Anytime Channel Coding with Feedback

by

Hidayet Tunc Simsek

B.S. (Middle East Technical University, Ankara, Turkey) 1996
M.S. (University of California, Berkeley) 2000

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

in

Engineering – Electrical Engineering
and Computer Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Pravin P. Varaiya, Chair
Professor Anant Sahai
Professor John K. Hedrick

Fall 2004

The dissertation of Hidayet Tunc Simsek is approved:

Chair Date

Date

Date

University of California, Berkeley

Fall 2004

Anytime Channel Coding with Feedback

Abstract

Anytime Channel Coding with Feedback

by

Hidayet Tunc Simsek
Doctor of Philosophy in Engineering – Electrical Engineering
and Computer Sciences

University of California, Berkeley

Professor Pravin P. Varaiya, Chair

We build on the Anytime information theory of Anant Sahai. Within that broad range we focus on the Anytime channel coding problem for discrete-time memoryless channels with noiseless feedback. Our main result is an upper bound on the exponent of anytime coding. For the special case of a Binary Erasure Channel with feedback the result is tight or, in other words, the upper bound is achievable. The upper bound turns out to have some interesting properties. For any discrete-time memoryless channel the bound is a concave function, it is strictly positive for all rates up to capacity and identically zero at capacity. Furthermore, for symmetric channels, the bound is expressed parametrically in terms of Gallager's exponent which is generally denoted as E_0. The discussion occupies Chapters 1-3.

In Chapter 4, we specialize to the Binary Symmetric Channel and design a family of 'time-sharing' codes that achieve an exponent close to the upper bound. In particular, the achievable region beats the sphere-packing bound which was thought to be a hard bound on performance with fixed delays [21]. This shows that the sphere-packing bound is misleading when it comes to the opportunities provided by feedback.

The two additional chapters may be read independently of the former material. In Chapter 5 we consider the problem of anytime relaying. We are interested in this problem because it represents the simplest instance of a more general problem of real-time data processing. The relaying node must be able to decode, interpret and re-encode the incoming data in a streaming fashion. We are unable to treat the relaying problem in general and therefore restrict our attention to the special case of a one-hop relay in which the link from the source node to the relaying node is an Erasure Channel with feedback. In this case we are able to give a precise characterization of the anytime capacity of the end-to-end relay.

In the final chapter we consider an infinite state Markovian jump system. The material presented in this chapter unfolded during one of our many attempts to model our anytime codes. We present stability criteria for jump systems. Our results may be viewed as a simpler and finite alternative to the traditional criteria which requires the solution of an infinite set of balance equations.

Professor Pravin P. Varaiya

Dissertation Committee Chair

To my mother, father and brother.

Contents

List of Figures

Acknowledgments

Its hard to find the right words to express my gratitude to my advisor Professor Pravin Varaiya. He has opened the doors for a brave new world, one in which I'm motivated by the self-contained beauty of scientific innovation and by rational thought and consideration for the basic problems. I've tried to follow along his footsteps and I can only hope that I will one day reach a tiny fraction of his wisdom and thoughtfulness. Most of all, I thank him for his friendship, through which I've become a better person.

I thank Professor Anant Sahai for his trust and sincere interest in my education and for showing me the 'scientific way' of doing things. This dissertation would not have been possible without his due diligence, thoroughness and razor sharp sense for the subject material. I hope that the results in this dissertation are found to be useful to the Anytime Information Theory which he published four years ago.

I thank Professor Karl Hedrick for being an invaluable mentor over the entire span of my graduate life. He taught me the right way to do cross-disciplinary research in the DARPA funded Mobies program.

I thank Professor Kemal Inan, without whom none of this would have been possible. Before I met him, science to me was something other people did. I am forever in his gratitude.

I would also like to thank my fellow students Rahul Jain and Duke Lee for being there through the good and the bad and for their sincere friendship.

Last but not least, I thank my loving mother, father, brother and my sweetheart Marcela. I cannot imagine myself going through with this degree without their unconditional love and support.

Chapter 1

Introduction

The reliable communication of real-time information over unreliable communication channels is central in modern technological applications. Examples include interactive streaming multimedia found in the on-line entertainment industry and interactive trading systems such as distributed auctions. There is no shortage of demand for real-time communications and, fortunately, on the technological side, there is no shortage of supply. We witness rapid development of integrated communication systems that match and often surpass the requirements of the driving applications. In fact, we often see technology driving the applications.

At the theoretical foundations of reliable communication of streaming data lies the simply stated problem of estimating an unstable system from measurements taken over a noisy data-rate limited communications link. In contrast to advances in technology, this seemingly simple problem, originally motivated by Schalkwijk and Kailath [26, 24] in the 1960s, has since evaded a sufficiently general understanding and a general method of design. Dynamic programming emerged as the favorite approach but, ultimately, proved not to be a broadly applicable methodology. In the absence of additional assumptions on the channel, the dynamic programming approach did not lead to tractable and useful formulations.

A recent and more fundamental approach is contained in the Anytime information transmission theorem of A. Sahai [22] in 2001. Contrary to previous belief, Sahai's theorem demonstrates the existence of encoders and decoders that are capable of tracking a random walk remotely over an arbitrary discrete-time memoryless channel. No additional assumptions are needed on the channel. To see what's going on consider the situation illustrated in Figures 1.1 and Figure 1.2. The source is a random walk described by the recursion $x_{n+1} = x_n + s_n$ in which $x_0 = 0$ and the $\{s_n\}$ are i.i.d. coin tosses with $\mathbb{P}(s_n = 1) = \mathbb{P}(s_n = -1) = 1/2$. In contrast to the standard assumptions on the source, the random walk is not stationary and the variance of x_n tends to infinity. The communication channel is modeled by a conditional probability distribution $Q_n(b|a)$ defined over the channel inputs $a = 1, 2, \cdots, A$ and the channel outputs $b = 1, 2, \cdots, B$. The tracking problem is to design an encoder that causally observes x_n and a decoder that, at time n, produces an estimate y_n of x_n such that,

$$\sup_{n\geq 0} \mathbb{E}|x_n - y_n|^2 < \infty. \tag{1.1}$$

The diagram in Figure 1.1 depicts the traditional framework, originally proposed by Witsenhausen [35]. The idea is to formulate the problem as a controller design problem. The decoder is modeled as a dynamic system $y_n = \mathcal{D}_n(y^{n-1}, b_n)$ and the encoder is modeled as a controller which steers the decoder through the noisy control signal b_n. As pointed out by Walrand and Varaiya [33] the design problem may be casted as a dynamic program with partial information. The information state crucially depends on the channel and admits no tractable representation in the absence of additional simplifying assumptions.

The diagram in Figure 1.2 depicts the Anytime framework proposed by Sahai. In the spirit of Shannon the idea is to break up the system into a source encoder-decoder and channel encoder-decoder and to think of these as two separate design problems. For a simple random walk the source encoder-decoder is trivial. In particular, the source encoder is described by the functions $\mathcal{E}_n^S(x^n) = s_n,\ n \geq 0$, and the source decoder is described by the functions $\mathcal{D}_n^S(\tilde{s}(n)) = \sum_{i=1}^n \tilde{s}_n,\ n \geq 0$. The heart of the matter lies in the channel encoder $\mathcal{E}_n^a(s^n, b^{n-1})$ and the channel decoder $\tilde{s}(n) = \mathcal{D}_n^a(b^n)$ in which $\tilde{s}(n)$ is a vector-valued estimate of the entire history of the bits that have entered the encoder up to time n,

$$\tilde{s}(n) = \big(\tilde{s}_1(n), \tilde{s}_2(n), \cdots, \tilde{s}_n(n)\big),\ n \geq 0.$$

Observe that the estimate of a particular bit may change at 'any' future time. The anytime channel coding theorem states that for a discrete-time memoryless channel with Shannon capacity greater than one there exists a channel encoder-decoder $(\mathcal{E}^a, \mathcal{D}^a)$ with parameters $\alpha > 0$ and $\beta > 0$ such that for all times $n \geq 0$ and for all delays $0 \leq d \leq n$ we have,

$$\mathbb{P}\big(s_{n-d} \neq \tilde{s}_{n-d}(n)\big) \leq \beta 2^{-\alpha d}. \tag{1.2}$$

So the probability of decoding error for a past bit drops off exponentially in the amount of time that the system had to decode that bit. The tracking property (1.1) then follows as,

$$\begin{aligned}
\mathbb{E}|x_n - y_n|^2 &= \sum_{s^n} \big|x_n - \mathcal{D}_n^S(\tilde{s}(n))\big|^2 \, \mathbb{P}(\tilde{s}(n)) \\
&= \sum_{s^n} \big|x_n - \sum_{i=1}^n \tilde{s}_n\big|^2 \, \mathbb{P}(\tilde{s}(n)) \\
&= \sum_{s^n} \big|\sum_{i=1}^n s_n - \sum_{i=1}^n \tilde{s}_n\big|^2 \, \mathbb{P}(\tilde{s}(n)) \\
&\leq \sum_{0 \leq d \leq n} \beta 2^{-\alpha d} \\
&\leq \frac{\beta}{1 - 2^{-\alpha}} \\
&< \infty.
\end{aligned}$$

We note the anytime channel coding theorem also justifies the use of the term 'anytime'. More precisely, the estimate of any given bit is permitted to change at any time and the estimate monotonically improves in time.

In contrast with the traditional approach, the anytime approach imposes no additional requirements on the channel except that its Shannon capacity is bigger than one. It should be pretty clear that if the capacity is less than one then there is no encoder-decoder that can track the random walk. The additional flexibility of the anytime approach is because the decoder is allowed to re-examine its past estimates based on its current knowledge. The traditional decoder is strictly progressive.

This brings us to the main subject of this dissertation. We attempt to contribute to the basic problem by determining bounds on the anytime exponent α and to design anytime codes that achieve or come close to these bounds. We will not worry about the magnitude of the coefficient β in our calculations. If $\alpha_1 > \alpha_2$, it is quite meaningful to say that one code with parameters (α_1, β_1) is better than another code with parameters (α_2, β_2): For any particular bit the probability of bit decoding error in (1.2) is governed by the exponent rather than the coefficient.

1.1 Summary of main results

We consider the anytime channel problem as described above. In addition to what is described above the rate at which new bits enter the encoder is permitted to be different than one and is denoted by R ($R > 0$). This roughly means that by time n a total of $j_n = \lceil nR \rceil$ source bits have entered the encoder and, correspondingly, the decoder's estimate at time n is given by,

$$\tilde{s}(n) = \big(\tilde{s}_1(n), \tilde{s}_2(n), \cdots, \tilde{s}_{j_n}(n)\big).$$

Consequently the anytime exponent also depends on the rate R and we write $\alpha(R)$ to denote the dependence. The basic problem is to determine bounds on the anytime exponent $\alpha(R)$ and design anytime codes that achieve or come close to these bounds over all rates R for which the anytime exponent is non-zero.

The presentation consists of three parts. In the first part of our presentation we survey the literature and present relevant results for the reliability function of block coding with and without channel feedback. The purpose in doing so is to understand and, if possible, leverage the relationship between the anytime exponent and the traditional reliability function of block coding.

The case of block coding with no feedback is well-understood and is treated in standard textbooks. By contrast, when feedback is permitted, we come across a number of surprising results that have not quite made it into the textbooks. We place emphasis on two such results (Theorems 2.1.1 and 2.2.1) which, at a very high level, state that feedback cannot be used to improve the reliability function for symmetric channels and that the reliability function can be significantly improved if the block length is allowed to depend on the feedback. We subsequently make use of the first result to upper bound (Theorem 3.3.1) the anytime exponent for symmetric channels and we use the second result in our code designs to improve the region of achievable exponents (Theorem 4.4.1).

Our upper bound turns out to have some desirable properties. In the second part of our presentation we focus on the upper bound and demonstrate that it is concave, positive for all rates R up to capacity C and identically zero at capacity (Theorem 3.3.2). For symmetric channels, we are able to express our bound parametrically in terms of Gallager's exponent which is usually denoted by E_0. So our upper bound is a 'nice' function. We subsequently argue that the bound is also a nice bound. In particular, it is tight for a Binary Erasure Channel with feedback. The upper bound, denoted $E_a^+(R)$, is plotted in Figure 1.3 for a Binary Erasure Channel with erasure probability $\varepsilon = 1/4$. The curve is described as the greatest lower bound for a family of curves parameterized by $0 \leq \lambda \leq 1$.

We conclude the main subject of our presentation by designing a time-sharing anytime code for discrete-time memoryless channels with perfect feedback (Theorems 4.3.1 and 4.4.1). The performance of the code is measured in terms of the region of achievable exponents. For the Binary Erasure Channel the time-sharing reduces to a trivial bit-by-bit repetition scheme that achieves the upper bound. For the Binary Symmetric Channel there is no such trivial reduction. We present a family of 'N-up J-down' schemes that are fairly simple and achieves a region close to the upper bound. In particular, the achievable region beats the sphere-packing bound which was thought to be a hard bound on performance with fixed delays [21]. This shows that the sphere-packing bound is misleading when it comes to the opportunities provided by feedback. The Anytime exponent for various N-up J-down codes for a Binary Symmetric Channel with crossover probability $\epsilon = 1/4$ are sketched in Figure 1.4. The reliability region

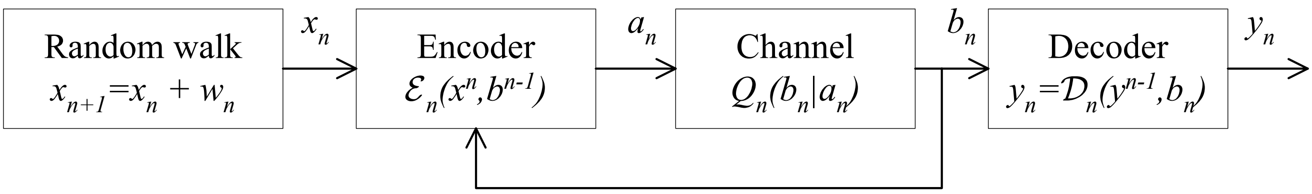

Figure 1.1: The dynamic programming framework for tracking a random walk.

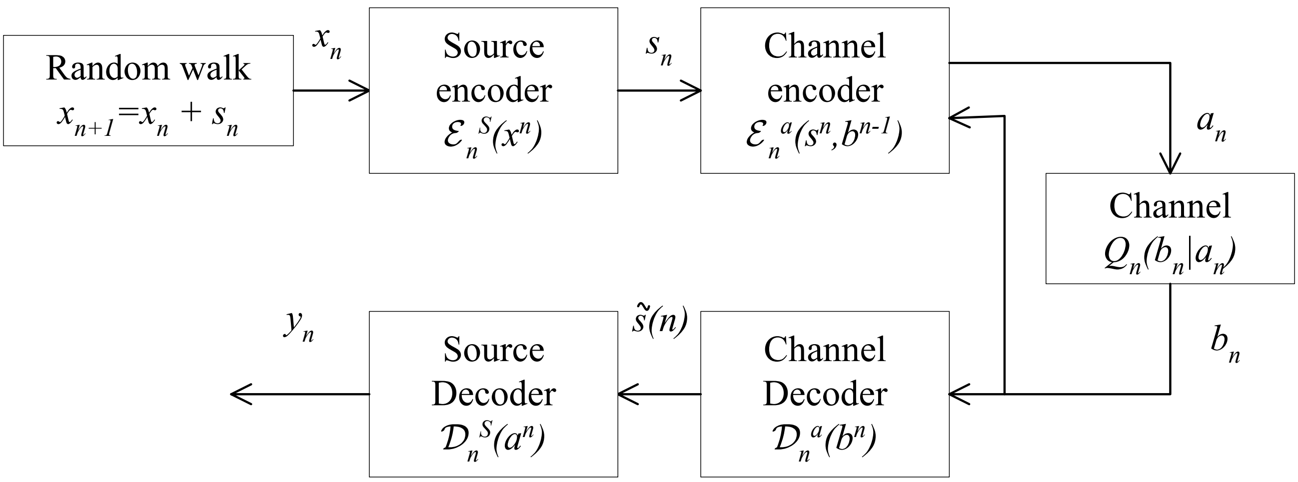

Figure 1.2: The anytime framework for tracking a random walk.

