the decoder.

We first consider protocols that use the feedback channel exclusively to signal the source whether transmission should continue or should cease. This requires a binary feedback signal $\mathcal{Z} = \{\text{ACK}, \text{NAK}\}$. As long as the destination outputs the 0 symbol the NAK symbol is fed back and transmission continues. At the first n such that $g_n(y^n) = \hat{m} \neq 0$, an ACK is sent back, and transmission terminates. The realized rate is $\log M/n$. For compactness, we denote the codeword transmitted up to time n , i.e., the concatenation of $f_0(m), f_1(m, z_1), \dots, f_{n-1}(m, z^{n-1})$ as $f^n(m)$. Note that we do not indicate the dependence of f^n on z^{n-1} since decoding at time n implies that the destination has not ACKed the source until time n .

At each time n the decoding region of the m th codeword is $\mathcal{R}_n(m) = g_n^{-1}(m) \subseteq |\mathcal{Y}|^n$ where $m \in \{1, 2, \dots, M\}$. Different decoding regions are disjoint, $\mathcal{R}_n(m) \cap \mathcal{R}_n(\tilde{m}) = \emptyset$ for $m \neq \tilde{m}$. The zeroth decoding region $\mathcal{R}_n(0) = \{\mathbf{y} \text{ s.t. } \mathbf{y} \notin \cup_{m \in \mathcal{M}} \mathcal{R}_n(m)\}$. If $\mathbf{y} \in \cup_{m \in \mathcal{M}} \mathcal{R}_n(m)$, then $g_n(\mathbf{y}) = \hat{m} \neq 0$, and we send the ACK message. If $\mathbf{y} \in \mathcal{R}_n(0)$, then $g_n(\mathbf{y}) = 0$, and the destination waits for the next decoding opportunity before revisiting the decision whether to decode. A core difference between ML and this decoding rule is that in the former $\cup_{m \in \mathcal{M}} \mathcal{R}_n(m) = |\mathcal{Y}|^n$ while for this rule it need not be.

We now give definitions of a random code, akin to those of Csiszár and Körner [6, pg. 209], appropriately modified for the current context. We use $\mathcal{C}(f, g)$ to indicate the family of pairs of mappings as defined by (10) and (11). A random code is a random variable taking values in this family. If this random variable, (f, g) , has distribution Q , then for any decoding time $n \leq N$ and channel $W^n \in \mathcal{W}^n$, a random code defines a channel $T_n : \mathcal{M} \rightarrow \mathcal{M} \cup \{0\}$ by

$$T_n(\tilde{m}|m) = \sum_{(f,g)} Q(f, g) W^n(g_n^{-1}(\tilde{m})|f^n(m)).$$

If message $m \in \mathcal{M}$ were transmitted and the channel $W^n \in \mathcal{W}^n$ were realized, then because decoding regions are disjoint the probability of erroneous decoding at time n is

$$e_m(W^n, Q) = \begin{cases} 1 - T_n(0|m) - T_n(m|m) & \text{if } n < N, \\ 1 - T_N(m|m) & \text{if } n = N. \end{cases}$$

Note that decisions to continue transmitting, which occur whenever $\mathbf{y} \in \mathcal{R}_n(0)$, are not counted as errors. The major difference from the definitions given in [6] is that there is now a set of check times. Therefore, the stopping time t is not known a priori. The error probability is a function both of the stopping time and

the resulting decision

$$\begin{aligned}
e_{m,\text{tot}}(W^N, Q) &= \Pr[(y^1 \notin \mathcal{R}_1(m) \cup \mathcal{R}_1(0)) \cap t = 1] \cup \dots \cup (y^N \notin \mathcal{R}_N(m) \cap t = N) | W^N, m = m] \\
&\leq \sum_{n=1}^{N-1} \Pr[(y^n \notin \mathcal{R}_n(m) \cup \mathcal{R}_n(0)) \cap t = n | W^N, m = m] + \Pr[(y^N \notin \mathcal{R}_N(m)) \cap t = N | W^N, m = m] \\
&= \sum_{n=1}^N e_m(W^n, Q) \Pr[t = n | W^N, m = m]. \tag{12}
\end{aligned}$$

To ensure reliability we consider the worst case state sequence. Furthermore, in this paper we consider the average error probability criterion

$$e(W^N, Q) = \max_{W^N \in \mathcal{W}^N} \frac{1}{M} \sum_{m \in \mathcal{M}} e_{m,\text{tot}}(W^N, Q).$$

Because the decoding time is not pre-set, the realized rate is unknown a priori and may be random. We define the average rate in terms of the random stopping time t as

$$R = \frac{\log M}{E[t | s^N]}, \tag{13}$$

where the expectation is taken both with respect to the behavior of the mixture of channels defined by the state sequence and, in the case of a randomized code, with respect to the code ensemble.

4 Incremental transmission and ML decoding

In this section we prove the following result:

Theorem 1 *Let \mathcal{W} denote a family of channels indexed by $s \in \mathcal{S}$, a finite set, and let $p_x(x)$ be a distribution on \mathcal{X} such that $\min_{s_0} I(x, y | s = s_0) > 0$, where the mutual information is defined by the joint distribution $p_x(x) \mathbf{1}(s = s_0) W(y | x, s)$. Then, for every $0 < \alpha < 1$ and for every $\epsilon_{\max} > 0$ there exists a deterministic code of size M and maximum block-length $N > 0$ such that for all $W^N \in \mathcal{W}^N$ decoding occurs at some time $n \leq N$ such that*

1. *the realized rate is bounded as*

$$(1 - \alpha) \frac{n-1}{n} R(W^{n-1}, p_x) \leq \frac{\log M}{n} < (1 - \alpha) R(W^n, p_x).$$

where s_i is the state of the channel during the i th channel use, $R(\cdot, \cdot)$ is defined in (1), and

2. *for all $m \in \{1, 2, \dots, M\}$,*

$$e_{m,\text{tot}}(W^N, Q) \leq \epsilon_{\max}.$$

Remark: The assumptions that $\min_{s_0} I(x; y | s = s_0) > 0$ implies that transmission need never exceed a maximum block length N . Together with the finiteness of \mathcal{S} this make it simple to show the existence of a good deterministic code.

Remark: When there is a single capacity-achieving input distribution for all the channels in \mathcal{W} we can get $R(W^n, p_x) = C_{\text{CSIT}}(W^n)$, and we have a capacity result—no higher rate of communication is possible. This follows in a straightforward manner since the capacity of each of the constituent channels is achieved. Variable-length strategies are still important since we do not know ahead of time the fraction of time each channel will be in use, and thus cannot set a block-length. The BSC example of (9) is an example of this situation.

4.1 Decoding rule

We now specify our decoding rule by defining the decision regions $\mathcal{R}_n(m) =$. Recall that \mathbf{s} , \mathbf{y} , and \mathbf{x}_m are, respectively, the state sequence, the observation, and the m th codewords, up to time n .

Definition 2 For any $0 \leq \alpha \leq 1$ the decoding regions $\{\mathcal{R}_n(m)\}$ are defined in the following way:

$$\begin{aligned} \text{If } \log M > (1 - \alpha) \sum_{i=1}^n I(x; y | s = s_i) \text{ then } \mathcal{R}_n(m) = \emptyset \text{ for } m \in \mathcal{M} \text{ and } \mathcal{R}_n(0) = \mathcal{Y}^n, \\ \text{else } \mathcal{R}_n(m) = \left\{ \mathbf{y} \mid W^n(\mathbf{y} | \mathbf{x}_m, \mathbf{s}) \geq W^n(\mathbf{y} | \mathbf{x}_{\tilde{m}}, \mathbf{s}) \text{ for all } \tilde{m} \neq m \right\}. \end{aligned} \quad (14)$$

Note that we suppress the dependence of $\mathcal{R}_n(m)$ on the state sequence \mathbf{s} .

The decoding rule of (14) is the stopping rule of (6) combined with ML decoding. As discussed in Section 2 since the channel state is observed at the receiver, the receiver can calculate the rate the channel can support through time n . Until the communication rate is somewhat below this amount (the margin is controlled by α), the destination will not decode. Once the information rate is sufficiently low, it decodes using maximum likelihood.

4.2 Randomized coding performance

We now show that the decoder defined in (14) can give an arbitrarily small error probability for any state sequence. We calculate the random coding error probability using Gallager's ML decoding bounds [13]. The analysis is akin to Gallager's for parallel channels [13, pg. 149]. However, in the current setting we do not know a priori how many symbols will be transmitted along each channel. The code ensemble (defined by \mathcal{Q}) that we use is the standard ensemble consisting of length- N codewords generated independently and in an i.i.d. manner according to some $p_x(x)$.

The decoding time n defined by (14) is not a function of the statistical behavior of the channel. Rather, it is only a function of the empirical state sequence. This means that for a given state sequence s^N the

decoding time $n \leq N$ is deterministic. In other words $\Pr[t = n|W^N, m = m]$ equals one for one of the terms in (12) and zero for the rest. We start with Gallager's bound on blocks (before he uses the memoryless character of the channel). Assuming decoding happens at some time n , Gallager's bound [13, pg. 135] is:

$$e_{m,\text{tot}}(W^N, Q) \leq (M-1)^\rho \sum_{\mathbf{y}} \left[\sum_{\mathbf{x}} p_{\mathbf{x}}(\mathbf{x}) W^n(\mathbf{y}|\mathbf{x}, \mathbf{s})^{1/(1+\rho)} \right]^{1+\rho}.$$

where $e_{m,\text{tot}}(W^N, Q) = e_m(W^n, Q)$ for the time n such that $\Pr[t = n|W^N, m = m] = 1$. The parameter $\rho \in [0, 1]$ is Gallager's Chernoff parameter. Since the channel is memoryless the expression simplifies as

$$\begin{aligned} e_{m,\text{tot}}(W^N, Q) &\leq (M-1)^\rho \prod_{i=1}^n \left[\sum_{y_i} \left[\sum_{x_i} p_x(x_i) W(y_i|x_i, s_i)^{1/(1+\rho)} \right]^{1+\rho} \right] \\ &= (M-1)^\rho \prod_{s \in \mathcal{S}} \left[\sum_y \left[\sum_x p_x(x) W(y|x, s)^{1/(1+\rho)} \right]^{1+\rho} \right]^{n\Lambda_n(s)} \end{aligned} \quad (15)$$

$$\begin{aligned} &\leq \exp \left\{ -n \left[- \sum_s \Lambda_n(s) \log \sum_y \left[\sum_x p_x(x) W(y|x, s)^{1/(1+\rho)} \right]^{1+\rho} - \rho \log M/n \right] \right\} \\ &= \exp \left\{ -n \left[\sum_s \Lambda(s) E_s(\rho, p_x) - \rho \log M/n \right] \right\}, \end{aligned} \quad (16)$$

where $E_s(\rho, p_x)$ is the same as Gallager's $E_o(\rho, p_x)$ function. We have changed the subscript to denote the particular channel dependence. Note that the random coding performance in (16) is no longer a function of the particular state sequence \mathbf{s} , but rather its type $\Lambda_n(s)$. This is because all permutations of a given codeword are equally likely under an i.i.d. generating distribution.

The decoding rule constrains the decoding time to be such that $n > \log M / (1 - \alpha) \sum_s \Lambda_n(s) I(x; y|s = s)$ where we recall that $\Lambda_n(s)$ is the type of the state sequence up to time n . Using this in (16) and maximizing over ρ to get the tightest bound gives

$$e_{m,\text{tot}}(W^N, Q) \leq \exp \left\{ -n \max_{\rho \in [0,1]} \left[\sum_s \Lambda_n(s) [E_s(\rho, p_x) - \rho(1 - \alpha) I(x; y|s = s)] \right] \right\}. \quad (17)$$

Each of the terms $[E_s(\rho, p_x) - \rho(1 - \alpha) I(x; y|s = s)]$ in (17) is non-negative for ρ small enough (see discussion of (5.6.28) of [13]). Therefore, the argument of the exponent in (17) is negative and we can upper bound the error probability by using a second application of the lower bound on decoding time n . Defining

$$I(\Lambda) = \sum_s \Lambda(s) I(x; y|s = s), \quad (18)$$

where the dependence of $I(\Lambda)$ on $p_x(x)$ and $W(y|x, s)$ is understood from the context, gives

$$e_{m,\text{tot}}(W^N, Q) \leq \exp \left\{ - \frac{\log M}{(1 - \alpha) I(\Lambda_n)} \max_{\rho \in [0,1]} \left[\sum_s \Lambda_n(s) [E_s(\rho, p_x) - \rho(1 - \alpha) I(x; y|s = s)] \right] \right\}.$$

At this point, the analysis looks the same as if the channel states had been generated according to the distribution $p_s(s) = \Lambda_n(s)$. Of course, if the state sequence really had been generated according to $p_s(s)$,

then the channel would be ergodic, and block coding would suffice. The bound holds for all empirical state distributions, some yielding lower error probabilities than others.