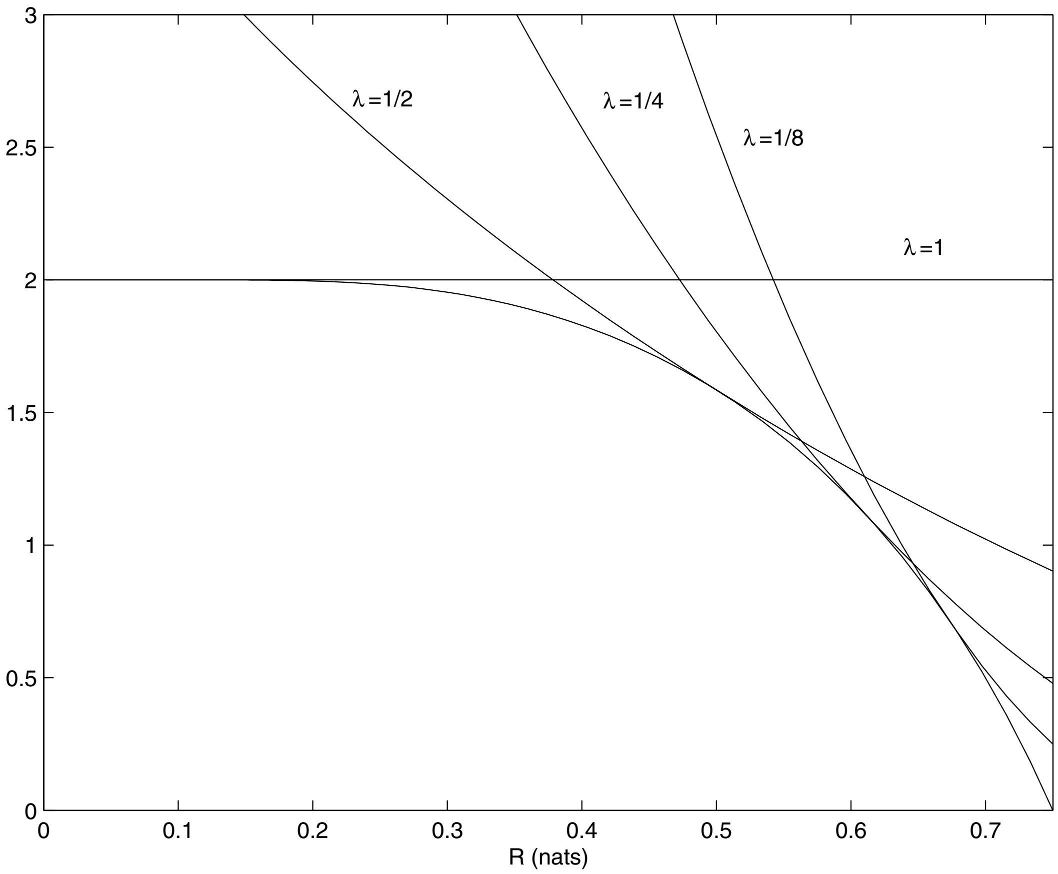

Figure 1.3: Plot of the Anytime upper bound $E_a^+(R)$ as a function of the transmission rate R for a Binary Erasure Channel with erasure probability $\varepsilon = 1/4$. The curve is described as the greatest lower bound for a family of curves parameterized by $0 \leq \lambda \leq 1$.

achievable by the family of N-up J-down curves is understood to be greater than the least upper bound of these curves. The curve $E_a^+(R)$ is the Anytime upper bound and $E_{sp}(R)$ is the sphere-packing bound.

Two subsequent chapters may be read independently of the former material. In Chapter 5 we consider the problem of anytime relaying. We are interested in this problem because it represents the simplest instance of a more general problem of real-time data processing. The relaying node must be able to decode, interpret and re-encode the incoming data in a streaming fashion. We are unable to treat the relaying problem in general and therefore restrict our attention to the special case of a one-hop relay in which the link from the source node to the relaying node is an Erasure Channel with feedback. In this case we are able to give a precise characterization of the anytime capacity of the end-to-end relay (Theorem 5.2.1).

In the final chapter we consider an infinite state Markovian jump system. The material presented in this chapter unfolded during one of our many attempts to model our anytime codes. We present sufficient conditions (Theorems 6.2.1 and 6.2.2) under which the jump system is stable. These conditions may be viewed as a simpler and finite alternative to the traditional formulation which requires the solution of an infinite set of balance equations.

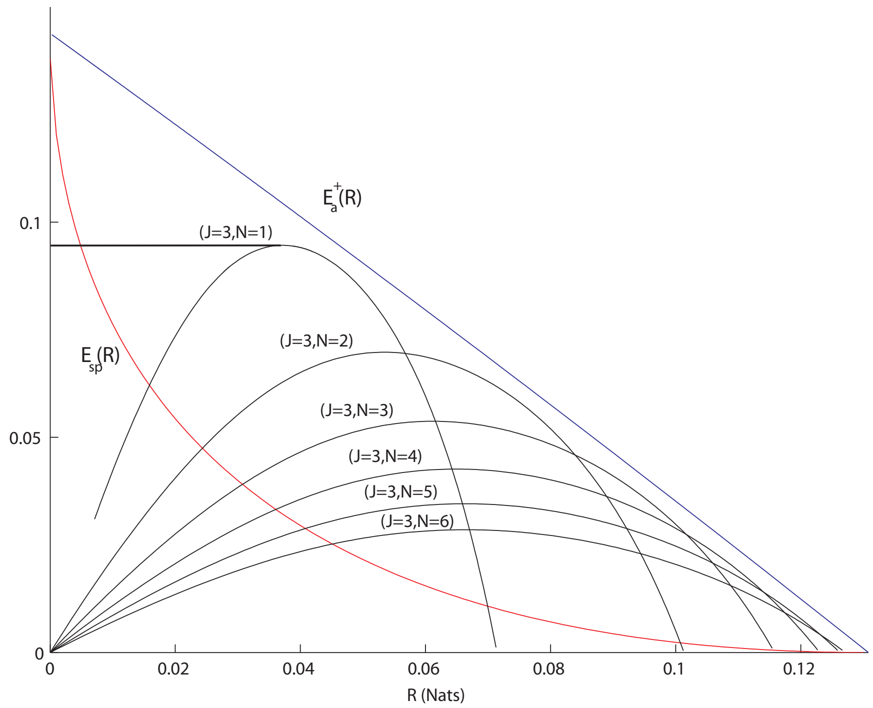

Figure 1.4: The Anytime exponent for various N-up J-down codes for a Binary Symmetric Channel with crossover probability $\epsilon = 1/4$. The curve $E_a^+(R)$ is the Anytime upper bound and $E_{sp}(R)$ is the sphere-packing bound.

1.2 List of notation

To assist with the presentation, we have compiled the following list of symbols which appear in the subsequent material.

1.2.1 Conventional symbols used in equations and derivations

x^k Shorthand for $x^k = x_1^k = (x_1, x_2, \cdots, x_k)$. Similarly for other symbols $y, z, \ldots$

$\delta_1, \delta_2, \delta_3$ Arbitrary small numbers used in approximations.

λ, ρ, η Arbitrary optimization parameters.

$\mathbb{P}(\mathcal{X})$ Probability of the event $\mathcal{X}$.

$\mathbb{E}$ Expectation of the random variable X.

1.2.2 Symbols used to describe channels

ϵ Binary Symmetric Channel (BSC) crossover probability.

ε Binary Erasure Channel (BEC) erasure probability.

p_{ij}, q_{ij} Channel transition probabilities for an arbitrary discrete memoryless channel.

C Shannon capacity.

C_1 Relative entropy. See Theorem 2.2.1.

$C^{\mathcal{A}}$ Anytime capacity.

1.2.3 Notation relating to fixed-length and variable-length block-coding

$R, \bar{R}$ Transmission rate and average transmission rate.

T, N Block length for fixed-length block coding.

L Block length for fixed-length transmission of a binary message.

$\bar{T}, \bar{N}$ Average block length for variable-length block coding.

a_k, b_k The random variables denoting the channel input and output at time k.

$\mathcal{E}_k, \mathcal{D}_k$ The functions that describe the encoder and decoder.

P_e Probability of decoding error.

$P_{e,m}$ Probability of decoding error given that the message $m \in \{1, 2, \cdots, M\}$ was transmitted.

1.2.4 Notations specific to Anytime coding

d Decoding delay.

$P_{e,s}(j, n)$ Probability of decoding error for the j-th bit at time n given that the bit sequence $s = s_1^n = (s_1, s_2, \cdots, s_n)$ was transmitted.

x_k, σ The state and system coefficient of an unstable system, Theorem 3.4.1 and (4.1).

y_k, $\hat{x}_k$ Decoder's estimate of x_k at time k.

ϕ_k, ψ_k Functions that describe the state encoder and estimator.

I_k Decoder's interval of confidence for estimate $\hat{x}_k$.

X_k The worst-case estimation error.

1.2.5 List of reliability functions

E Reliability function for fixed-length block coding.

E_r Random coding lower bound for the fixed-length block coding reliability function.

E^+ Information theoretic upper bound for the reliability function of an arbitrary fixed-length block code using feedback. See (2.2).

E_{sp} Sphere-packing upper bound for the reliability function of an arbitrary fixed-length block code using feedback over a symmetric channel. See Theorem 2.1.1.

E_{opt} Burnashev's reliability function for variable-length block codes with full information feedback. See Theorem 2.2.1.

E_f Forney's reliability function for variable-length block codes with binary decision feedback. See Theorem 2.4.1.

E_L Viterbi's reliability function for tree codes using feedback. See Theorem 3.5.1.

E_a Anytime exponent or the Anytime reliability function.

E_a^+ Upper bound on the Anytime exponent. See Theorem 3.3.1 and 3.3.2. $E_a^+ = E_L$ for symmetric channels.

E_0 Gallager's exponent used in the parametric representation of E_{sp} and E_a^+. See Theorem 2.1.1.

Chapter 2

The role of feedback in traditional communication

The basic and generally adopted use of feedback is to improve the performance of a feed-forward communication system. Schalkwijk and Kailath [26, 24] considered the Additive White Gaussian Noise (AWGN) channel with an 'average power' constraint and they demonstrated that feedback can be used in such a way that the probability of decoding error drops as a 'double' exponential in the block length. This result is interesting because without feedback, by the sphere-packing bound ([14], 5.8.2), the dependence is strictly exponential.

Therefore we are surprised to learn of Berlekamp's result [3] which says that the sphere-packing bound remains valid for the Binary Symmetric Channel even with access to noiseless and delayless feedback. More generally, the sphere-packing bound remains valid for any 'sufficiently symmetric' discrete memoryless channel.

The gap between Schalkwijk's result and that of Berlekamp's is explained by Wyner [36] who showed that the probability of decoding error for an AWGN with a 'peak power' constraint (instead of the average power constraint) cannot drop off as a double exponential in the block length: the dependence is strictly exponential. This leads us to think that the single exponential dependence and, in particular, the sphere-packing bound, is not an artifact of symmetric discrete channels but rather an artifact of the block-coding framework. Perhaps block-coding is not the 'best' way to transmit information when feedback is available. In the rest of the chapter, we consider discrete memoryless channels with perfect feedback and we bring into sharp focus the key results that describe the reliability function for such channels (Theorems 2.1.1 and 2.2.1). Among these we pay special attention to the technique of Yamamoto and Itoh [37, 38]. This is a simple variable-length block-coding strategy which achieves reliabilities above the sphere-packing bound.

2.1 The reliability function for block coding

We begin with the case of fixed-length block-coding. This means that the decoder is asked to make a decision exactly after N uses of the channel. More precisely, consider a set of M messages. A message m ($1 \leq m \leq M$) is fed into the encoder. The encoder emits N channel symbols a_n, $n = 1, 2, \cdots, N$ to convey the message m to the decoder. The decoder receives b_n, $n = 1, 2, \cdots, N$. There is a perfect feedback link from the decoder to the encoder. So the encoder is described by functions,

$$a_n = \mathcal{E}_n\big(m, b_1, b_2, \cdots, b_{n-1}\big),\ n = 1, 2, \cdots, N.$$

The channel is described by a ‘memoryless’ probability distribution,

$$\begin{aligned}\mathbb{P}\big(b_n = b \mid\ a_n = a, a_{n-1}, a_{n-2}, \cdots, a_1,&\\ b_{n-1}, b_{n-2}, \cdots, b_1\big)&\\ = \ \mathbb{P}\big(b_n = b \mid a_n = a\big)&\\ = \ p_{ab}, \quad a = 1, 2, \cdots, A,\ b = 1, 2, \cdots, B.&\end{aligned} \tag{2.1}$$

The decoder is described by the function,

$$m' = \mathcal{D}\big(b_1, b_2, \cdots, b_N\big).$$

Such a configuration defines a rate $R = (\log M)/N$ (nats) communication system with feedback. For a family $(N, M = 2^{RN})$ of such fixed-length block codes the quantity that we are interested in is the reliability function,

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

in which P_e is the probability of decoding error,

$$P_e = \max_{m\in\{1,\cdots,M=2^{RN}\}} \mathbb{P}\Big\{m' \neq m \mid m \text{ was transmitted}\Big\}.$$

The result that we are interested in says that if the channel is ‘sufficiently’ symmetric then the sphere-packing bound remains an upper bound for the reliability function with feedback [1],

$$E(R) \leq E_{sp}(R),\ 0 \leq R \leq C.$$

The channel is said to be sufficiently symmetric if the set of all output symbols can be divided into subsets in such a way that for each subset the submatrix of channel transition probabilities $[p_{ij}]$ possesses the property that each row is a permutation of another row and each column is a transposition of another column. The Binary Erasure and Binary Symmetric Channels are symmetric according to this definition.

If the channel is not symmetric then the reliability function with feedback can be larger than the sphere-packing bound [39]. In this case we know that the following quantity is an upper bound on the reliability function with feedback,

$$E^{+}(R) = \min_G \left\{ \max_i \sum_j g_{ij} \log \frac{g_{ij}}{p_{ij}} \ :\ \max_q I(G,q) \leq R \right\} \tag{2.2}$$

where $G = \{g_{ij}\}$ is a probability matrix and $I(G,q)$ measures the information as,

$$I(G,q) = \sum_i \sum_j q_i g_{ij} \log \frac{g_{ij}}{\sum_i q_i g_{ij}}$$

These results are summarized in the following Theorem.

Theorem 2.1.1 (Arutyunyan,[1]) *Consider a family $(N, 2^{RN})$ of rate R block codes with access to noiseless and delayless feedback. Then,*

$$E(R) \leq E^{+}(R),\ 0 \leq R \leq C,$$

where E^+ is defined in (2.2). If the channel is symmetric then E^+ reduces to the sphere-packing bound $E^+(R) = E_{sp}(R)$, $0 \le R \le C$. The sphere-packing bound may be calculated by the following expression given in [1],

$$E_{sp}(R) = \max_q \min_G \left\{ \sum_i q_i \sum_j g_{ij} \log \frac{g_{ij}}{p_{ij}} \; : \; I(G,q) \le R \right\}, \; 0 \le R \le C,$$

or the following parametric Gallager form [14],

$$E_{sp}(R) = \max_{\rho > 0} \big[E_0(R,\rho) - \rho R \big]$$

$$E_0(R,\rho) = -\max_q \sum_j \left[\sum_i q_i p_{ij}^{\frac{1}{1+\rho}} \right]^{(1+\rho)}, \; 0 \le R \le C.$$

Remark 2.1.1 *An earlier version of Theorem 2.1.1 was given by Dobrushin [9]. He considered a restricted class of symmetric discrete-time memoryless channels with feedback. For this class he showed that feedback cannot be used to improve the reliability exponent for rates R larger than a critical threshold.*

2.2 The reliability function for variable-length block coding

In variable-delay coding we allow the decoder to select when it wants to stop decoding. A message m ($m = 1, 2, \cdots, M$) enters an encoder. The encoder is described by the process,

$$a_n = \mathcal{E}_n\big(m, b_1, b_2, \cdots, b_{n-1}\big), \; n = 1, 2, \cdots, \tag{2.3}$$

in which n is an index that denotes the n-th channel use and $\{a_n, b_n\}$ are channel inputs and outputs described by a memoryless probability distribution,

$$\mathbb{P}\big(b_n = b \mid a_n = a\big) = p_{ab}, \; a = 1, 2, \cdots, A, \; b = 1, 2, \cdots, B,$$

At each time n the decoder receives b_n. Based on what it has received up to time n the decoder either makes an estimate m' of m at time n or refrains from making a decision at time n. So the decoder is described by a process,

$$\mathcal{D}_n(b_1, b_2, \cdots, b_n) \in \{1, 2, \cdots, M\} \bigcup \{\emptyset\}, \; n = 1, 2, \cdots,$$

where $\emptyset$ denotes that no decision has been made.