The error probability $e_{m,\text{tot}}(W^N, Q)$ is only a function of the type $\Lambda_n(s)$, and not of the actual state sequence. We bound the error probability by maximizing over all state distributions (not necessarily types). Since the resulting bound holds for all $m \in \mathcal{M}$, it also holds for the average giving

$$e(W^N, Q) \leq \max_{\Lambda} \exp \left\{ -\frac{\log M}{(1-\alpha)I(\Lambda)} \max_{\rho \in [0,1]} \left[\sum_s \Lambda(s) [E_s(\rho, p_x) - \rho(1-\alpha)I(x; y|s = s)] \right] \right\}. \quad (19)$$

The argument of the exponent is negative and therefore it is monotonically decreasing in M . To guarantee that we meet some target error probability ϵ_{\max} uniformly over all possible state sequences, we choose the codebook size $|\mathcal{M}| = M^*$ large enough that

$$e(W^N, Q) \leq \max_{\Lambda} \exp \left\{ -\frac{\log M^*}{(1-\alpha)I(\Lambda)} \max_{\rho \in [0,1]} \left[\sum_s \Lambda(s) [E_s(\rho, p_x) - \rho(1-\alpha)I(x; y|s = s)] \right] \right\} < \epsilon_{\max}. \quad (20)$$

We now determine the maximum block length N . The mixture of channels $\Lambda(s)$ that results in the longest decoding time is just the channel with the lowest mutual information under $p_x(x)$, i.e., $\min_{s_0 \in \mathcal{S}} I(x, y|s = s_0)$. This is the rate that the compound channel defined by \mathcal{W} can support under input distribution p_x . If maximized over p_x , this would be the capacity of the compound channel [6]. With M^* constrained as in (20), we must select the maximum code-length N to satisfy

$$N = \left\lceil \frac{\log M^*}{(1-\alpha) \min_{s_0 \in \mathcal{S}} I(x; y|s = s_0)} \right\rceil. \quad (21)$$

This choice ensures that the message is eventually decoded, regardless of the state sequence. And, by the choice of M^* in (20), the probability of error is less than ϵ_{\max} .

4.3 Deterministic coding performance

In this section we specify a deterministic coding strategy. We have thus far bounded the worst-case average error probability over the ensemble of codes defined by the i.i.d. input distribution $p_x(x)$. Unlike the case of a fixed channel, we cannot argue that at least one code in the ensemble has performance equal to the ensemble average for all state sequences. While one code may be good for a given state sequence it might not be good for another (and there are exponentially many possible state sequences— $|\mathcal{W}|^N$).

One way to address this issue is to have a set of codebooks and, for each message, to pick a codebook to use randomly so that the probability of hitting a bad codebook for the realized state sequence is small. Implementing this choice requires a source of common randomness at encoder and decoder. One approach is to distribute a random seed with the codebook that informs the encoder and decoder which codebook to use for each data block. Distributing a seed with the codebook does not work for AVCs. This is because

the jammer would learn the seed. It would then condition its jamming signal on the codebook in use. One way to address this problem is to send a prefix code with the seed prior to data transmission (the jammer is assumed not to actively listen to and adapt to the transmission).

In our context the state selector is not malicious, so the seed could be distributed with the codebook. However, if a sequence of messages is to be sent, the size of the random seed would have to increase in proportion to the number of messages. If there is no bound on the number of messages we might use the code to send, we cannot bound the amount of randomness we would have to distribute a priori. An alternate approach is to distribute the seeds sequentially, on a per-message basis. We do this in a way that does not diminish the data communication rate.

We first use a “de-randomization” technique of Ahlswede to show that to get the ensemble performance over an arbitrarily varying channel one need not randomize over all block-codes in the ensemble. Rather, one can randomize uniformly over a much smaller set. Setting up the choice of which code to use to transmit each message will then require much less randomness. Ahlswede’s de-randomization technique is based on the fact that as long as the ensemble error probability is suitably small, the error performances of the codes in the ensemble concentrate about the mean. A derivation of his result is given in [6, Lemma 6.8]. We restate the result here:

Lemma 1 (Ahlswede) *Let \mathcal{W}^N be a finite family of channels and Q a probability distribution on $\mathcal{C}(f, g)$. Then for any ϵ_{\max} and L satisfying*

$$\epsilon > 2 \log(1 + e(\mathcal{W}^N, Q)), \quad L > \frac{2}{\epsilon} [\log |\mathcal{M}| + \log |\mathcal{W}^N|] \quad (22)$$

there exist L codes $(f[l], g[l]) \in \mathcal{C}(f, g)$ for $l = 1, \dots, L$, such that

$$\frac{1}{L} \sum_{l=1}^L e_{m, \text{tot}}(W^N, f[l], g[l]) < \epsilon \text{ for all } m \in \mathcal{M}, W^N \in \mathcal{W}^N, \quad (23)$$

where $e_{m, \text{tot}}(W^N, f[l], g[l])$ is defined as $e_{m, \text{tot}}(W^N, Q)$, but now for the particular code specified by $f[l]$ and $g[l]$, rather than for the ensemble Q .

To use Lemma 1 we first note that $e(\mathcal{W}^N, Q) \leq \epsilon_{\max}$ by (20). Turning to the number of bits of randomness that need to be shared between encoder and decoder to choose from the L codebooks, we have

$$\begin{aligned} \log L &> \log \frac{2}{\epsilon} (\log |\mathcal{M}| + N \log |\mathcal{W}|) \\ &\geq \log \frac{2}{\epsilon} \left(\log M^* + \frac{\log M^* \log |\mathcal{W}|}{(1 - \alpha) \min_{s_0 \in \mathcal{S}} I(x; y | s = s_0)} \right) \\ &= \gamma \log \log M^*. \end{aligned}$$

where $N = \log M^* / (1 - \alpha) \min_{s_0 \in \mathcal{S}} I(x; y | s = s_0)$ from (21), and γ is a constant in terms of ϵ and $\min_{s_0 \in \mathcal{S}} I(x; y | s = s_0)$. This tells us that we need not share more than $\gamma \log \log M^*$ nats between encoder

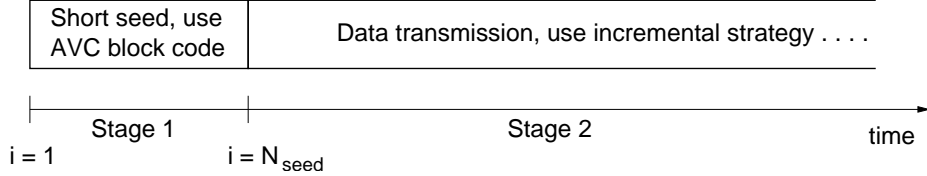


Figure 4: Common randomness can be established by using Stambler’s block code to transmit the seed to the decoder. A randomized variable-length strategy can then be used to make data transmission robust to channel variations.

and decoder to achieve the random coding results. In other words, there exist $\exp\{\gamma \log \log M^*\}$ codebooks such that if we uniformly choose one of these codebook to use for each transmission, the error probability is upper bounded by ϵ for all $m \in \mathcal{M}$.