Both the encoder and decoder processes terminate at the first time instant when the decoder makes a decision. We will write T to denote this time instant,

$$T = \min\{n \ge 1 \; : \; \mathcal{D}_n(b_1, b_2, \cdots, b_n) \ne \emptyset\}.$$

Keep in mind that T is a random variable and both the encoder and decoder are random processes with random termination times. For the set-up to make sense we need to know that the process terminates. Typically, the following condition is required,

$$\mathbb{P}\big(T < \infty\big) = 1.$$

We would like to know more about the distribution of T, but such an analysis is quite difficult except for a few special cases. Write $\bar{T} = \mathbb{E}T$. Keeping in mind that $\bar{T}$ can be infinite we define the rate of a variable-length code as,

$$\bar{R} = \frac{\log M}{\bar{T}},$$

and we define the probability decoding error,

$$P_e = \max_{m=1,2,\cdots,M} \mathbb{P}\big(\mathcal{D}_n(b_1, \cdots, b_n) \neq m \mid T = n \text{ and message } m \text{ was transmitted}\big)$$

where we assumed equally likely messages. Finally, for a family $(\bar{R}, 2^{\bar{T}\bar{R}})$ of rate $\bar{R}$ encoder-decoders, we define the reliability function as,

$$E(\bar{R}) = \lim_{P_e \to 0} \frac{\log P_e^{-1}}{\bar{T}},$$

when the limit exists.

Theorem 2.2.1 (Burnashev, [5]) *Consider a discrete-time memoryless channel with feedback and transition probabilities* $\{p_{ij} \neq 0\}$*. For a family* $(\bar{T}, 2^{\bar{T}\bar{R}})$ *of rate* $\bar{R}$ *encoder-decoders, the reliability is upper bounded as,*

$$E(\bar{R}) \leq E_{opt}(\bar{R}) = C_1\left(1 - \frac{\bar{R}}{C}\right), \; 0 \leq \bar{R} \leq C,$$

where,

$$C_1 = \max_{i,k} \sum_l p_{il} \log \frac{p_{il}}{p_{kl}}.$$

Furthermore, there exists a family that achieves $E_{opt}(R)$ *for all* $0 \leq \bar{R} \leq C$*.*

There are a few notable properties of this result. We mention two of these. First, C_1 is extremely easy to calculate for any discrete-time memoryless channel. Second, the performance exceeds that of the sphere packing bound. A plot of $E_{opt}(R)$ and $E_{sp}(R)$ for a binary symmetric channel with cross-over probability $\epsilon = 0.1$ is illustrated in Figure 2.1. Finally, that E_{opt} is a straight line suggests a 'time-sharing' property of the family of encoders-decoders that achieve E_{opt}. This turns out to be a correct observation and is developed further in the following section.

Horstein [17] considered a slight variation of the problem for the special case of the binary symmetric channel. Instead of the message decoding error he considered the bit decoding error. He was able to show that, on the average, one could achieve a reliability of orders of magnitude better than the sphere packing bound. In particular, Horstein claims significant reliability in the near neighborhood of capacity rather than having the reliability go to zero. However, in his approach, the 'per bit' transmission time is counted only after a bit becomes active at the encoder. He did not count the time from when the bit became available there.

2.3 An asymptotically optimal variable-length block coding technique for the Binary Symmetric Channel

We describe a very simple 'time-sharing' coding method which achieves the optimal variable-length block coding exponent E_{opt} for a Binary Symmetric Channel with cross-over probability $0 \leq \epsilon < 1/2$. The material described here was originally sketched out and the results conjectured by A. Sahai in a discussion with the author. After this text was written Sahai discovered that this method was previously described in Yamamoto and Itoh [37, 38].

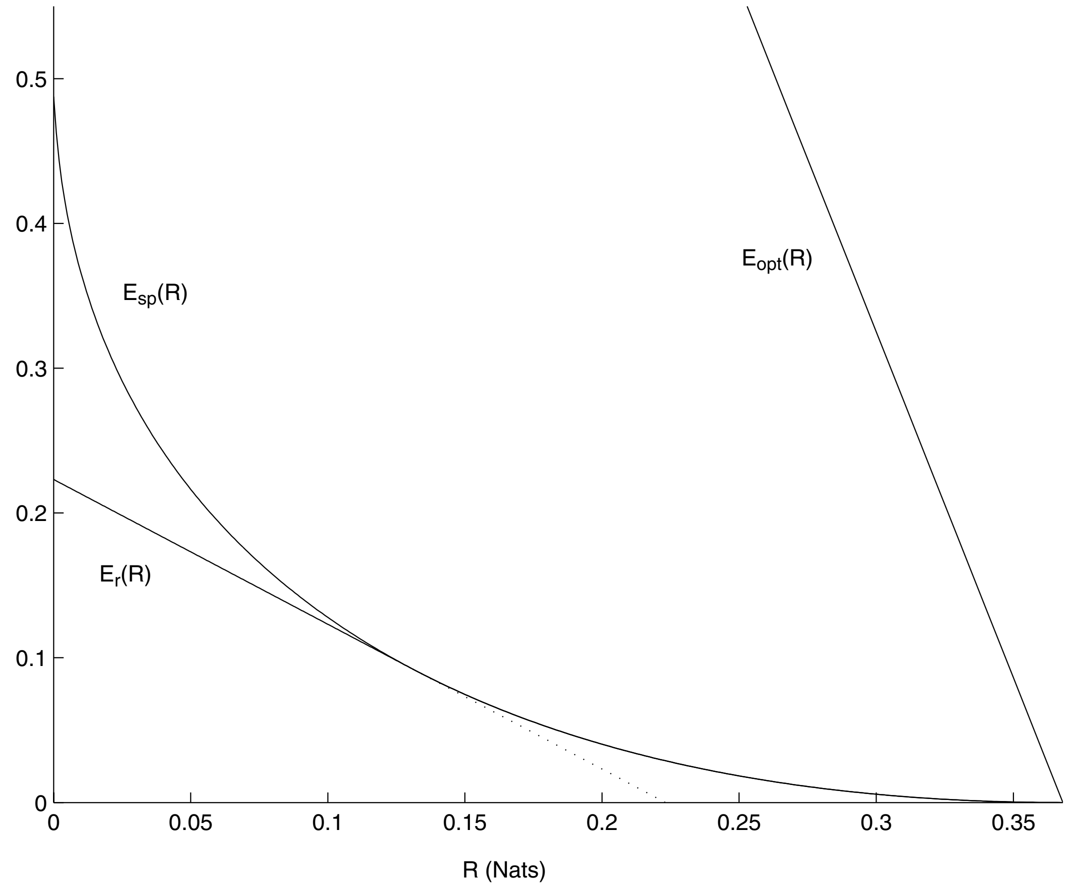

Figure 2.1: The reliability function of a Binary Symmetric Channel ($\epsilon = .1$) with feedback.

Let $0 \le \bar{R} \le C$ be a given average rate of transmission. Let $N, \rho > 1$ and $\delta > 0$ be parameters to be determined later. For simplicity we will always assume that $2^{\bar{R}N}$ is an integer. For notational convenience we also introduce an auxiliary parameter L defined as,

$$L = N - \frac{\bar{R}N}{C-\delta}.$$

The encoder begins by transmitting the message m using a length $N-L$ block code that works well at rates near the channel capacity. Based on what it has received, a maximum-likelihood (ML) decoder produces an estimate m' of the true message m. Observe that the encoder can also determine m'. If $m' = m$ then the encoder sends L ones to indicate a 'confirmation' and if $m' \neq m$ it sends L zeros to indicate a 'deny'. At time N, the decoder makes a decision in favor of message m' if $\big(L - \lceil \rho\epsilon L \rceil\big)$ or more ones were received. Otherwise the decoder makes no decision. This situation is illustrated in Figure 2.2. The procedure continues in transmission of blocks of length N. At time kN, if no decision has been made then the decoder forgets about what it has observed up to time kN and the procedure is repeated starting at time $kN+1$.

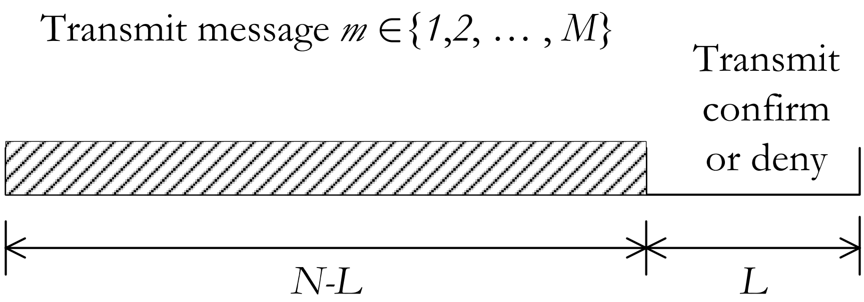

Figure 2.2: Illustration of one block of transmission for the asymptotically optimal encoder-decoder.

Let Q denote the probability of the decoder making a decision at time N and write $P = 1 - Q$. Observe that at time N no decision is made either if (a) $m' = m$ and $\rho\epsilon L$ or more bits were flipped by the binary symmetric channel in the subsequent transmission of the confirmation signal or (b) $m' \neq m$ and fewer than $(L - \rho\epsilon L)$ bits were flipped by the channel in the subsequent transmission of the deny signal. This translates to the following bound,

$$\begin{aligned}
P &\le \mathbb{P}\big(m' = m \mid \text{m was transmitted}\big) \times \sum_{i=\lfloor \rho\epsilon L \rfloor}^{L} \binom{L}{i} \epsilon^i (1-\epsilon)^{L-i} \\
&\quad + \mathbb{P}\big(m' \neq m \mid \text{m was transmitted}\big) \times \sum_{i=0}^{L-\lfloor \rho\epsilon L \rfloor} \binom{L}{i} \epsilon^i (1-\epsilon)^{L-i} \\
&\le \sum_{i=\lfloor \rho\epsilon L \rfloor}^{L} \binom{L}{i} \epsilon^i (1-\epsilon)^{L-i} + \mathbb{P}\big(m' \neq m \mid \text{m was transmitted}\big) \\
&\le e^{N\left[\frac{\lfloor \rho\epsilon L \rfloor}{L} \log \frac{L\epsilon}{\lfloor \rho\epsilon L \rfloor} + \left(\frac{L - \lfloor \rho\epsilon L \rfloor}{L}\right) \log \frac{L(1-\epsilon)}{L - \lfloor \rho\epsilon L \rfloor}\right]} + P_e^b
\end{aligned}$$

$$\leq 2\max\left\{e^{N\left[\frac{\lfloor\rho\epsilon N\rfloor}{N}\log\frac{N\epsilon}{\lfloor\rho\epsilon N\rfloor}+(\frac{N-\lfloor\rho\epsilon N\rfloor}{N})\log\frac{N(1-\epsilon)}{N-\lfloor\rho\epsilon N\rfloor}\right]},\ P_e^b\right\} \tag{2.4}$$

where we used the Chernoff bound (pp. 531, [14]) to approximate the summation and we wrote P_e^b to denote the probability $\mathbb{P}\big(m' \neq m \mid \text{m was transmitted}\big)$ of the block decoding error for message m. By the random coding bound this probability is bounded by,

$$\begin{aligned} P_e^b &\leq e^{-(N-L)E_r(\frac{\bar{R}N}{N-L})} \\ &= e^{-(N-L)E_r(C-\delta)} \end{aligned}$$

where E_r is Gallager's random coding exponent and we know that $E_r(x) > 0$ for all $x < C$.

Based on the bound on P we can calculate approximate the distribution of the decision time T. Because the decoder interprets the k-th block independently of the first $k-1$ blocks it has received it follows that the probability of making a decision at time kN is given by,

$$\mathbb{P}\big(T = kN\big) = P^{k-1}Q,\ k = 1, 2, \cdots$$

which we recall has the generating function,

$$\mathbb{E}e^{\theta T} = \frac{Qe^{\theta n}}{1 - Pe^{\theta N}},\ (e^{\theta N}P < 1), \tag{2.5}$$

and the mean,

$$\bar{T} = \frac{N}{1-P}.$$

Theorem 2.3.1 *For all $\delta > 0$ and $\rho > 1$ there exists a sufficiently large N such that,*

$$N < \bar{T} < N + \delta. \tag{2.6}$$

Proof:

Recall the log-sum inequality which states that for non-negative numbers $a_1, a_2, \cdots, a_n$ and $b_1, b_2, \cdots, b_n$ the following relationship holds,

$$\sum_{i=1}^{n} a_i \log\frac{b_i}{a_i} \leq \left(\sum_{i=1}^{n} a_i\right)\log\frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} a_i}$$

with equality if and only if a_i/b_i are constant for $i = 1, 2, \cdots, n$. Take,

$$\begin{aligned} a_1 &= \frac{\lfloor\rho\epsilon N\rfloor}{N}, \\ a_2 &= \frac{L - \lfloor\rho\epsilon N\rfloor}{N}, \\ b_1 &= \epsilon, \\ b_2 &= 1 - \epsilon, \end{aligned}$$

and use the log-sum inequality to show that,

$$\frac{\lfloor\rho\epsilon N\rfloor}{N}\log\frac{N\epsilon}{\lfloor\rho\epsilon N\rfloor} + \left(\frac{N - \lfloor\rho\epsilon N\rfloor}{N}\right)\log\frac{N(1-\epsilon)}{N - \lfloor\rho\epsilon N\rfloor} \leq 0. \tag{2.7}$$

Now observe that $\lfloor \rho\epsilon N \rfloor / N$ tends to $\rho\epsilon$ as N tends to infinity. So we can pick N large enough such that $\lfloor \rho\epsilon N \rfloor / N$ is strictly larger than ϵ. Consequently we can pick N large enough so that $a_1/b_1 \neq a_2/b_2$. This means that the inequality (2.7) is strict. In other words, we can select N large enough so that,

$$Ne^{N\left[\frac{\lfloor \rho\epsilon N \rfloor}{N}\log\frac{N\epsilon}{\lfloor \rho\epsilon N \rfloor}+(\frac{N-\lfloor \rho\epsilon N \rfloor}{N})\log\frac{N(1-\epsilon)}{N-\lfloor \rho\epsilon N \rfloor}\right]} > \delta/4 < 1/2. \tag{2.8}$$

Then, by observing that the random coding exponent $E_r(C-\delta)$ is strictly positive and observing also that $N - L = \frac{\bar{R}N}{C-\delta}$ we can pick N even larger, if necessary, so that,

$$\begin{aligned} N\,P_e^b &\leq Ne^{-(N-L)E_r(C-\delta)} \\ &= Ne^{-N\frac{\bar{R}E_r(C-\delta)}{C-\delta}} \\ &\leq \delta/4 < 1/2. \end{aligned} \tag{2.9}$$

Finally, select N large enough to satisfy both inequalities (2.8) and (2.9) the desired result follows by,

$$\begin{aligned} N < \bar{T} &= \frac{N}{1-P} \\ &\leq N(1+2P) \\ &= N + 2NP \\ &< N + \delta. \end{aligned}$$

□

This is the first result we were after. We turn our attention now to the probability of decoding error. First consider the probability of the decoder making a decision error $m' \neq m$ at time N. For this to happen first there must have been block decoding error in the first $N-L$ time slots and second the decoder must have understood the 'deny' signal as a 'confirm' signal. For the latter the binary symmetric channel must have flipped $(L - \lfloor \rho\epsilon L \rfloor)$ or more bits in the L time slots during which the deny signal was being transmitted. By Bayes law,

$$\begin{aligned} &\mathbb{P}\big(m' \neq m \mid \; m \text{ was transmitted, decision made at time } N\;\big) \\ &\qquad\qquad\qquad = \frac{P_e^b \times \sum_{\ell=L-\lfloor \rho\epsilon L \rfloor}^{L} \epsilon^\ell (1-\epsilon)^{L-\ell}}{Q} \end{aligned}$$

Because the decoder interprets the k-th block independently of the first $k-1$ blocks it had received the probability for the decoder ever making a decision error is given by the geometric series,

$$\begin{aligned} P_e &= \sum_{k=1}^{\infty} \frac{P_e^b \times \sum_{\ell=L-\lfloor \rho\epsilon L \rfloor}^{L} \epsilon^\ell (1-\epsilon)^{L-\ell}}{Q}\mathbb{P}\big(T = kN\big) \\ &\leq \sum_{k=1}^{\infty} \frac{\sum_{\ell=L-\lfloor \rho\epsilon L \rfloor}^{L} \epsilon^\ell (1-\epsilon)^{L-\ell}}{Q}\mathbb{P}\big(T = kN\big) \\ &\leq \frac{e^{L\left[(\frac{L-2\lfloor \rho\epsilon L \rfloor}{L})\log\frac{\lfloor \rho\epsilon L \rfloor}{L(1-\epsilon)}\right]}}{1-P} \end{aligned} \tag{2.10}$$

where the last inequality follows from the Chernoff bound. This leads to the following result.