The easiest way to establish the required common randomness at encoder and decoder is for the decoder to generate the random choice on its own. It could then use the noiseless feedback link to communicate that choice to the encoder. If the feedback channel cannot support this, or we don’t want to use it for this purpose, we describe an alternative strategy based on prefixing the data transmission with a short message indicating which codebook to use.

Ahlswede originally developed the type of prefixing strategy we use to show that the average-error deterministic-coding capacity of an AVC is either zero or it equals the random coding capacity of the channel (see, e.g. [6, Theorem 6.11, pg. 214]). Applied to our context, the encoder first uses Stambler’s feedback-free coding scheme to transmit a seed of $\gamma \log \log M$ nats to the decoder. The rate of this code is set to be just below $C_{\text{AVC}}(\mathcal{W})$. Since by the assumptions of Theorem 1 this rate is positive, we can successfully communicate the seed. Furthermore, in terms of the codebook size M , the amount of time it takes to transmit the common seed scales as

$$N_{\text{seed}} = \frac{\gamma \log \log M}{C_{\text{AVC}}(\mathcal{W})}.$$

However, by the decoding rule (14), the time to transmit the data through any mixture of channels scales as $\log M$. Thus the fraction of time spent sending the seed scales as $\log \log M / \log M$, which goes to zero with increasing M . Therefore, communicating the seed can be made to have a negligible effect on the overall rate, regardless of the state sequence. Following the prefix, the source transmits the data using the variable-length coding scheme implemented with the codebook indexed by the transmitted seed. This two-stage strategy is diagrammed in Figure 4.

4.4 Sparse Check Times

To this point we have assumed that the feedback channel is available after every channel use. We now show that analogous results hold even when the feedback channel is available only much more rarely. Instead

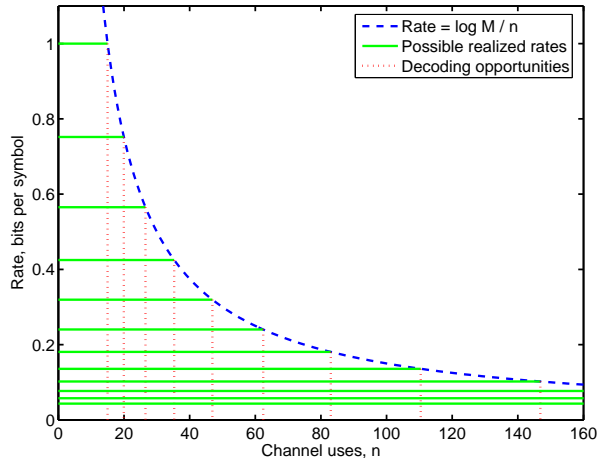


Figure 5: Decoding opportunities for a binary input alphabet and codebook size $M = 2^{15}$. In this example, the fractional rate decrement per decoding opportunity $\alpha = 1/3$. The possible realized rates $R_{n_1}, R_{n_2}, \dots, R_{n_{12}}$ and corresponding decoding opportunities are shown. In both cases the spacing of decoding opportunities becomes increasingly sparse as the rate drops, or as M is increased.

of checking the decoding rule (14) after each channel use, we check it only at a sparse set of times. We enumerate these decoding opportunities in ascending order as $n_1 < n_2 < \dots < n_K$, where $n_K = N$.

A set of check times that works for any family of channels \mathcal{W} and ensures communication efficiency $(1 - \alpha)$ is defined by setting

$$n_k = (1 - \alpha)^{k-1} \frac{\log M}{\min\{|\mathcal{X}|, |\mathcal{Y}|\}}. \quad (24)$$

An example of this choice is given in Figure 5. Note that decoding at time n_1 will only occur if the set \mathcal{W} includes a noiseless channel. Since many families of channels will not include a noiseless channel the first check can often be delayed. For the choice of check times (24) we solve for the total number of check times by substituting the definition for $n_K = N$ into (21) and solving for K , which gives

$$K = \left\lceil \frac{\log \frac{\min\{|\mathcal{X}|, |\mathcal{Y}|\}}{\min_{s_0 \in \mathcal{S}} I(x; y | s = s_0)}}{\log(1 - \alpha)} \right\rceil. \quad (25)$$

5 Incremental transmission and erasure decoding

In the last section we used a state-dependent stopping time and a ML decoding rule to show reliable communication can be achieved at a rate close to $R(W^n, p_x)$. The stopping time n was chosen only as a function of the state sequence and not as a function of the stochastic behavior of the channel. In this section we build upon the erasure decoding ideas of Forney [12] to give a paired stopping and decoding rule that is both state and channel-behavior dependent. The result is an improved error exponent.

In [12] Forney shows that when feedback is available an error exponent that far exceeds the sphere packing bound can be achieved if only the average block-length of the code is constrained—as opposed to the fixed block-length constraint of block coding. The sphere packing bound upper bounds the error exponent of block coding. It therefore bounds the reliability function of the scheme presented in Section 4 where the block length is the transmission duration n . To improve the reliability function of our scheme we want to apply these ideas to our setting.

In [12] the improvement in error exponent is made possible through the use of erasure decoding. In a nutshell, erasure decoding wraps an algorithm that attempts to detect errors (and asks for a retransmission if one is detected) around a regular block code. Information is first transmitted using the block code. The decoder calculates the likelihoods of each codeword. If the ML codeword is not enough more likely than the rest of the codewords (a threshold parameter controls what is meant by “enough”), instead of decoding the message, the decoder uses the feedback channel to request a retransmission. Implementing the erasure-decoding rule requires the communication of a single ACK/NAK bit. The decision not to decode is termed an erasure because after retransmission the decoder ignores the first block of received data, i.e., the decoder treats it as erased. In general, the design threshold controls the trade off between the probability of erasure and the probability of decoding error.