Theorem 2.3.2 *For all $\delta > 0$ and $\rho > 1$ we may select L large enough so that,*

$$P_e \leq e^{L\left[(1-2\rho\epsilon)\log\frac{\rho\epsilon}{1-\epsilon}+\delta\right]}(1+\delta). \tag{2.11}$$

where $L = N - \frac{N\bar{R}}{C-\delta}$.

Proof: First follow the steps in the proof of Theorem 2.3.1 to show that for large N there exists a $\delta' > 0$ such that $P < e^{-\delta' N}$. Next, we use the fact that $\lfloor \rho\epsilon N \rfloor / N$ can be selected arbitrarily close to $\rho\epsilon$ by choosing N even larger. So for large enough N we have,

$$\left(\frac{N - 2\lfloor \rho\epsilon N \rfloor}{N}\right)\log\frac{\lfloor \rho\epsilon N \rfloor}{N(1-\epsilon)} \leq (1-2\rho\epsilon)\log\frac{\rho\epsilon}{1-\epsilon} + \delta \tag{2.12}$$

Finally, we select N even larger, if necessary, so that,

$$\begin{aligned}\frac{1}{1-P} &\leq 1 + 2P \\ &\leq 1 + \delta'\end{aligned} \tag{2.13}$$

The desired result follows by using (2.12) and (2.13) in (2.10). □

This gives us the second result we were after. Finally write $\bar{T}(\rho, \delta)$ and $P_e(\rho, \delta)$ to denote the mean decision time and probability of decoding error which is determined by having specified $\rho > 1$ and $\delta > 0$ and having selected L large enough to satisfy both inequalities (2.6) and (2.11). It now follows that,

$$\lim_{\substack{\rho \to 1, \\ \delta \to 0}} \frac{\log P_e(\rho, \delta)}{\bar{T}(\rho, \delta)} = (1-2\epsilon)\log\frac{\epsilon}{1-\epsilon}\left(1 - \frac{\bar{R}}{C}\right). \tag{2.14}$$

By Burnashev [5] this is the largest achievable variable-length block coding exponent for the Binary Symmetric Channel.

We also draw attention to the importance of the transmission of the single bit that represents the 'confirm' or 'deny' signals. The optimal exponent in (2.14) may be viewed as a time-sharing between the rate at which we transmit the confirmation signal and the rate at which we transmit the true message.

2.3.1 The more general case

The previous result also holds for more general discrete memoryless channels (DMC). In the following paragraphs we give a sketch of the proof. Consider an arbitrary DMC $[p_{ij}]$ and an encoder-decoder pair that works in two phases as illustrated in Figure 2.2. The encoder is fed a message $m \in \{1, 2, \cdots, M\}$ which it transmits to the decoder by means of a 'good' length $N - L$ block code at a rate $R = C - \delta_1$ close to capacity. Based on what it has received, the decoder produces an estimate m' of m. This is the first phase of the transmission method.

Recall that the encoder can also calculate m' and let i and k be the channel letters to be determined later. During the second phase of transmission, if $m' = m$, the encoder 'confirms' the decoders estimate by transmitting L copies of the i-th letter. If $m' \neq m$, the decoder 'denies' the estimate by transmitting L copies of the k-th letter. Based only on the L control letters it has received, the decoder decides either in favor of 'confirm' or 'deny'. If 'confirmed', the estimate m' is finalized and transmission terminates. On the other hand, if 'denied', the decoder forgets about everything it has received and the entire procedure is repeated.

Observe that an error occurs only if the estimate m' was different from m and a subsequent 'deny' was misunderstood as a 'confirm' (ie. false-positive). On the other hand, the transmission repeats either if the estimate m' was accurate but the subsequent 'confirm' was misunderstood as a 'deny' (false-negative) or if the estimate m' was different from m and a subsequent deny was correctly decoded. By Sanov's Theorem [8], there exists a decoder for which the probability of a false-positive and of a false-negative is bounded as,

$$\begin{aligned} \mathbb{P}\big(\text{false positive}\big) &\leq \delta_2 \\ \mathbb{P}\big(\text{false negative}\big) &\leq 2^{C_1(i,k)L} \end{aligned} \tag{2.15}$$

and

$$\mathbb{P}\big(m' \neq m\big) \leq \delta_3$$

where $\delta_2, \delta_3 > 0$ can be made arbitrarily small by choosing N large and $C_1(i,k)$ is the relative entropy of i and k,

$$C_1(i,k) = \sum_{\ell} p_{i\ell} \log \frac{p_{i\ell}}{p_{k\ell}}.$$

On the other hand, because the transmission of blocks are independent, the average total transmission time is the mean of a geometric random variable,

$$\bar{T} \approx \frac{N}{1 - \delta_2 - \delta_3}. \tag{2.16}$$

So the average rate of transmission is calculated to be,

$$\begin{aligned} \bar{R} &= \frac{R(N-L)}{\bar{T}} \\ &= (C - \delta_1)(1 - \delta_2 - \delta_3)\left(1 - \frac{L}{N}\right) \end{aligned} \tag{2.17}$$

Finally, by selecting $C_1 = \max_{i,k} C_1(i,k)$ and using (2.17) in (2.15) we obtain the following bound for the probability of decoding error,

$$\begin{aligned} P_e &\leq 2^{C_1 L} \\ &= 2^{C_1 N \frac{L}{N}} \\ &= 2^{C_1 N \left(1 - \frac{\bar{R}}{(C-\delta_1)(1-\delta_2-\delta_3)}\right)}. \end{aligned} \tag{2.18}$$

It then follows that the proposed method can approach Burnashev's bound arbitrarily closely by choosing N large.

An interesting situation occurs if at least one of the channel transition probabilities $p_{i^\star k^\star}$ is identically zero. In this case we design the 'confirmation' signal to consist of L copies of the letter $i^\star$ and we design the 'deny' signal to consist of L copies of the letter $k^\star$. Because $p_{i^\star k^\star} = 0$ we can design the decoder so that the probability of a false positive is identically zero. This can be done by selecting the decoder so that it never decides in favor of 'confirm' unless it sees at least one $k^\star$. The implication of this result is that any desired level of reliability is achievable for $0 \leq \bar{R} \leq C$. For example, the achievable reliability region for a Binary Erasure Channel is illustrated in Figure 2.3.

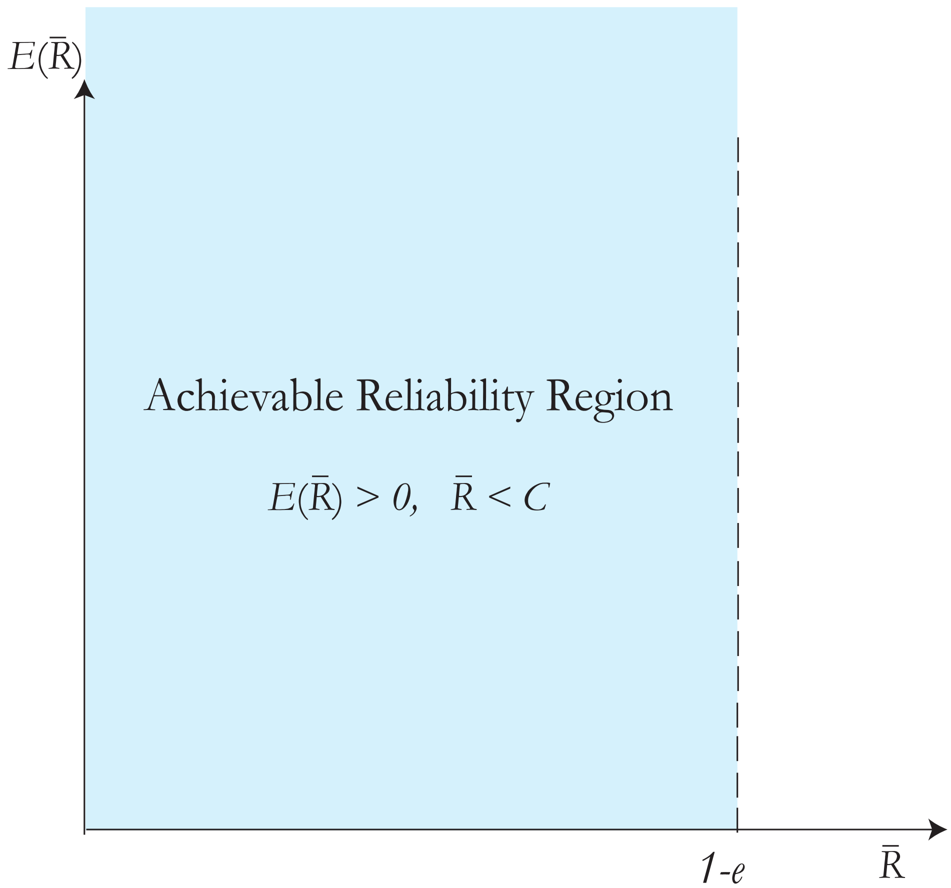

Figure 2.3: The achievable reliability region for a Binary Erasure Channel with erasure probability $0 \leq e \leq 1$.

2.4 Discussion

We presented two basic results for discrete memoryless channels. The first is a negative result says that feedback cannot be used to improve the reliability function of a family of fixed-length block codes. The second result gives an upper bound on the reliability function of any family of feedback codes. The bound has three important properties: (1) It is easy to calculate, (2) It is achievable by means of a simple time-sharing argument and (3) It is significantly larger than the sphere-packing bound and can even be infinite. We focused on these two results to establish how feedback can be used to improve the asymptotic performance of a feed-forward system and the adversarial role of the traditional fixed-length block coding framework.

There are many other ways in which feedback can be used to improve the performance. Before we conclude we would like to touch on two such 'practical' uses.

2.4.1 Schalkwijk's one-up J-down method

In a practical setting we are necessarily lead to think about 'short' codes with small delays. In such a situation the previously described transmission methods are not useful. In this section we describe a very simple coding technique which leads to dramatic improvements for very short average delays. This technique will also be useful in subsequent chapters in the design of 'good' anytime codes. The material presented here was developed independently by the author but then turned out to be a special case of Schalkwijk's one-up J-down method known in the literature.

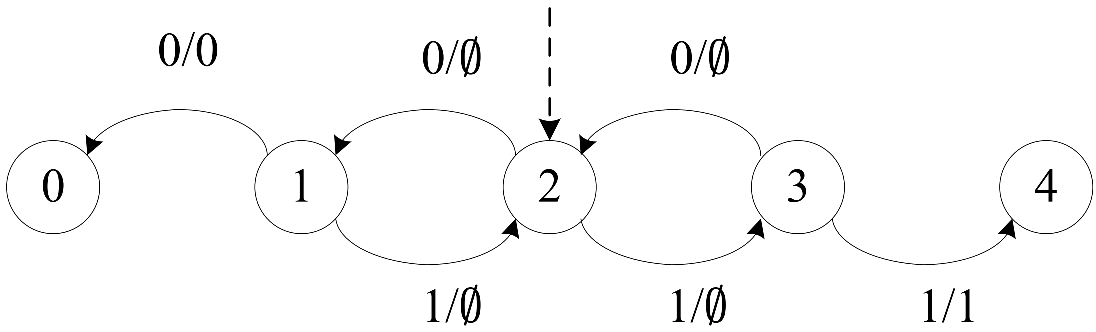

Figure 2.4: Illustration of 1-up J-down method for $J = 2$.

For simplicity of presentation consider a Binary Symmetric Channel with cross-over probability $0 \leq \epsilon < 1/2$ and let the message set be $m \in \{0, 1\}$. The encoder repeats its binary message m until the decoder makes a decision at a random time T. Because of feedback this time is known to the encoder.

Let J be a positive integer to be determined later. The decoder is described by a Mealy finite state machine consisting of $2J + 1$ states numbered as $0, 1, 2, \cdots, 2J$. The decoder is illustrated in Figure 2.4. Initially, the decoder is in state J. Suppose at time n the decoder is in state $z_n \in \{1, 2, \cdots, 2J - 1\}$. If a zero is received the decoder jumps to state $z_{n+1} = z_n - 1$ and if a one is received the decoder jumps to state $z_{n+1} = z_n + 1$. At time n the decoder makes a decision in favor of message $m = 0$ if $z_{n+1} = 0$ or it makes a decision in favor of message $m = 1$ if $z_{n+1} = 2J$. Otherwise the decoder makes no decision and outputs the symbol $\emptyset$ to indicate that no decision has been made. The decision time T is given by the following stopping time,

$$T = \min \{ n \geq J \, : \, z_{n+1} = 0 \text{ or } z_{n+1} = 2J \}.$$

Observe that a decision can be made no earlier than J time steps.

The structure of the binary symmetric channel induces a natural probability model for the decoder. Letting $0 \leq \epsilon < 1/2$ to denote the BSC cross-over probability we can think of the decoder as a finite Markov chain. Letting Z_n denote the state of the chain at time n we write the one-step transition probabilities,

$$\begin{aligned} q_{i,i+1}(m) &= \mathbb{P}(Z_{n+1} = i + 1 \mid Z_n = i, m) \\ &= m - \epsilon \\ q_{i,i-1}(m) &= \mathbb{P}(Z_{n+1} = i - 1 \mid Z_n = i, m) \\ &= m - (1 - \epsilon) \end{aligned}$$

for $m = 0, 1$ and $i = 1, 2, \cdots, 2J - 1$. The decision states zero and $2J$ are modeled as absorbing states.

To calculate the probability of decoding error we need to calculate the probability of ending up in the absorbing state zero when the message $m = 1$ or equivalently the probability of ending up in the absorbing state $2J$ when the message $m = 0$. This is a standard exercise in a first course in stochastic processes (pp.26, [18]). We reorder the states as $\{0, 2J, 1, 2, \cdots, 2J-1\}$ so that the two absorbing states are up front. Based on this ordering we write the one-step transition matrix $Q(m)$,

$$Q(m) = \begin{bmatrix} I_{2\times 2} & | & 0 \\ - - - - & - & - - - - \\ S(m) & | & R(m) \end{bmatrix}, \; m = 0, 1,$$

where

$$S(m) = \begin{bmatrix} m-(1-\epsilon) & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & m-\epsilon \end{bmatrix}, \; m = 0, 1,$$

and

$$R(m) = \begin{bmatrix} 0 & m-\epsilon & & & & \\ m-(1-\epsilon) & 0 & m-\epsilon & & & \\ & m-(1-\epsilon) & 0 & m-\epsilon & & \\ & & & \ddots & & \\ & & & & m-(1-\epsilon) & 0 \end{bmatrix}, \; m = 0, 1.$$

The matrix $S(m)$ is of dimension $2J-1 \times 2$ and $R(m)$ is of dimension $2J-1 \times 2J-1$. Given that the chain starts in state J the probability of ending up in state zero for message $m=1$ is known to be the $(J,1)$-th entry for the matrix $(I-R(1))^{-1}S(1)$. Consequently,

$$P_{e,m=1} = (N, 1)\text{-th entry of } \big(I - R(1)\big)^{-1} S(1). \tag{2.19}$$

Because of symmetry we also have that $P_{e,m=0} = P_{e,m=1}$. The inverse $\big(I-R(1)\big)^{-1}$ exists because Q is a probability matrix.