As an example, consider the BEC decoding rule given in (5). If the set of parity equations is not invertible, then there are at least two codewords that both have (equal) positive conditional probabilities (conditioned on the observations). In such a case an erasure decoding rule would request a retransmission. In reality, of course, we would use an incremental-erasure decoding rule that would simply wait for a few more parities to arrive before again trying to decode. However, as long as the probability of erasure can be made to decrease to zero as the block length is increased (which can be achieved, e.g., by any positive erasure exponent), the average number of transmitted blocks converges to one which implies that the average communication rate approaches that of the underlying block code. Therefore, ignoring previous blocks during (the rare) retransmissions does not noticeably effect the average rate.

The code structure is nearly the same as specified in Section 3. The only change to the protocol is that instead of only telling the sender whether to continue transmitting or to stop, the feedback message can also indicate a request for a retransmission. (These ideas can be combined with the rare check times discussed in Section 4.4.)

Theorem 2 *Let \mathcal{W} denote a family of channels indexed by $s \in \mathcal{S}$, a finite set, and let $p_x(x)$ be a distribution on \mathcal{X} such that $\min_{s_0} I(x; y | s = s_0) > 0$, where the mutual information is defined by the joint distribution $p_x(x)\mathbf{1}(s = s_0)W(y|x, s)$. Then, for every $0 < \alpha < 1$ and $\epsilon' > 0$ there exists a randomized code of size M such that for all $W^\infty \in \mathcal{W}^\infty$*

1. the expected communication rate is bounded as

$$\frac{\log M}{E[t|s^\infty]} \geq \frac{(1-\alpha)}{1+\epsilon'} R(W^{t_1-1}, p_x)$$

where t_1 is defined as the first time such that $\log M < (1-\alpha) \sum_{i=1}^{t_1} I(x; y|s = s_i)$ with s_i being the state of the channel during the i th channel use, $R(W^n, p_x)$ is defined in (1), and

2. the probability of error is upper bounded by

$$\max_{\Lambda} \exp \left\{ -\frac{\log M}{(1-\alpha)I(\Lambda)} \left[\max_{\beta \geq 1} \sum_s \Lambda(s) [E_{s,f}(\beta, p_x) - \beta(1-\alpha)I(x; y|s = s)] - \delta \right] \right\}$$

for all $\delta > 0$ where

$$E_{s,f}(\beta, p_x) = \sum_y \sum_x p_x(x) W(y|x, s) \left[\log W(y|x, s) - \log \left(\sum_{x'} p_x(x') W(y|x', s)^{1/\beta} \right)^\beta \right],$$

and $I(\Lambda)$ is defined in (18).

5.1 Bounding the expected decoding time

As in Section 4, the destination's first potential stopping time is the first n such that $\log M < (1-\alpha) \sum_{i=1}^n I(x; y|s = s_i)$. At that point, instead of necessarily decoding, the destination tries to detect whether the channel noise has been atypically large. If it has been, the destination requests a retransmission. When this occurs, transmission starts anew and the decoder ignores the previous block. Any state sequence s^∞ determines a set of possible stopping times $t_1 < t_2 < t_3 < \dots$. These t_i are defined as the minimal times such that $\log M < (1-\alpha) \sum_{i=t_{k-1}+1}^{t_k} I(x; y|s = s_i)$ for all $k = 1, 2, \dots$, where $t_0 = 0$.

The question of when transmission ends depends both on the state sequence (which specifies the t_i) and the channel noise (which determines whether retransmissions occur after each t_i). We bound the expected decoding time $E[t|s^\infty]$ as

$$E[t|s^\infty] = t_1 + \Pr[\text{retrans}|s_1^{t_1}] \left((t_2 - t_1) + \Pr[\text{retrans}|s_{t_1+1}^{t_2}] \left((t_3 - t_2) + \dots \right) \right) \quad (26)$$

$$\leq t_1 + \epsilon \left((t_2 - t_1) + \epsilon \left((t_3 - t_2) + \dots \right) \right) \quad (27)$$

$$\leq t_1 + \epsilon \left(N + \epsilon \left(N + \dots \right) \right) \quad (28)$$

$$= t_1 + \frac{\epsilon}{1-\epsilon} N. \quad (29)$$

The expected end-of-transmission time in (26) is parameterized by the state sequence and is defined both in terms of the state-dependent decoding times $t_1 < t_2 < \dots$, and the probability of retransmission following

each possible decoding time. To get (27) we will show that the probability of retransmission can be upper bounded by any $\epsilon > 0$ regardless of the state sequence $s_{t_{k-1}+1}^{t_k}$. In (28) we use the fact that since the channel is assumed to have a positive worst-case capacity, the duration of each transmission $t_k - t_{k-1} < N$ were the maximum block length N is bounded as in (21). The definition of t_1 tells us that $\log M \geq (1 - \alpha) \sum_{i=1}^{t_1-1} I(x; y|s = s_i) = (1 - \alpha)(t_1 - 1)R(W^{t_1-1}, p_x)$. Using this together with (28) in (29) gives

$$\begin{aligned} E[t|s^\infty] &\leq \frac{\log M}{(1 - \alpha)R(W^{t_1-1}, p_x)} + 1 + \frac{\epsilon}{1 - \epsilon} \frac{\log M}{(1 - \alpha) \min_s I(x; y|s = s)} \\ &\leq \frac{\log M}{(1 - \alpha)R(W^{t_1-1}, p_x)} \left[1 + \frac{(1 - \alpha) \max_s I(x; y|s = s)}{\log M} + \frac{\epsilon}{1 - \epsilon} \frac{\max_s I(x; y|s = s)}{\min_s I(x; y|s = s)} \right], \end{aligned} \quad (30)$$

since $R(W^{t_1-1}, p_x) \leq \max_s I(x; y|s = s)$. We will show that we can pick $\epsilon > 0$ as small as desired, so with $\log M$ picked suitably large, the average rate

$$\frac{\log M}{E[t|s^\infty]} \geq \frac{(1 - \alpha)R(W^{t_1-1}, p_x)}{1 + \epsilon'}$$

where $(1 + \epsilon')$ is the term within the square brackets in (30), and so ϵ' can be made as small as desired.

Our analysis combines a state-dependent erasure decoding analysis for the first decoding time with a worst-case analysis for the follow-on transmissions. A natural question to ask is why we ignore the accumulated information of earlier blocks at later decoding times. The answer is that as we will choose system parameters to make retransmissions exceptionally rare, ignoring that information does not have a significant impact on the bound on $E[t|s^\infty]$, and the difficulty of the analysis is eased considerably.

5.2 Bounding the reliability function

We now specify our decoding rule.

Definition 3 For any $0 < \alpha < 1$ and $\theta > 0$ the decoding regions $\{\mathcal{R}_n(m)\}$ are defined as follows.

$$\begin{aligned} \text{If } \log M > (1 - \alpha) \sum_{i=1}^n I(x; y|s = s_i) \text{ then } \mathcal{R}_n(m) &= \emptyset \text{ for } m \in \mathcal{M} \text{ and } \mathcal{R}_n(0) = \mathcal{Y}^n, \\ \text{else } \mathcal{R}_n(m) &= \left\{ \mathbf{y} \mid \frac{W^n(\mathbf{y}|\mathbf{x}_m, \mathbf{s})}{\sum_{\tilde{m} \neq m} W^n(\mathbf{y}|\mathbf{x}_{\tilde{m}}, \mathbf{s})} \geq \exp\{n\theta\} \right\} \text{ for } m \in \mathcal{M}. \end{aligned} \quad (31)$$

Recall that $\mathcal{R}_n(0) = \{\mathbf{y} \text{ s.t. } \mathbf{y} \notin \cup_{m \in \mathcal{M}} \mathcal{R}_n(m)\}$.

This is Forney's decoding rule [12] with the addition of state-dependence. Forney shows that the decoding regions implied by this rule give the optimal trade off (in a Neyman-Pearson sense) between the probabilities of error and erasure. When $\theta > 0$ at most one codeword satisfies the rule. Choosing θ negative leads to list decoding.

Say decoding occurs at time n . We define E_1 to be the event that the observation \mathbf{y} is not in $\mathcal{R}_n(m)$, where \mathbf{x}_m is the first n symbols of the transmitted codeword. The probability of this event upper bounds

the probability of erasure (i.e., the probability that a retransmission is requested).

$$\Pr[E_1|\mathbf{s}] = \sum_m \sum_{\mathbf{y} \notin \mathcal{R}_n(m)} W^n(\mathbf{y}|\mathbf{x}_m, \mathbf{s}) \Pr[\mathbf{x}_m]. \quad (32)$$

An error occurs if $\mathbf{y} \in \mathcal{R}_n(m)$ and codeword $\mathbf{x}_{\tilde{m}}$ was transmitted where $\tilde{m} \neq m$. We define this event as E_2 , the probability of which is parameterized by the state sequence up to that time:

$$\Pr[E_2|\mathbf{s}] = \sum_m \sum_{\mathbf{y} \in \mathcal{R}_n(m)} \sum_{\tilde{m} \neq m} W^n(\mathbf{y}|\mathbf{x}_{\tilde{m}}, \mathbf{s}) \Pr[\mathbf{x}_{\tilde{m}}]. \quad (33)$$

We now bound $\Pr[E_1|\mathbf{s}]$ and $\Pr[E_2|\mathbf{s}]$ for a random coding ensemble. We can following Forney's derivation until the point where we use the memoryless property of the channel. In particular, picking the codewords in an independent and identically distributed manner, where each symbol of each codeword is selected i.i.d. from $p_x(x)$, we get the following bound (see [12] pg. 219) on the average moment generating function $\overline{h(-\gamma)}$ (averaged over the ensemble of codebooks):

$$\overline{h_m(-\gamma|\mathbf{s})} \leq \exp\{\rho n R\} \left\{ \sum_{\mathbf{y}} \left[\sum_{\mathbf{x}} p_x^n(\mathbf{x}) W^n(\mathbf{y}|\mathbf{x}, \mathbf{s})^{1-\gamma} \right] \left[\sum_{\mathbf{x}} p_x^n(\mathbf{x}) W^n(\mathbf{y}|\mathbf{x}, \mathbf{s})^{\gamma/\rho} \right]^\rho \right\}, \quad 0 \leq \gamma \leq \rho \leq 1, \quad (34)$$

where the rate $R = \log M/n$. At the first decoding opportunity, $n = t_1$, at the second, $n = t_2 - t_1$, at the k th, $n = t_k - t_{k-1}$. The only modification of Forney's results thus far is the conditioning on the state sequence. The moment generating function is related to the probabilities of erasure and error as, respectively,

$$\overline{\Pr[E_1|\mathbf{s}]} = \exp\{n\theta\gamma\} (1/M) \sum_m \overline{h_m(-\gamma|\mathbf{s})}, \quad (35)$$

$$\overline{\Pr[E_2|\mathbf{s}]} = \exp\{n\theta(1-\gamma)\} (1/M) \sum_m \overline{h_m(-\gamma|\mathbf{s})}. \quad (36)$$

The derivation continues by extending Forney's analysis to a time-varying channel model where the number of uses of each of the channels is determined by the empirical distribution of the state sequence. We use the memoryless property of the channel and the i.i.d. nature of the input distribution to express (34) as

$$\begin{aligned} \overline{h_m(-\gamma|\mathbf{s})} &\leq \exp\{\rho n R\} \prod_{i=1}^n \left\{ \sum_{y_i} \left[\sum_{x_i} p_x(x_i) W(y_i|x_i, s_i)^{1-\gamma} \right] \left[\sum_{x_i} p_x(x_i) W(y_i|x_i, s_i)^{\gamma/\rho} \right]^\rho \right\}, \\ &= \exp\{\rho n R\} \prod_{s \in \mathcal{S}} \prod_{i \text{ s.t. } s_i=s} \left\{ \sum_{y_i} \left[\sum_{x_i} p_x(x_i) W(y_i|x_i, s_i)^{1-\gamma} \right] \left[\sum_{x_i} p_x(x_i) W(y_i|x_i, s_i)^{\gamma/\rho} \right]^\rho \right\}, \\ &= \exp\{\rho n R\} \prod_{s \in \mathcal{S}} \left\{ \sum_y \left[\sum_x p_x(x) W(y|x, s)^{1-\gamma} \right] \left[\sum_x p_x(x) W(y|x, s)^{\gamma/\rho} \right]^\rho \right\}^{n\Lambda_n(s)}, \\ &= \exp\{\rho n R\} \exp \left\{ -n \sum_s \Lambda_n(s) E_s(\gamma, \rho, p_x) \right\}. \end{aligned} \quad (37)$$