The mean decoding time is calculated along the same lines. We consider the mean time until the chain enters either the absorbing state zero or the absorbing state $2J$. Following the steps above we obtain,

$$\begin{aligned} \bar{T} &= \mathbb{E}T \\ &= \sum_{i=1}^{2J-1} \big(I - R(0)\big)^{-1}_{J,i}. \end{aligned} \tag{2.20}$$

This completes our calculations. To evaluate the results we consider the case of $J=2$ and a Binary Symmetric Channel with cross-over probability $\epsilon = 10^{-3}$. Using these parameters in (2.19) and (2.20) gives,

$$P_{e,m}\Big|_{\epsilon=10^{-3}, N=2} = 1.002 \times 10^{-6}, \; m = 0, 1.$$

and

$$\bar{T}\Big|_{\epsilon=10^{-3}, N=2} = 2.004.$$

So for $J=2$ the 1-up 2-down method gives a probability of decoding error of approximately ϵ^2. It should be clear that we can do no better. This result stands in sharp contrast to the case of block coding which is useless for a block length of two.

2.4.2 Limited feedback

In the preceding discussion we assumed a 'perfect' feedback channel and 'complete information' at the encoder and we used this to achieve the tight performance bounds of Theorem 2.2.1. It is also natural to wonder whether we can use feedback to achieve such significant performance gains subject to practical constraints on feedback channel and, consequently, limitations on the information known to the encoder. Surprisingly, it turns out that even very limited feedback can be used to drastically improve the performance of a feed-forward system. In the following paragraphs we make this statement precise.

Consider a rate R (nats) block-code of length N that consists of the $M = e^{NR}$ codewords $x_1, x_2, \cdots, x_M$. Based on this $(N, M = e^{NR})$ block-code, transmission occurs as follows. A message $m \in \{1, 2, \cdots, M\}$ enters the encoder which it subsequently emits using the m-th codeword x_m. At time N, upon receiving the word y, the decoder produces the estimate m' of m if there exists a true codeword such that,

$$\forall m'' \neq m', \; \frac{\mathbb{P}\big(x_{m'}, y\big)}{\mathbb{P}\big(x_{m''}, y\big)} > e^{N\rho}$$

where $\rho \geq 0$. If no such codeword exists, the decoder forgets what it has received and requests a retransmission. The transmission repeats in this fashion until the decoder produces an estimate. The parameter ρ governs the relative magnitudes of the probability of decoding error and the probability of a retransmission: as ρ increases, the probability of a decoding error is reduced but retransmissions are more likely. This scheme, originally due to Forney [13], describes a variable-length method of transmission with the following property,

Theorem 2.4.1 (Forney, [13]) *For any $\delta > 0$ and rate $R < C$ there exists a decision threshold $\rho \geq 0$ and a block-code (N, e^{NR}) such that,*

$$C_0 \leq R - \delta \leq \bar{R} \leq C,$$
$$P_e \leq e^{-N(E_f - \delta)}$$

in which the exponent E_f is given by the following expression,

$$E_f(\bar{R}) = E_{sp}(\bar{R}) - \bar{R} + C,$$

and C_0 is the rate at which E_{sp} becomes infinite.

There are two things to keep in mind. First, the achievable exponent E_f is strictly larger than the sphere-packing bound over the entire spectrum of finite values. In particular, E_f is significantly superior at rates close to capacity. To see this, observe that as $\bar{R}$ tends to C, E_f approaches zero with a slope of negative one. In contrast, the sphere-packing curve approaches zero with a slope that also tends to zero. Second, only one bit of feedback is used for each block transmitted from the encoder to the decoder. Consequently, this demonstrates that even the 'least' amount of feedback information can be used to significantly improve the performance of the feed-forward system.

Chapter 3

Anytime channel coding with feedback

We formally present the anytime channel coding problem. We formulate the problem as an encoder-decoder design problem. The objective is to design encoder-decoder pairs with large 'exponents'. In order to evaluate the exponent we present lower and upper bounds that depend only on the channel. The main contribution of this chapter is to find an upper-bound. We conclude by reviewing a method of design that will be used extensively in subsequent chapters.

3.1 Problem set-up

A source produces one bit s_j of new information every $1/R$ time units for a prescribed rate $R > 0$. These bits are fed into an encoder. At time n ($n \geq 0$), based on what it has seen so far, the encoder emits a channel symbol a_n. At the same time n a decoder receives b_n. Based on what it has seen up to time n the decoder produces an estimate $\hat{s}(n)$ for 'all' of the true source bits s_1 through s_{j_n} where j_n is the number of bits produced by the source up to time n,

$$j_n = \lceil nR \rceil.$$

Observing that $\hat{s}(n)$ is a vector, we write,

$$\hat{s}(n) = \big(\hat{s}_1(n), \hat{s}_2(n), \cdots, \hat{s}_{j_n}(n)\big). \tag{3.1}$$

There is a perfect feedback link from the channel output to the encoder. So the encoder is described by functions,

$$\begin{aligned} a_n &= \mathcal{E}_n\big(s_1, s_2, \cdots, s_{j_n}, b_1, b_2, \cdots, b_{n-1}\big) \\ &\equiv \mathcal{E}_n\big(s_1^{j_n}, b_1^{n-1}\big), \end{aligned}$$

where we use the notation $s_1^{j_n} = (s_1, s_2, \cdots, s_{j_n})$ and similarly for other variables a, b, etc... The channel is described by a memoryless probability distribution,

$$\begin{aligned} &\mathbb{P}\big(b_n = b \mid a_n = a, a_{n-1}, a_{n-2}, \cdots, a_1, b_{n-1}, b_{n-2}, \cdots, b_1\big) \\ &\quad = \mathbb{P}\big(b_n = b \mid a_n = a\big) \\ &\quad = p_{ij}, \quad a = 1, 2, \cdots, A,\ b = 1, 2, \cdots, B. \end{aligned} \tag{3.2}$$

The integers A and B denote the number of distinct channel inputs and outputs respectively. Finally, the decoder is described by functions,

$$\begin{aligned} \hat{s}(n) &= \mathcal{D}_n\big(b_1, b_2, \cdots, b_n\big) \\ &= \mathcal{D}_n\big(b_1^n\big). \end{aligned}$$

The set-up is illustrated in Figure 3.1.

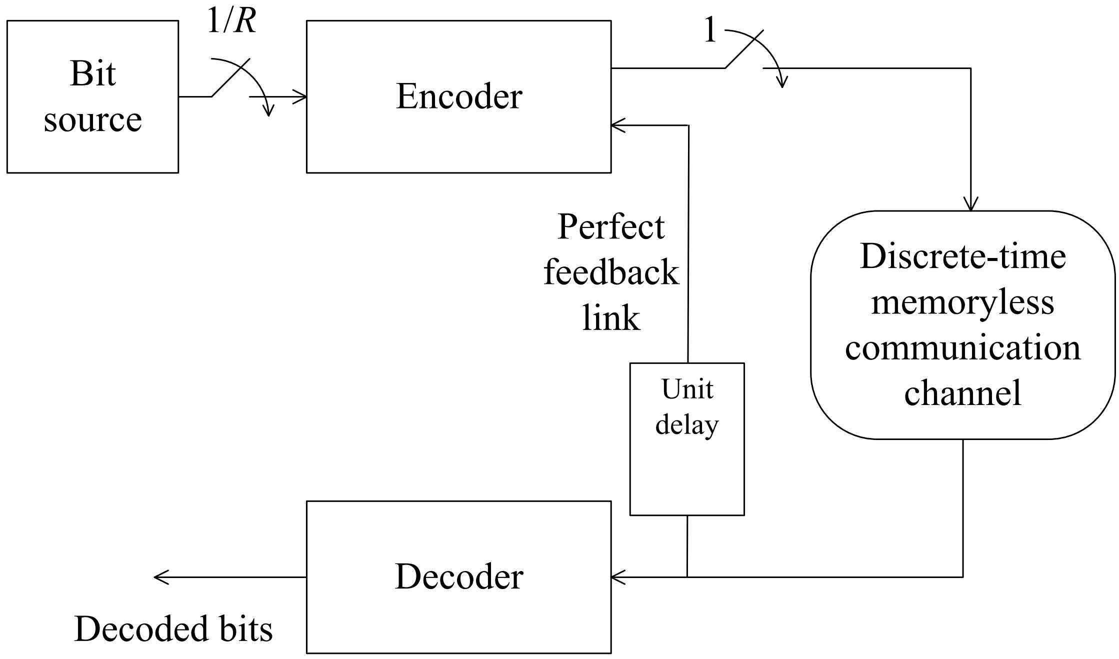

Figure 3.1: A block diagram illustrating the set-up of anytime coding with feedback.

For a prescribed communication channel and rate $R > 0$ the basic problem is to design an encoder-decoder pair $(\mathcal{E}, \mathcal{D})$ such that for some $\alpha, \beta > 0$ the probability of estimation error is bounded as,

$$\forall\big(n \geq 0, j \leq j_n \text{ and } s_1^{j_n}\big),\ \mathbb{P}\big(\hat{s}_j(n) \neq s_j \mid s_1^{j_n} \text{ was transmitted}\big) \leq \beta 2^{-\alpha d(n,j)} \tag{3.3}$$

where $d(n, j)$ denotes the time delay,

$$d(n, j) = n - j/R.$$

To keep the notation compact we write,

$$P_{e,s}\big(n, j\big) = \mathbb{P}\big(\hat{s}_j(n) \neq s_j \mid s_1^{j_n} \text{ was transmitted}\big).$$

The inequality in (3.3) is the anytime decoding property. The decoder is capable of producing an estimate of any bit b_j at any time $n > j/R$. The larger the time difference the better the quality of the estimate. An encoder-decoder pair $(\mathcal{E}, \mathcal{D})$ with this property is called an anytime encoder-decoder or simply an anytime code.

Remark 3.1.1 *In a more general sense, we may consider replacing the* $2^{-\alpha d(n,j)}$ *in (3.3) by some function* $f\big(d(n,j)\big)$ *that tends to zero as* $d(n,j)$ *tends to infinity. We will not do so here.*

Remark 3.1.2 *An anytime algorithm is an asymptotically convergent calculation that can be stopped at any time to produce an approximate answer to a difficult question. The later the algorithm is stopped, the closer the answer is to the asymptote. Anytime algorithms are in general non-trivial but are of interest because they offer a mechanism to trade off calculation time vs. quality of answer.*

For a given channel and rate R encoder-decoder $(\mathcal{E}, \mathcal{D})$, the largest exponent α for which inequality (3.3) holds is called the anytime exponent and is denoted as $E_a(R)$. Observe that the anytime exponent really depends on the channel and encoder-decoder. The notation $E_a(R)$ is used because the encoder-decoder is assumed to be associated with the rate R and the channel is assumed to be clear from the context.

Definition 3.1.1 *For a given channel and rate R encoder-decoder the anytime exponent is,*

$$E_a(R) = \sup_{\alpha} \left\{ \exists \beta > 0 \ s.t. \ \forall \big(n \geq 0, j \leq j_n, s_1^{j_n}\big), \ P_{e,s}(n,j) \leq \beta 2^{-\alpha d(n,j)} \right\}$$

where $d(n,k)$ *denotes the discrete time delay,*

$$d(n,j) = n - j/R.$$

Conversely, for a given exponent $\alpha > 0$ *we define the anytime capacity of the channel as,*

$$C^{\mathcal{A}}(\alpha) = \sup_{R \geq 0} \Big\{ \textit{Exists a } \beta > 0 \textit{ and rate } R \textit{ encoder-decoder } (\mathcal{E}, \mathcal{D}) \ s.t. \quad \forall \big(n \geq 0, j \leq j_n, s_1^{j_n}\big), \ P_{e,s}(n,j) \leq \beta 2^{-\alpha d(n,j)} \Big\}$$

The basic objective is to design good rate R anytime codes for which the anytime exponent $E_a(R)$ is large over a large range of R. For this objective to make sense we really need bounds on the anytime exponent. So given a channel we would like to have two functions $E_a^-(R)$ and $E_a^+(R)$ such that there exists anytime codes with exponents greater than or equal to $E_a^-(R)$ and such that the exponent of any coding scheme is less than or equal to $E_a^+(R)$. These two functions should depend only on the channel.

Sahai's lower bound holds for an arbitrary discrete-time memoryless channel without feedback. This bound remains valid with feedback since feedback can only improve the anytime exponent. This lower bound is presented in the following section. We are not aware of any known upper bounds. Following the review of Sahai's lower bound we design an upper bound using a time-sharing argument. The rest of the chapter is devoted to understanding the properties of our upper bound.

Before continuing we encourage the reader to compare the definition of the anytime exponent E_a with the way the reliability function is defined in Section 2.1. Observe that the reliability function is defined over a family of block codes of increasing block length. By contrast, the anytime exponent is associated with a single encoder-decoder pair.

3.2 The random coding bound

Beyond defining the concept of Anytime information theory, Sahai was able to bound the anytime capacity of an arbitrary discrete-time memoryless channel by a quantity well-known to information theorists. Specifically, for any given discrete-time memoryless channel and for any rate $0 \le R \le C$, there exists an encoder-decoder pair such that,

1. the encoder-decoder has access to a common source of randomness,
2. there is no feedback link from the decoder to encoder,
3. for all $0 \le R \le C$, $E_a(R) \ge E_r(R \log 2)/\log 2$,

where $E_r(R)$ is Gallager's random coding exponent (cf. Section 5.6, [14]) also discussed in Chapter 2. The 'well-known' quantities are, of course, the Shannon capacity C and the exponent $E_r(R)$.

Remark 3.2.1 *The log 2 factor is used to convert between the units of 'nats' and 'bits'. For cleanliness of the presentation the units of nats is preferable. To see this, compare the derivatives of e^x and 2^x. However, in our setup, because bits, rather than messages, enter into the decoder, there is no convenient way to define the anytime exponent and capacity in Nats.*

Without common randomization a slightly weaker result is true. For any discrete-time memoryless channel and rate $0 \le R \le C$, there exists an encoder-decoder pair such that,

1. the encoder-decoder does not have access to a common source of randomness,
2. there is still no feedback link from the decoder to encoder,
3. exists $\beta > 0$ such that for all $n \ge 0$ and $j \le j_n$,

$$\mathbb{E}_s P_{e,s}(n,j) \le \beta 2^{-d(n,j)E_r(R\log 2)/\log 2} \tag{3.4}$$

where the expectation is taken over all possible source messages. So without common randomization we cannot bound the probability of error uniformly as required by Definition 3.1.1. At each time n, there may be source sequences $s_1^{j_n}$ and $j < j_n$ for which the probability of decoding error $P_{e,s}(n,j)$ is much larger than the average in (3.4).

We know that E_r is strictly positive for any rate R strictly less than capacity. In this sense, the most important point about the above mentioned results is that there exists an anytime code with a strictly positive exponent for all rates strictly less than capacity. The converse follows from classical results: There is no code for which the anytime exponent is positive at rates above capacity.

3.3 A time-sharing upper bound

Consider the following scenario. We are given a rate R source $\{s_j\}$ and anytime encoder-decoder pair with anytime exponent $E_a(R)$. At some randomly selected time n we select a delay λn ($0 \le \lambda \le 1$) and ask the anytime decoder to produce an estimate $\hat{s}_1(n), \hat{s}_2(n), \cdots, \hat{s}_{j_{(1-\lambda)n}}(n)$ of the true encoded bits $s_1, s_2, \cdots, s_{j_{(1-\lambda)n}}$. By Definition 3.1.1, the probability of an error in one of more of the $j_{(1-\lambda)n}$ estimated bits is bounded as,

$$P_{e,s}\big(n,(1-\lambda)n\big) \le \beta 2^{-E_a(R)\lambda n}. \tag{3.5}$$

Keep in mind that the constant β is independent of the delay λn and in particular that this inequality is true for any choice of the time index n, the delay λn and the true source bits $s_1, s_2, \cdots, s_{j_{(1-\lambda)n}}$.

Consider also the alternative scenario where the decision time n and the number of bits $j_{(1-\lambda)n}$ is fixed ahead of time. In this scenario we are given the 'best' $\left(n, 2^{j_{(1-\lambda)n}}\right)$ block code. Up to time n the source bits $s_1, s_2, \cdots, s_{j_{(1-\lambda)n}}$ are fed into the block encoder. From times n to $2n$ the block encoder communicates these bits over the channel and at time $2n$ the decoder produces an estimate $\tilde{s}_1, \tilde{s}_2, \cdots, \tilde{s}_{j_{(1-\lambda)n}}$ of the true source bits. For even the best such block code the probability of block decoding error is bounded as,

$$P_{e,s} \geq \tilde{\beta} 2^{-nE^+((1-\lambda)R/\log_2 e)\log_2 e} \tag{3.6}$$

where E^+ is the upper bound on the reliability function of the channel (Theorem 2.1.1) and $\tilde{\beta}$ is some constant that is independent of the source bits s and the time index n. The factor $\log_2 e$ is used to convert between the units of Bits and Nats.

Compare the two scenarios. Because of the additional *a priori* knowledge and because the block encoder-decoder is the best possible we have by (3.5) and (3.6),

$$\beta 2^{-E_a(R)\lambda n} \geq \tilde{\beta} 2^{-nE^+((1-\lambda)R/\log_2 e)\log_2 e}$$

By taking logs,

$$E_a(R) \leq \frac{1}{\lambda} E^+\big((1-\lambda)R/\log_2 e\big)\log_2 e - \frac{1}{\lambda n}\log\frac{\tilde{\beta}}{\beta}.$$

But β and β' are both independent of n and n was arbitrary. So for any value of λ and an arbitrarily small $\delta > 0$ we can pick n so large so that,

$$E_a(R) \leq \frac{1}{\lambda} E^+\big((1-\lambda)R/\log_2 e\big)\log_2 e - \delta.$$

Because λ was also arbitrary this proves the following result.

Theorem 3.3.1 *Consider a discrete-time memoryless channel. The exponent of any rate R anytime encoder-decoder pair is bounded as,*

$$\begin{aligned} E_a(R) &\leq \min_{0<\lambda<1} \frac{E^+\big((1-\lambda)R/\log_2 e\big)\log_2 e}{\lambda} \\ &= E_a^+(R),\ 0 \leq R \leq C. \end{aligned}$$

in which E^+ is given by Theorem 2.1.1. If the channel is symmetric then E^+ may be replaced with the sphere-packing bound E_{sp}.

Remark 3.3.1 *Because of the term $(1-\lambda)R$ we call this the 'time-sharing' bound.*

In the following subsections we present some of the basic properties of the upper-bound E_a^+ of Theorem 3.3.1.

3.3.1 A parametric representation for symmetric channels

Consider a symmetric channel with channel transition matrix $[p_{ij}]$. For such a channel, the upper bound E_a^+ on the anytime exponent is expressed in terms of the sphere-packing bound as follows,

$$\begin{aligned} E_a^+(R) &= \min_{0<\lambda<1} \frac{E_{sp}\big((1-\lambda)R/\log_2 e\big)\log_2 e}{\lambda} \\ &= \min_{0<\lambda<1} \left[\max_{\rho>0} \frac{1}{\lambda}\left(E_0(\rho) - \frac{(1-\lambda)\rho R}{\log_2 e}\right)\right] \log_2 e \\ &= \min_{0<\lambda<1} \left[\max_{\rho>0} \frac{1}{\lambda}\Big(E_0(\rho)\log_2 e - (1-\lambda)\rho R\Big)\right]. \end{aligned} \tag{3.7}$$

in which $E_0(\rho)\log_2 e$ is Gallager's exponent base-2,

$$\begin{aligned} E_0(\rho)\log_2 e &= -\max_q \log \sum_j \left[\sum_i q_i p_{ij}^{\frac{1}{1+\rho}} \right]^{(1+\rho)} \log_2 e \\ &= -\max_q \log_2 \sum_j \left[\sum_i q_i p_{ij}^{\frac{1}{1+\rho}} \right]^{(1+\rho)} \end{aligned} \tag{3.8}$$

The exponent $E_0(\rho)$, illustrated in Figure 3.2, is well-studied (Theorem 5.6.3, [14]) and has desirable properties. It is positive concave and a line of slope C drawn from the origin is tangent to the curve at $\rho = 0$. So unless the channel is trivial (ie. has zero Shannon Capacity C) then $E_0(\rho) > 0$ for all $\rho > 0$. On the other hand, the form of $E_a^+(R)$ in (3.7) resembles the 'conjugate' or the 'dual' of E_0. This suggests that E_a^+ itself might have similar desirable properties. This last observation leads to the main result of this chapter.

Theorem 3.3.2 *The function $E_a^+(R)$ given in (3.7) may be written parametrically as,*

$$\begin{aligned} R &= E_0(\eta)\log_2 e/\eta \\ E_a^+(R) &= E_0(\eta)\log_2 e,\ \eta > 0. \end{aligned} \tag{3.9}$$

and has the following properties,

1. $\lim_{\eta\to 0} E_0(\eta)\log_2 e/\eta = C$ *and* $E_0(0) = 0$,
2. $\lim_{\eta\to\infty} E_0(\eta)/\eta = 0$ *and* $\lim_{\eta\to\infty} E_0(\eta) = E_{sp}(0)$,
3. *It is concave and decreases from* $E_a^+(0) = E_{sp}(0)\log_2 e$ *to zero at* $R = C$.

Remark 3.3.2 *The result is based on the Sphere-packing bound which is known to be the tightest block-coding bound at high rates for symmetric channels with feedback. At low rates, there may be tighter block-coding bounds such as the Straight-Line bound, [19], that could be used to improve the result of Theorem 3.3.2.*

Remark 3.3.3 *It is possible to get rid of the annoying* $\log_2 e$ *factors by redefining* E_0 *in base-2. However, this complicates the subsequent analysis which is more convenient base-e.*

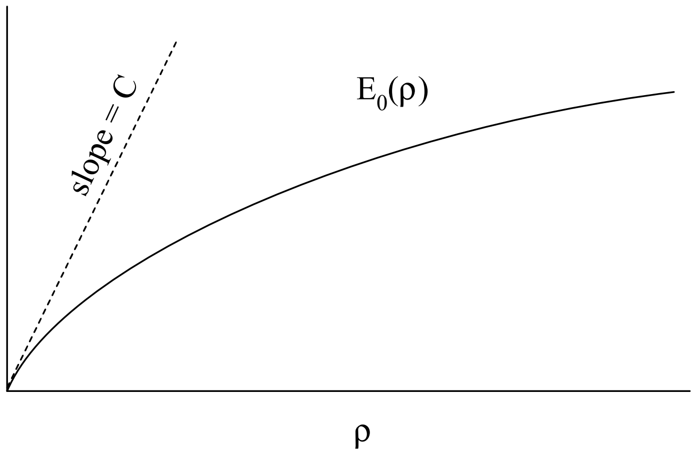

Figure 3.2: Sketch of $E_0(\rho)$.

Proof: The first property follows immediately from the properties of E_0. The third property follows from the first and second properties together with (3.9). So we only need to prove the second property and that E_a^+ can be represented parametrically as in (3.9).

First consider the limit,

$$\lim_{\eta\to\infty} \frac{E_0(\eta)}{\eta}.$$

By (3.8) and Jensen's inequality,

$$\begin{aligned}\lim_{\eta\to\infty} \frac{E_0(\eta)}{\eta} &= -\lim_{\eta\to\infty} \max_q \log \left[\sum_j \left[\sum_i q_i p_{ij}^{\frac{1}{1+\eta}} \right]^{(1+\eta)} \right]^{\frac{1}{\eta}} \\ &\leq -\lim_{\eta\to\infty} \max_q \log \sum_j \left[\sum_i q_i p_{ij}^{\frac{1}{1+\eta}} \right]^{\frac{1+\eta}{\eta}}\end{aligned}$$

For very large values of η, the quantities $p_{ij}^{\frac{1}{1+\eta}}$ can be made arbitrarily close to one and the exponent $\frac{1+\eta}{\eta}$ can also be made arbitrarily close to one. Consequently, for any probability distribution q, the summation on the right hand side can be made arbitrarily close to one. Because $\log 1 = 0$ this means that the right hand side converges to 0. On the other hand, $E_0(\eta) \geq 0$ for all $\eta \geq 0$ so the limit is identically zero.

To complete the proof of the second property we observe the Gallager form for the sphere-packing bound,

$$E_{sp}(R) = \max_{\rho>0} \big[E_0(\rho) - \rho R\big].$$

Taking $R = 0$ and keeping in mind that $E_0(\rho)$ increases in ρ we obtain the desired result.

It remains to show that the upper-bound E_a^+ can be written parametrically as in (3.9). Consider (3.7). Observe that the maximum in the brackets is a function of η and R and write $f\big(\eta, \rho(\eta, R)\big)$

to denote the maximum. The quantity $\rho(\lambda, R)$ is understood to be the value of ρ in (3.9) at which the maximum occurs. But E_0 is concave so that $\rho(\lambda, R)$ may be taken as any solution of,

$$\frac{\partial E_0(\rho)}{\partial \rho} - \frac{(1-\lambda)R}{\log_2 e} = 0. \tag{3.10}$$

A solution exists because E_0 is concave. If the solution is not unique we will pick the smallest solution.

Now use (3.10) in (3.7),

$$\frac{E_a^+(R)}{\log_2 e} = \min_{0<\lambda\leq 1} \frac{1}{\lambda}\left[E_0\big(\rho(\lambda,R)\big) - \rho(\lambda,R)\frac{\partial E_0\big(\rho(\lambda,R)\big)}{\partial \rho}\right]. \tag{3.11}$$

Write $g(\lambda, R)$ to denote the quantity in the brackets. Again, by concavity, the minimizing value of λ can be taken as any solution of,

$$\frac{\partial \frac{1}{\lambda} g(\lambda,R)}{\partial \lambda} = 0.$$

By the ratio rule of differentiation, this solution is any fixpoint of the following equation,

$$\lambda = \frac{g(\lambda,R)}{\frac{\partial g(\lambda,R)}{\partial \lambda}}, \tag{3.12}$$

where, by direct calculation and (3.10),

$$\frac{\partial g(\lambda,R)}{\partial \lambda} = -\rho(\lambda,R)\frac{R}{\log_2 e}.$$

Using this in (3.12) and subsequently in (3.11),

$$R = \frac{E_0\big(\rho(\lambda,R)\big)\log_2 e}{\rho(\lambda,R)},$$
$$E_a^+(R) = E_0\big(\rho(\lambda,R)\big)\log_2 e.$$

The desired result then follows by observing that $\rho(\lambda, R)$ must be strictly positive and renaming it as η. □

The point of this theorem is to demonstrate that the function E_a^+ has some desirable properties. So E_a^+ is a 'nice' function. We still need to see whether it is also a 'nice' bound. In particular, does it represent a reasonably close value to the true achievable anytime exponent? This is a non-trivial question to which we have no definitive answer. Instead, we compare E_a^+ with previously known truly achievable anytime exponents. We proceed to do so for the case of an optimal encoder-decoder for the Binary Erasure Channel with feedback.

3.3.2 The bound is tight for the Binary Erasure Channel with feedback

Consider a Binary Erasure Channel with feedback and a rate R source $\{s_j\}$. The optimal anytime encoder maintains an infinite binary queue. Every $1/R$ time units the next source bit is put at the end of the queue. At the n-th use of the channel, the encoder emits the bit at the head of the queue. If there were no errors, this bit is removed from the queue. If there were errors, the bit is retransmitted at time $n+1$. The function of the decoder is obvious.

The encoder-decoder pair described above is trivially optimal. We use the term optimal to mean that it achieves the largest possible anytime exponent. The exact formula for the achievable exponent is calculated in [22].

Theorem 3.3.3 *Consider a Binary Erasure Channel with feedback and erasure probability* $0 \leq \varepsilon \leq 1$. *The anytime exponent of any encoder-decoder is bounded by,*

$$E_a\left(1 - \frac{1}{\eta}\log_2\left[1+\varepsilon(2^\eta - 1)\right]\right) \tag{3.13}$$

$$= \eta - \log_2\left[1+\varepsilon(2^\eta-1)\right],\ \eta > 0. \tag{3.14}$$

The relationship holds with equality for the encoder-decoder described above.

Now consider the upper-bound calculated in Theorem 3.3.2. For the Binary Erasure Channel the channel transition probabilities are,

$$[p_{ij}] = \begin{bmatrix} 1-\varepsilon & \varepsilon & 0 \\ 0 & \varepsilon & 1-\varepsilon \end{bmatrix}$$

Using these figures in (3.8),

$$\begin{aligned}
E_0(\eta)\log_2 e &= -\max_q \log_2 \sum_j \left[\sum_i q_i p_{ij}^{\frac{1}{1+\eta}}\right]^{(1+\eta)} \\
&= -\max_q \log_2 \Bigg[\left(q_0(1-\varepsilon)^{\frac{1}{1+\eta}}\right)^{(1+\eta)} \\
&\qquad + \left(q_1(1-\varepsilon)^{\frac{1}{1+\eta}}\right)^{(1+\eta)} \\
&\qquad + \left(q_0\varepsilon^{\frac{1}{1+\eta}} + q_1\varepsilon^{\frac{1}{1+\eta}}\right)^{(1+\eta)}\Bigg] \\
&\overset{q=(.5,.5)}{=} -\log_2\left[(1-\varepsilon)\frac{1}{2^\eta} + \varepsilon\right] \\
&= \eta - \log_2\left[1+\varepsilon(2^\eta-1)\right]
\end{aligned}$$

Using this in (3.9) gives the following parametric representation for E_a^+,

$$\begin{aligned}
R &= 1 - \frac{1}{\eta}\log_2\left[1+\varepsilon(2^\eta-1)\right] \\
E_a^+(R) &= \eta - \log_2\left[1+\varepsilon(2^\eta-1)\right]
\end{aligned} \tag{3.15}$$

This expression coincides with the anytime exponent of Theorem 3.3.3. Because this exponent is achievable we conclude that the upper bound E_a^+ is tight for all $\eta > 0$ and consequently for all $0 \leq R \leq C$ for the Binary Erasure Channel with feedback.

For illustrative purposes (Figure 3.3) we plot $E_a^+(R)$ versus R for a Binary Erasure Channel with feedback and erasure probability $\varepsilon = 1/4$. In the figure we have also plotted the curve $(1/\lambda)E_{sp}\big((1-\lambda)R/\log_2 e\big)\log_2 e$ versus R for several distinct values of the parameter λ. For each value of λ the associated curve tangentially touches E_a^+ at the value predicted by (3.15).