As before $\Lambda_n(s)$ is the fraction of the n symbols observed across channel $W(y|x, s)$, and

$$E_s(\gamma, \rho, p_x) = -\log \sum_y \left[\sum_x p_x(x) W(y|x, s)^{1-\gamma} \right] \left[\sum_x p_x(x) W(y|x, s)^{\gamma/\rho} \right]^\rho.$$

Substituting (37) into (35) and (36) gives

$$\overline{\Pr[E_1|\mathbf{s}]} \leq \exp \left\{ -n \left(\sum_s \Lambda_n(s) E_s(\gamma, \rho, p_x) - \rho R - \gamma \theta \right) \right\}, \quad (38)$$

$$\overline{\Pr[E_2|\mathbf{s}]} \leq \exp \left\{ -n \left(\sum_s \Lambda_n(s) E_s(\gamma, \rho, p_x) - \rho R + (1 - \gamma) \theta \right) \right\} = \overline{\Pr[E_1|\mathbf{s}]} \exp\{-n\theta\}. \quad (39)$$

The right-hand-side of (39) indicates that as θ is selected larger, the probability of error $\overline{\Pr[E_2|\mathbf{s}]}$ becomes exponentially smaller than the probability of erasure $\overline{\Pr[E_1|\mathbf{s}]}$. The tightest bounds are derived by maximizing the exponents over $0 \leq \gamma \leq \rho \leq 1$, giving the erasure and error exponents:

$$E_{\text{era}}(R, \theta, \Lambda, p_x) = \max_{0 \leq \gamma \leq \rho \leq 1} \sum_s \Lambda(s) E_s(\gamma, \rho, p_x) - \rho R - \gamma \theta$$

$$E_{\text{err}}(R, \theta, \Lambda, p_x) = \max_{0 \leq \gamma \leq \rho \leq 1} \sum_s \Lambda(s) E_s(\gamma, \rho, p_x) - \rho R + (1 - \gamma) \theta.$$

We solve for θ and express the error exponent in terms of the erasure exponent:

$$E_{\text{err}}(R, \theta, \Lambda, p_x) = \max_{0 \leq \gamma \leq \rho \leq 1} \frac{1}{\gamma} \sum_s \Lambda(s) E_s(\gamma, \rho, p_x) - \frac{\rho}{\gamma} R - \frac{1 - \gamma}{\gamma} E_{\text{era}}(R, \theta, \Lambda, p_x).$$

As the erasure exponent $E_{\text{era}}(R, \theta, \Lambda, p_x)$ approaches zero, the maximum error exponent $E_{\text{err}}(R, \theta, \Lambda, p_x)$ is achieved for a fixed R by letting γ and ρ both go to zero at the constant ratio $\beta = \rho/\gamma \geq 1$. Since both γ and $E_s(\gamma, \rho, p_x)$ are going to zero, L'Hôpital's rule gives the resulting function $E_{s,f}(\beta, p_x)$ for each channel individually (see [12], pg. 213):

$$E_{s,f}(\beta, p_x) = \lim_{\beta \rightarrow 0} \frac{1}{\gamma} E_s(\gamma, \rho, p_x)$$

$$= \sum_y \sum_x p_x(x) W(y|x, s) \left[\log W(y|x, s) - \log \left(\sum_{x'} p(x') W(y|x', s)^{1/\beta} \right)^\beta \right]. \quad (40)$$

The same derivation holds in the current context except that now the maximization over β must take into account the empirical state distribution as

$$E_f(R, \Lambda, p_x) = \max_{\beta \geq 1} \sum_s \Lambda(s) E_{s,f}(\beta, p_x) - \beta R, \quad (41)$$

By the decoding rule the decoding rate at the first decoding opportunity t_1 is bounded as $R = \log M/n > (1 - \alpha)R(W^{t_1-1}, p_x) = (1 - \alpha)I(\Lambda_{t_1-1})$. For any state distribution $\Lambda(s)$ and any $\delta > 0$ we choose $\theta = \theta_\Lambda$ as a function of $\Lambda(s)$ to get

$$E_{\text{err}}((1 - \alpha)I(\Lambda), \theta_\Lambda, \Lambda, p_x) = E_f((1 - \alpha)I(\Lambda), \Lambda, p_x) - \delta,$$

which gives

$$E_{\text{era}}((1 - \alpha)I(\Lambda), \theta_\Lambda, \Lambda, p_x) > 0.$$

To ensure that retransmissions are sufficiently rare we choose the codebook to be large enough. To determine this choice we maximize the probability of E_1 over all state distributions and choose $M > M^*$ where M^* is large enough that for any $\epsilon > 0$,

$$\overline{\Pr[E_1]} \leq \max_{\Lambda} \exp \left\{ -\frac{\log M^*}{(1-\alpha)I(\Lambda)} E_{\text{era}}((1-\alpha)I(\Lambda), \theta_{\Lambda}, \Lambda, p_x) \right\} < \epsilon. \quad (42)$$

We use (42) to bound the probability of retransmission in (27), giving a bound on the expected decoding time. This leads to the bound on the average rate of communication.

The error probability is upper bounded by

$$\overline{\Pr[E_2]} \leq \max_{\Lambda} \exp \left\{ -\frac{\log M^*}{(1-\alpha)I(\Lambda)} [E_f((1-\alpha)I(\Lambda), \Lambda, p_x) - \delta] \right\} \quad (43)$$

for any state sequence, thereby proving Theorem 2. The smallest error exponent is bounded by

$$\min_{\Lambda} E_f((1-\alpha)I(\Lambda), \Lambda, p_x).$$

5.3 Discussion

The improvement in the reliability function (error exponent) of erasure decoding as compared ML decoding can be seen by examining (41). In particular, consider the slope of the reliability function as we let $\alpha \rightarrow 0$. This limit corresponds to R approaching $\sum_s \Lambda(s) I(x; y|s = s)$. In this limit the exponent $E_f(R, \Lambda, p_x)$ approaches zero at slope at least -1 (since $\beta \geq 1$). This is a huge improvement over the behavior of ML decoding. For the ML decoding rule presented in Section 4 the slope of the reliability function as $\alpha \rightarrow 0$ would generally be zero.

The larger error exponent impacts the choice of M . If have a target probability of error that we want to achieve, and the choice of M^* specified by (42) isn't large enough to get (43) to meet that target, we can increase M further until it does. Since the error exponent of the erasure-decoding scheme is much larger than that of ML decoding, a much smaller codebook size M will be required to meet the target.