3.4 Anytime coding and estimation of unstable systems

Anytime codes are interesting because they can be used to estimate the state of an unstable scalar system over a noisy communication channel with or without feedback. This result was proven

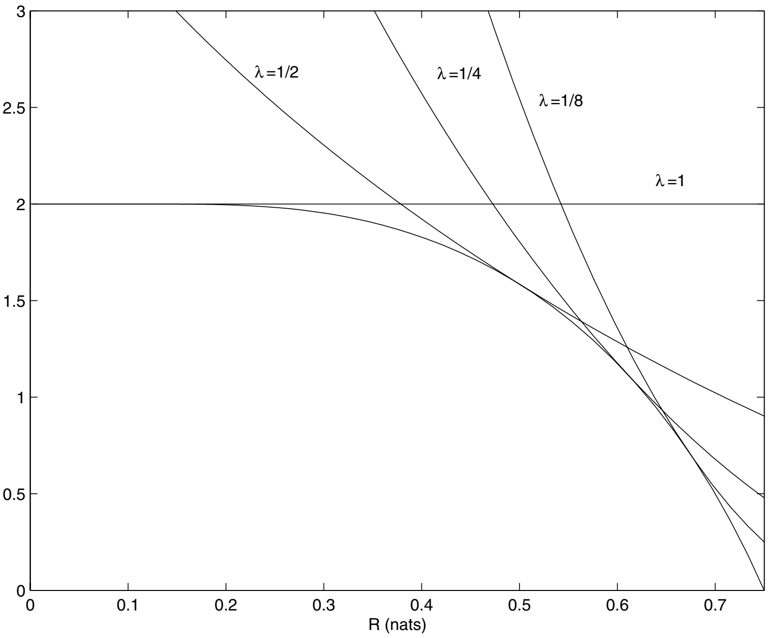

Figure 3.3: Plot of $E_a^+(R)$ vs. R for a Binary Erasure Channel with erasure probability $\varepsilon = 1/4$.

by Sahai [22]. The converse statement is also true. If we can reliably estimate the state of an unstable system over a noisy channel then we can construct an anytime code for that channel. Although less obvious the converse statement is more interesting because it provides a method to design anytime codes. Constructing an anytime code by designing an estimator is analogous to solving an integer program by solving its linear program relaxation. The idea is to avoid typical problems in integer counting and case enumeration.

In this section we prove Sahai's theorem that shows how to construct an anytime code from a stable estimator. In subsequent chapters we will use this result to design anytime codes. The proof is repeated here to avoid excessive cross referencing.

Theorem 3.4.1 (Sahai) *Consider the unstable system,*

$$x_n = \sigma x_{n-1} + w_n,\ n \geq 1,\ x_0 = 0,\ \sigma > 1$$

where w_n is a bounded but otherwise unknown disturbance. Suppose there exists an encoder that observes x_n at time n and communicates its observations over a noisy channel to a decoder in such a way that the decoder produces an estimate y_n of x_n for which,

$$\forall n \geq 1, \sup_n \mathbb{E}|x_n - y_n|^\eta < \infty.$$

Under these conditions we can construct a rate $R = \log_2 \sigma$ anytime encoder-decoder pair for which,

$$E_a(R) \geq \eta R.$$

In other words,

$$C^{\mathcal{A}}(\eta \log_2 \sigma) \geq \log_2 \sigma.$$

Remark 3.4.1 *The result holds for any noisy channel with or without feedback or memory.*

Proof: Let $\delta > 0$. We construct a rate $R = \log_{2+\delta} \sigma$ anytime code with the necessary anytime exponent. A pre-encoder simulates a system,

$$x_n = \sigma x_{n-1} + w_n,\ n \geq 1,\ x_0 = 0,\ \sigma > 1$$

and feeds the state x_n into the encoder of the theorem. At time $n\tau$, the pre-encoder examines the source bits,

$$s_{\lfloor (n-1)R \rfloor + 1}, \cdots, s_{\lfloor nR \rfloor}$$

and selects,

$$w_n = \sigma^n \sum_{m=\lfloor (n-1)R \rfloor + 1}^{\lfloor nR \rfloor} (2+\delta)^{-m} s_m$$

if $\lfloor nR \rfloor \geq \lfloor (n-1)R \rfloor + 1$ or selects $w_n = 0$ otherwise. Observe that w_n is bounded for all $n \geq 1$. To see this, note that $\lfloor a \rfloor - \lfloor b \rfloor \leq \lfloor a - b \rfloor \leq \lfloor a \rfloor - \lfloor b \rfloor + 1$ and simplify,

$$\begin{aligned}
w_n &= \sigma^n \sum_{m=\lfloor (n-1)R \rfloor + 1}^{\lfloor nR \rfloor} (2+\delta)^{-m}\, s_m \\
&= \Big(\lfloor nR \rfloor - \lfloor (n-1)R \rfloor - 1 \Big)\, \sigma^{nR}\, (2+\delta)^{-\lfloor (n-1)R \rfloor - 1} \\
&\leq \big(\lfloor R \rfloor + 1 \big)\ \sigma^{nR}\, (2+\delta)^{-\lfloor (n-1)R \rfloor - 1} \\
&\leq \big(\lfloor R \rfloor + 1 \big)\ (2+\delta)^{nR}\, (2+\delta)^{-\lfloor (n-1)R \rfloor - 1} \\
&\leq \big(\lfloor R \rfloor + 1 \big)\, (2+\delta)^{-\lfloor R \rfloor - 1}\, (2+\delta)^{nR - \lfloor nR \rfloor} \\
&\leq \big(\lfloor R \rfloor + 1 \big)\, (2+\delta)^{-\lfloor R \rfloor}.
\end{aligned}$$

So w_n is a bounded but otherwise arbitrary sequence as required by the assumptions of the theorem. The reason for selecting w_n in this awkward manner is obvious once we write out the state evolution,

$$x_n = \sigma^n \Big((2+\delta)^{-1} s_1 + (2+\delta)^{-2} s_2 + \cdots + (2+\delta)^{-\lfloor nR \rfloor} s_{\lfloor nR \rfloor} \Big).$$

The post-decoder guesses $\hat{s}_1(n), \hat{s}_2(n), \cdots, \hat{s}_{\lfloor nR \rfloor}(n)$ such that,

$$y_n = \sigma^n \sum_{m=-1}^{\infty} (2+\delta)^{-m} \hat{s}_m(n).$$

So if the first j bits $\hat{s}_1(n), \cdots, \hat{s}_j(n)$ were guessed incorrectly then necessarily,

$$\begin{aligned}
|y_n - x_n| &\geq \sigma^{n-j/R} \min_{\hat{s}_{j+1}^{\lfloor nR \rfloor}} \left[\sum_{m=j+1}^{\lfloor nR \rfloor} (2+\delta)^{-m} s_m - \sum_{m=j+1}^{\lfloor nR \rfloor} (2+\delta)^{-m} \hat{s}_m(n) \right] \\
&\geq \sigma^{n-j/R} 2 \sum_{m=j+1}^{\infty} (2+\delta)^{-m} \\
&= \sigma^{n-j/R} \frac{2\delta}{1+\delta}.
\end{aligned}$$

Consequently, by Markov's inequality,

$$\begin{aligned}\mathbb{P}\big(\hat{s}_j(n) \neq s_j\big) &\leq \mathbb{P}\big(\hat{s}_1^j(n) \neq s_1^j\big) \\ &\leq \mathbb{P}\big(|y_n - x_n| > \sigma^{n-j/R} c_1\big) \\ &\leq \frac{\mathbb{E}|y_n - x_n|^\eta}{c_1 \sigma^{\eta(n-j/R)}} \\ &\leq c_2 \sigma^{-\eta(n-j/R)}\end{aligned}$$

which proves the theorem. □

The main purpose of this theorem was to demonstrate a design method for anytime codes. We first design a stable estimator for an unstable system with parameter σ. The proof of the theorem then gives a construction which leads to a rate $R = \log_2 \sigma$ anytime encoder-decoder. It is interesting, but perhaps irrelevant, that the anytime exponent of the constructed encoder-decoder is similar in form to the parametric representation of the bound E_a^+ of Theorem 3.3.2.

To conclude we state without proof the converse of Theorem 3.4.1.

Theorem 3.4.2 *Consider a discrete-time channel whose capacity is* C *and for which* E_a^* *is an achievable anytime exponent. Consider also the system,*

$$x_n = \sigma x_{n-1} + w_n,\ n \geq 1,\ x_0 = 0,\ \sigma > 1.$$

If $\log_2 \sigma < C$ *then there exists an encoder-decoder pair such that the encoder observes* x_n *at time* n *and communicates its observations to the decoder in such a way that the decoder produces an estimate* y_n *of* x_n *for which,*

$$\sup_{n \geq 0} \mathbb{E}|x_n - y_n|^\eta < \infty$$

for any $\eta < E_a^*(\log_2 \sigma)/\log_2 \sigma$.

3.5 Discussion: Tree codes

Consider a discrete memoryless channel and a rate $R = \frac{p}{q}$ anytime encoder. Because the rate is purely rational and because an anytime encoder is also a 'sequential' encoder, we can describe its operation by means of a coding tree. The coding tree has 2^p branches, each containing q symbols from the input alphabet of the channel as illustrated in Figure 3.5. The source bits $s_1, s_2, s_3, \cdots$ determine a path along the tree. The encoder moves along this path (one node each q time steps) and emits the channel symbols it sees along the way. At time n, based on the n channel symbols it has received, the decoder produces an estimate $\hat{s}(n) = (\hat{s}_0, \hat{s}_1, \cdots, \hat{s}_{j_n})$ where,

$$j_n = \left\lceil n\frac{p}{q} \right\rceil,\ n = 1, 2, \cdots$$

It there is a feedback channel from the encoder to the decoder, the branch labels depend on the history of symbols transmitted by the encoder and the history of symbols received by the decoder.

Suppose that we drop the anytime property and that we are interested in an encoder-decoder pair that produces the 'best' estimate $\hat{s}(N)$ for a given time N. Suppose also that we drop the feedback assumption. Then the tree code reduces to a convolutional code for which Viterbi proved the following result in his seminal paper [32],

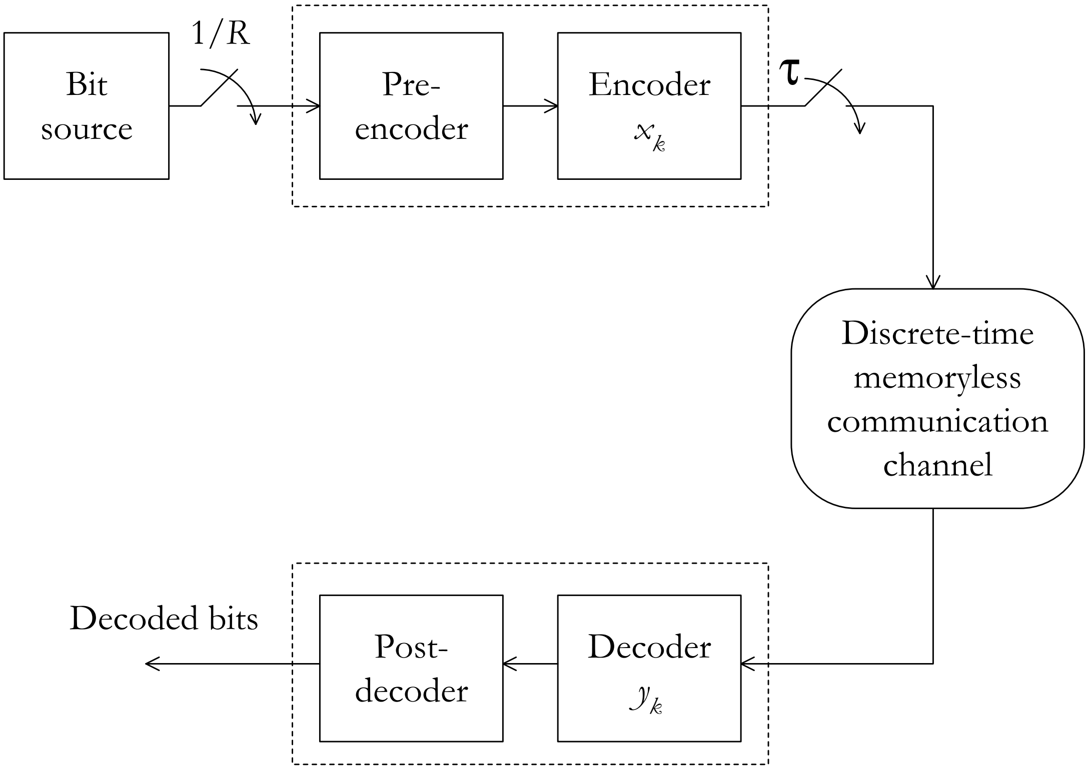

Figure 3.4: A block diagram illustrating the construction of an anytime code from a state estimator.

Theorem 3.5.1 (Viterbi, [32]) *Consider a rate $R = p/q$ tree code with total transmission time N. Let the source sequence be 'zero padded' such that the last d $(0 \leq d \leq N)$ bits $s_{\lceil NR \rceil}^{\lceil NR \rceil - d+1} = 0$. Then the probability of decoding error is bounded by,*

$$\mathbb{P}\left(\hat{s}_1^{\lceil NR \rceil - d} \neq s_1^{\lceil nR \rceil - d} \right) > 2^{-d[E_L(R)+o(d)]}$$

where $E_L(R)$ is determined by solving for $\eta > 0$ in the following equations,

$$\begin{aligned} R &= E_0(\eta) \log_2 e/\eta \\ E_L(R) &= E_0(\eta) \log_2 e. \end{aligned} \tag{3.16}$$

Furthermore, the bound is tight for $R > E_0(1)$.

There are three important aspects of Viterbi's bound. First, he is not assuming any feedback information. Second, the bound is tight for a region of rates close to capacity. Lastly, the bound (3.16) is identical to the parametric bound in (3.9). Putting all of these together we obtain the following surprising result,

> "Feedback cannot be used to improve the reliability of a tree code for symmetric channels and for the region of rates $R > E_0(1)$."

To put this into perspective recall the analogous result for block-codes. (Theorem 2.1.1) Without feedback, the family of tree-codes of length N is a strict subset of the family of length N block-codes. But with feedback, the two families are one in the same. Consequently, the result for tree-codes is stronger than its analogue for block-codes.

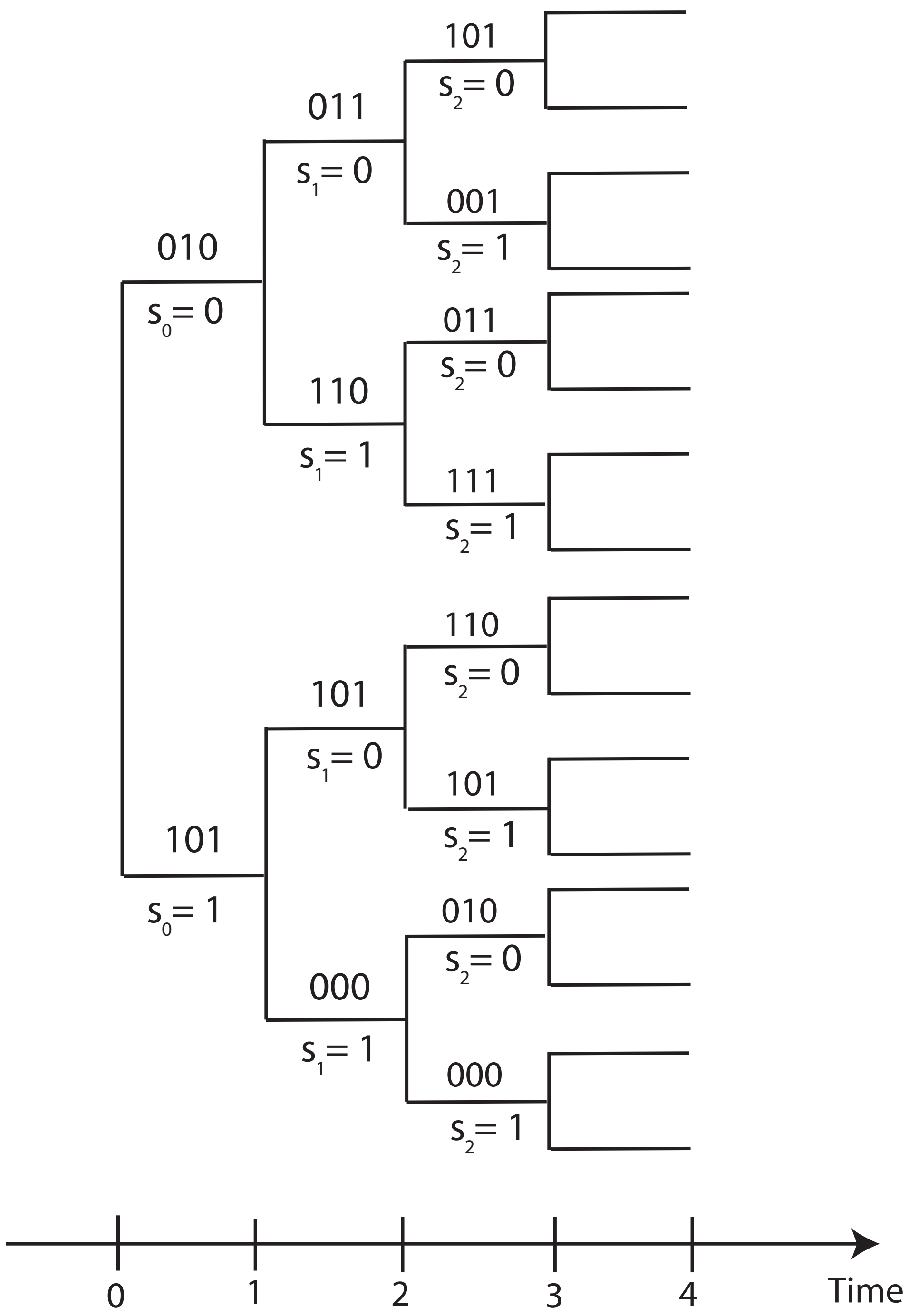

Figure 3.5: A rate $R = 1/3$ Tree-code with two branches, each labeled with $q = 3$ channel symbols. The source bits s_n, $n = 0, 1, 2, \cdots$ are indicated below branches and the channel symbols are indicated above the branches.