We emphasize that in this paper feedback is used in a very limited ways. In both this section and in Section 4 we use feedback to implement variable-length coding. The initial objective is to adjust the rate of communication to adapt to the non-ergodic nature of the channel. The result is a state-dependent rate function and an arbitrarily small probability of decoding error. However, the ML decoding rule of Section 4 does not take into account the stochastic behavior of the channel. Its error probability is determined by the likelihood that the channel behaves atypically. Therefore, as a secondary objective, in this section we further use the feedback to improve upon the reliability function of the ML scheme. By using erasure decoding in the place of ML, the code can often detect atypical channel behavior and avoid decoding at those times. Instead it requests a retransmission. The result is a coding strategy with a much higher reliability function.

6 Conclusions

In this paper we show that rateless codes find a natural application in achieving reliable and efficient communication over a non-ergodic channel. We examined the special case of systems with receiver channel state information. We specified two decoding rules: one based on ML decoding and one that gave an improved reliability function by using erasure decoding in the place of ML.

An extension of this work that we are now pursuing is to channels without decoder CSI. As discussed in Section 2, to make this extension we combine decoders appropriate for AVCs without CSI with variable-length decoding rules.

In this paper we use variable-length transmission to implement variable-rate communication. When the channel state information is arbitrary and known non-causally by the transmitter, but not by the receiver, variable-rate transmission can be implemented with a fixed block length. This problem is closely related to the arbitrarily varying version of the Gel'fand-Pinsker problem [14] as studied in [2] by Ahlswede. As long as the worst-case capacity is positive, the encoder can signal to the decoder its observed state type. It can then follow up with a codebook whose rate depends on the type of the state sequence. As in [2], the decoding rule would also be tuned to the type of the state sequence.

References

- [1] R. Ahlswede. Channels with arbitrarily varying channel probability functions in the presence of noiseless feedback. *Z. Wahrscheinlichkeitsth. Verw. Gebiete*, 25:239–252, 1973.
- [2] R. Ahlswede. Arbitrarily varying channels with states sequence known to the sender. *IEEE Trans. Inform. Theory*, 32:621–629, September 1986.
- [3] R. Ahlswede and I. Csiszár. Common randomness in information theory and cryptography – Part II: CR capacity. *IEEE Trans. Inform. Theory*, 44:225–240, January 1998.
- [4] J. Byers, M. Luby, M. Mitzenmacher, and A. Rege. A digital fountain approach to reliable distribution of bulk data. In *Proc. ACM SIGCOMM*.
- [5] G. Caire and D. Tuninetti. The throughput of Hybrid-ARQ protocols for the Gaussian collision channel. *IEEE Trans. Inform. Theory*, 47:1971–1988, July 2001.
- [6] I. Csiszár and J. Körner. *Information Theory, Coding Theorems for Discrete Memoryless Systems*. Akadémiai Kiadó, 1981.
- [7] S. C. Draper. Universal incremental Slepian-Wolf coding. In *Proc. 42nd Allerton Conf. on Communication, Control and Computing*, October 2004.

- [8] S. C. Draper, B. J. Frey, and F. R. Kschischang. Efficient variable length channel coding for unknown DMCs. In *Proc. Int. Symp. Inform. Theory*, page 379, June 2004.
- [9] S. C. Draper and A. Sahai. Noisy feedback improves communication reliability. In *Submitted to Proc. Int. Symp. Inform. Theory*.
- [10] U. Erez, G. W. Wornell, and M. D. Trott. Faster-than-Nyquist coding: The merits of a regime change. In *Proc. 42nd Allerton Conf. on Communication, Control and Computing*, Allerton House, Monticello, IL, September 2004.
- [11] O. Etesami, M. Molkarai, and A. Shokrollahi. Raptor codes on symmetric channels. In *Proc. Int. Symp. Inform. Theory*, page 38, Chicago, IL, June 2004.
- [12] G. D. Forney. Exponential error bounds for erasure, list, and decision feedback schemes. *IEEE Trans. Inform. Theory*, 14:206–220, March 1968.
- [13] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley and Sons, 1968.
- [14] S. I. Gel'fand and M. S. Pinsker. Coding for channels with random parameters. *Problems of Control and Information Theory*, 9:19–31, 1980.
- [15] M. Horstein. Sequential transmission using noiseless feedback. *IEEE Trans. Inform. Theory*, 1963.
- [16] A. Lapidoth and P. Narayan. Reliable communication under channel uncertainty. *IEEE Trans. Inform. Theory*, 44:2148–2177, October 1998.
- [17] M. Luby. LT codes. In *the 43rd Annual IEEE Symposium on Foundations of Computer Science*, 2002.
- [18] R. Mantha and F. R. Kschischang. A capacity-approaching hybrid ARQ scheme using turbo codes. In *Proc. IEEE GLOBECOM*, pages 2341–2345, 1999.
- [19] R. Palanki and J. S. Yedidia. Rateless codes on noisy channels. In *Proc. Int. Symp. Inform. Theory*, page 37, Chicago, IL, June 2004.
- [20] D. N. Rowitch and L. B. Milstein. On the performance of hybrid FEC/ARQ systems using rate-compatible punctured turbo RCPT codes. *IEEE Trans. Commun.*, 48:948–959, June 2004.
- [21] S. Sesia, G. Caire, and G. Vivier. Incremental redundancy Hybrid ARQ schemes based on low-density parity-check codes. *submitted to IEEE Trans. Commun.*
- [22] O. Shayevitz and M. Feder. Communicating using feedback over a binary channel with arbitrary noise sequence. In *Proc. Int. Symp. Inform. Theory*, 2005.

- [23] A. Shokrollahi. Fountain codes. In *Proc. 41st Allerton Conf. on Communication, Control and Computing*, pages 1290–1297, October 2003.
- [24] N. Shulman. *Communication over an Unknown Channel in Common Broadcasting*. PhD thesis, Tel Aviv Univ. 2003.
- [25] E. Soljanin, N. Varnica, and P. Whiting. Incremental redundancy Hybrid ARQ with LDPC and Raptor codes. *submitted to IEEE Trans. Inform. Theory*, 2005.
- [26] S. Z. Stambler. Shannon’s theorems for a complete class of discrete channels whose state is known at the output. *Prob. Peredachi Informatsii*, 11(4):3–12, 1975.
- [27] A. Tchamkerten and E. Telatar. Optimal feedback schemes over unknown channels. In *Proc. Int. Symp. Inform. Theory*, page 378, Chicago, IL, June 2004.