Chapter 4

A time-sharing anytime code

We design an anytime code with feedback. The design has a simple time-sharing property. The time horizon is broken into segments of length N. Each segment is then divided into a 'information' and 'control' part. The information part is used to transmit new bits of data from the encoder to the decoder. For an anytime code no errors can go uncorrected. When an error occurs the control part is used to inform the decoder about the situation. The basic idea may appear to be a special case of the time-sharing argument for variable-length block codes described previously in Section 2.3. However, there is a crucial difference. Because the decoder reinterprets every bit upon receiving new data it is possible that bits that were correctly decoded in the past can be corrupted by errors that occur in the present. These kinds of back-ups are a very real and difficult bottleneck for Anytime codes.

The code has an interesting 'renewal' property that helps us deal with the back-up problem. There are time instants when we say the code is renewed. At a renewal time the encoder-decoder dynamics are reset. Between renewals times, new bits of information are transmitted, if there were no errors, or no new information is transmitted, if there were errors that were then recaptured by a sequence of distinguishing symbols.

To avoid problems of integer counting and enumeration we employ Theorem 3.4.1. So rather than designing an anytime code we design a stable estimator for an unstable scalar system. The reader who has skipped forward may wish to look at Section 3.4 before proceeding.

Our time-sharing codes achieve an Anytime exponent $E_a(R)$ close to the upper bound E_a^+ derived in the previous chapter. As a consequence, the achievable exponent beats the sphere-packing bound. Therefore, observing that the Anytime exponent E_a can be used to lower bound the performance of a convolutional or tree code targeted at a fixed decoding delay, we have disproved Pinsker's result [21] which argues that the sphere-packing bound is a hard bound on fixed delay decoding with feedback. In particular, this disproves the general belief that feedback is useless for the Binary Symmetric Channel: the sphere-packing bound is misleading when it comes to the opportunities provided by feedback.

4.1 The estimation problem

We consider a one-dimensional, unstable discrete-time system,

$$x_{n+1} = \sigma\, x_n + w_n,\ \sigma > 1,\ x_0 = 0,\ n \geq 0, \tag{4.1}$$

where $w_n \in [-W, W]$ is a bounded but otherwise unknown disturbance. An encoder observes x_n. Based on its observations up to time n it emits a channel symbol a_n. At the same time a decoder receives b_n.

Based on what it has received up to time n the decoder produces an estimate $\hat{x}_n$ of x_n. Formally, the encoder is described by functions ϕ_n,

$$z_n = \phi_n\big(x^n, b^{n-1}\big),\ n \geq 0,$$

the channel is described by the probability distribution (3.2) and the decoder is described by functions ψ_n,

$$\hat{x}_n = \psi_n(b^n),\ n \geq 0.$$

By Theorem 3.4.1 the pair (ϕ, ψ) is a rate $R = \log_2 \sigma$ anytime encoder-decoder with anytime exponent $\eta \log_2 \sigma$ provided that the error in the estimate is η-stable,

$$\sup_{n \geq 0} \mathbb{E}|x_n - \hat{x}_n|^\eta < \infty. \tag{4.2}$$

The design objective is:

> Fix the system coefficient $\sigma > 1$ and design an encoder-decoder pair (ϕ, ψ) such that the error in the estimate is η-stable for the largest possible η.

4.2 The best estimator for the perfect channel

Let's begin with a special case. Suppose the channel is the perfect binary channel. So the channel input a_n is a binary signal and the channel output b_n is equal to the channel input,

$$b_n = a_n,\ n \geq 0.$$

The following theorem is originally due to Tatikonda.

Theorem 4.2.1 *If $\sigma < 2$ there is a bounded-error encoder-decoder pair so that,*

$$\sup_{k \geq 0} |x_k - \hat{x}_k| < \infty.$$

If $\sigma > 2$ no such pair exists.

Remark 4.2.1 *A generalization of this theorem also holds for higher dimensional linear systems for which x_k is a vector and σ is a matrix. See Tatikonda [29] for details.*

Proof: We construct the required pair. At each n, the decoder determines an uncertainty interval $[l_n, u_n]$ that is guaranteed to contain x_n and selects the estimate to be the mid-point of this interval, $\hat{x}_n = \frac{1}{2}(l_n + u_n)$, starting with $[l_0, u_0] = [-1, 1]$. Observe that at each time n, the encoder can also calculate $\hat{x}_n$. The encoder selects a_{n+1} as $a_{n+1} = 1$ if $x_{n+1} \geq \sigma \hat{x}_n$ and $a_{n+1} = 0$, otherwise. Upon receiving a_{n+1}, the decoder selects $[l_{n+1}, u_{n+1}]$ by

$$[l_{n+1}, u_{n+1}] = \begin{cases} [\sigma \hat{x}_n, \sigma u_n + W], & \text{if } z_{n+1} = 1, \\ [\sigma l_n - W, \sigma \hat{x}_n] & \text{if } z_{n+1} = 0 \end{cases} \tag{4.3}$$

Since $x_{n+1} = \sigma x_n + w_n \in \sigma[l_n, u_n] + [-W, W]$ it follows that $x_{n+1} \in [l_{n+1}, u_{n+1}]$. The size of the uncertainty interval evolves as

$$u_{n+1} - l_{n+1} = \frac{\sigma}{2}(u_n - l_n) + W,$$

which remains bounded if $\sigma < 2$.

On the other hand, at time n, $x_n = \sum_{j=0}^{n-1} \sigma^{n-1-j} w_j$ can be any point in the following interval,

$$X_n = \frac{\sigma^n - 1}{\sigma - 1}[-W, W].$$

By time n the decoder has observed 2^n possible values of a^n, based on which it could have made 2^n possible point estimates $\psi_n(a^n)$. Hence the worst estimation error is at least

$$\begin{aligned} \max_{x \in X_n} \min_{\psi_n} |x - \psi_n(a^n)| &\geq |X_n| \\ &= 2W \frac{\sigma^n - 1}{\sigma - 1} \frac{1}{2^n}, \end{aligned}$$

which is unbounded if $\sigma > 2$. □

The construction is very simple but demonstrates a fundamental technique. The idea is to maintain an interval at the decoder and to 'cut' the interval at the fastest possible rate. We proceed now to the general case. We attempt to carry over the basic technique developed here. Because of channel noise, however, we can no longer guarantee the interval to contain the true state.

Observe that for the perfect channel,

$$\sup_{n \geq 0} \mathbb{E}|x_n - \hat{x}_n|^\eta < \infty$$

for any $\eta > 0$. By invoking Theorem 3.4.1 we see that the anytime capacity of the perfect binary channel is given by,

$$\mathrm{C}^{\mathcal{A}}(\alpha) = 1, \ \alpha > 0.$$

Conversely, the anytime exponent for the perfect channel is infinite for any rate up to one,

$$E_a(R) = \infty, \ 0 \leq R \leq 1.$$

4.3 A decent estimator for a non-perfect channel

The decoder maintains an interval $I_n = [l_n, u_n]$. Contrary to our previous construction we cannot guarantee I_n to contain the true state at all times. For this reason we will call this the 'interval of confidence'. The decoder believes that the true state is in the interval so it produces the estimate $\hat{x}_n$ of the true state x_n as the mid-point of the interval,

$$\hat{x}_n = \frac{l_n + u_n}{2}. \tag{4.4}$$

At time $n+1$, the encoder also computes the interval I_n. Once again, the encoder can do this because it has perfect knowledge of the decoder.

4.3.1 The decoder

At time n, the decoder adaptively listens for a variable-length codeword. Let N, L and M be parameters to be determined later. The decoder first receives L channel symbols b_n^{n+L-1}. These symbols are used according to the maximum *a posterior* probabilities to distinguish between a binary hypothesis: 'back-up' or 'continue'.

Say that the decoder makes a decision in favor of the hypothesis 'continue'. The decoder then receives $N-L$ channel symbols b_{n+L}^{n+N-1} that are used to select one of M possible messages according to the maximum a posterior probabilities. The decoder concludes that the true state was in the m-th cut $[p_m, p_{m+1}]$ of an M partition of $I_{n-1} = [l_{n-1}, u_{n-1}]$,

$$l_{n-1} = p_1 \leq p_2 \leq \cdots \leq p_M = u_{n-1}.$$

Consequently, at time $n+N-1$ the decoder updates the interval of confidence as,

$$I_{n+N-1} = \sigma^N [p_m,\, p_{m+1}] + [-\Gamma_N^\sigma(W),\, \Gamma_N^\sigma(W)], \tag{4.5}$$

where

$$\Gamma_N^\sigma(W) \equiv \sum_{i=0}^{N-1} \sigma^i W. \tag{4.6}$$

The decoder saves the index m in a stack S for future use and at time $n+N$ listens for a new variable-length codeword.

For the first $N-1$ time steps, while the decoder is receiving the codeword, the interval is updated by,

$$\begin{aligned} I_n &= \sigma[l_{n-1},\, u_{n-1}] + [-W, W], \\ I_{n+1} &= \sigma[l_n,\, u_n] + [-W, W], \\ &\vdots \\ I_{n+N-2} &= \sigma[l_{n+N-3},\, u_{n+N-3}] + [-W, W]. \end{aligned}$$

Suppose now that the decoder had decided in favor of the 'back-up' hypothesis at time $n+L-1$. The decoder concludes that the true state x_{n-1} was not in the interval I_{n-1} so the decoder pops a partition index m from the top of the stack S and enlarges the interval as,

$$\begin{aligned} I_{n+L-1} = \sigma^L \left[\, l_{n-1} - m|I_{n-1}|,\, u_{n-1} + \big(M - (m+1)\big)|I_{n-1}|\,\right] \\ + \big[- \Gamma_L^\sigma(W),\, \Gamma_L^\sigma(W) \big], \end{aligned} \tag{4.7}$$

where Γ is the function defined in (4.6). In words, the interval is extended to the left by $m|I_{n-1}|$ and to the right by $(2^N - (m+1))$ and the resulting interval is propagated by the system equation (4.1). At time $n+L$ the decoder waits for a new variable-length codeword.

Observe that independent of the value of the partition index m the interval is updated such that,

$$\sigma^{(N+L)} I_{n-1} + \Big[- \Gamma_{N+L}^\sigma(W),\, \Gamma_{N+L}^\sigma(W) \Big] \tag{4.8}$$

$$\subseteq I_{N+L-1}, \tag{4.9}$$

This concludes the description of the decoder.

4.3.2 The encoder

We prescribe the probability of transmission errors. P_{e,h^0} denotes the probability of decision error for the 'back-up' hypothesis by L uses of the channel and P_{e,h^1} denotes the probability of decision error for the 'continue' hypothesis by L uses of the channel. We also prescribe the M probabilities $P_{e,m}$, $m = 1, 2, \cdots, M$, which denote the probability of error in transmitting a M-ary message by N uses of the channel. Write $P_e = \max_m P_{e,m}$.

Based on our previous description of the decoder and on the prescribed probabilities, the encoder can exhibit exactly three distinct behaviors as illustrated in Figure 4.1:

i) At time $n-1$ if $x_{n-1} \in I_{n-1}$ and the decoder is waiting for a new codeword we select $\phi_n, \cdots, \phi_{n+L-1}$ to be the encoder that transmits the 'continue' hypothesis. So with probability $(1-P_{e,h^1})$ the decoder decides in favor of the 'continue' hypothesis in which case we select $\phi_{n+L}, \cdots, \phi_{n+N-1}$ to be the encoder that transmits the appropriate cut of the partition of I_{n-1} that contains the true state x_{n-1}. With probability greater than $(1-P_e)$ the decoder receives the true cut and updates the interval according to (4.5) such that $x_{n+N-1} \in I_{n+N-1}$. This is the good case.

ii) With probability P_{e,h^1} the decoder decides in favor of the false back-up hypothesis. So at time $n+L-1$ the decoder mistakenly enlarges the interval according to (4.7). The situation is not so grim because at time $n+L-1$ we still have $x_{n+L-1} \in I_{n+L-1}$.

iii) Finally, the worst case occurs when the decoder decides in favor of the true continue hypothesis but makes a mistake in deciding the true cut. At time $n+N-1$ the decoder mistakenly updates the interval according to (4.5) but the true state x_{n+N-1} is no longer contained in the updated interval I_{n+N-1}. Starting at the troublesome time $n+N$ the encoder is designed in such a way as to bias the decoder to always favor the 'back-up' decision. The encoder goes back to normal operation at the smallest time $n' > n+N$ such that the decoder has decided in an equal number of 'back-up' and 'continue' decisions in the time interval $\{n, n+1, \cdots, n'\}$.

Observe that the smallest such time can be represented as $n' = n + K(N+L)$ for some integer $K > 0$. This is because each continue decision leads to a run of N time steps and each back-up decision leads to a run of L time steps. Furthermore, for any such $K > 0$ it follows by (4.8) that,

$$\sigma^{K(N+L)} I_{n-1} + [-\Gamma^\sigma_{K(N+L)}(W), \Gamma^\sigma_{K(N+L)}(W)] \subseteq I_{n+K(N+L)}. \tag{4.10}$$

This concludes the design of the encoder-decoder pair. What we really need now is a useful mathematical model that describes the dynamics of the interval.

Observe that in each of the three possible cases there is a common property. The true state is contained in the interval $x_{n-1} \in I_{n-1}$ and the decoder awaits a new codeword at the start and end times. This suggests a natural mathematical abstraction which we pursue next.

4.3.3 Mathematical model of encoder-decoder

Say that the pair is 'renewed' each time the interval contains the true state and the decoder awaits a new codeword. At each renewal time, the encoder-decoder dynamics are reset. In between renewal times, the interval either shrinks, if there were no errors, or the interval grows, if there was an error that was later recaptured by a sequence of back-ups. In the following paragraphs we make these statements precise.

Let's begin with some notation that we've found to be useful. Consider a stochastic process $\Upsilon = (\Upsilon_1, \cdots, \Upsilon_T)$ with random termination time T. Let $\Upsilon^1, \Upsilon^2, \cdots$ be independent copies of Υ. Write,

$$\Upsilon^i = \left(\Upsilon^i_1, \Upsilon^i_2 \cdots, \Upsilon^i_{T^i}\right), \; i \geq 1.$$

Construct a new 'renewal' process by concatenating these independent copies of Υ,

$$(\Upsilon^1_1, \Upsilon^1_2, \ldots, \Upsilon^1_{T^1}, \Upsilon^2_1, \Upsilon^2_2, \ldots, \Upsilon^2_{T^2}, \ldots) \tag{4.11}$$

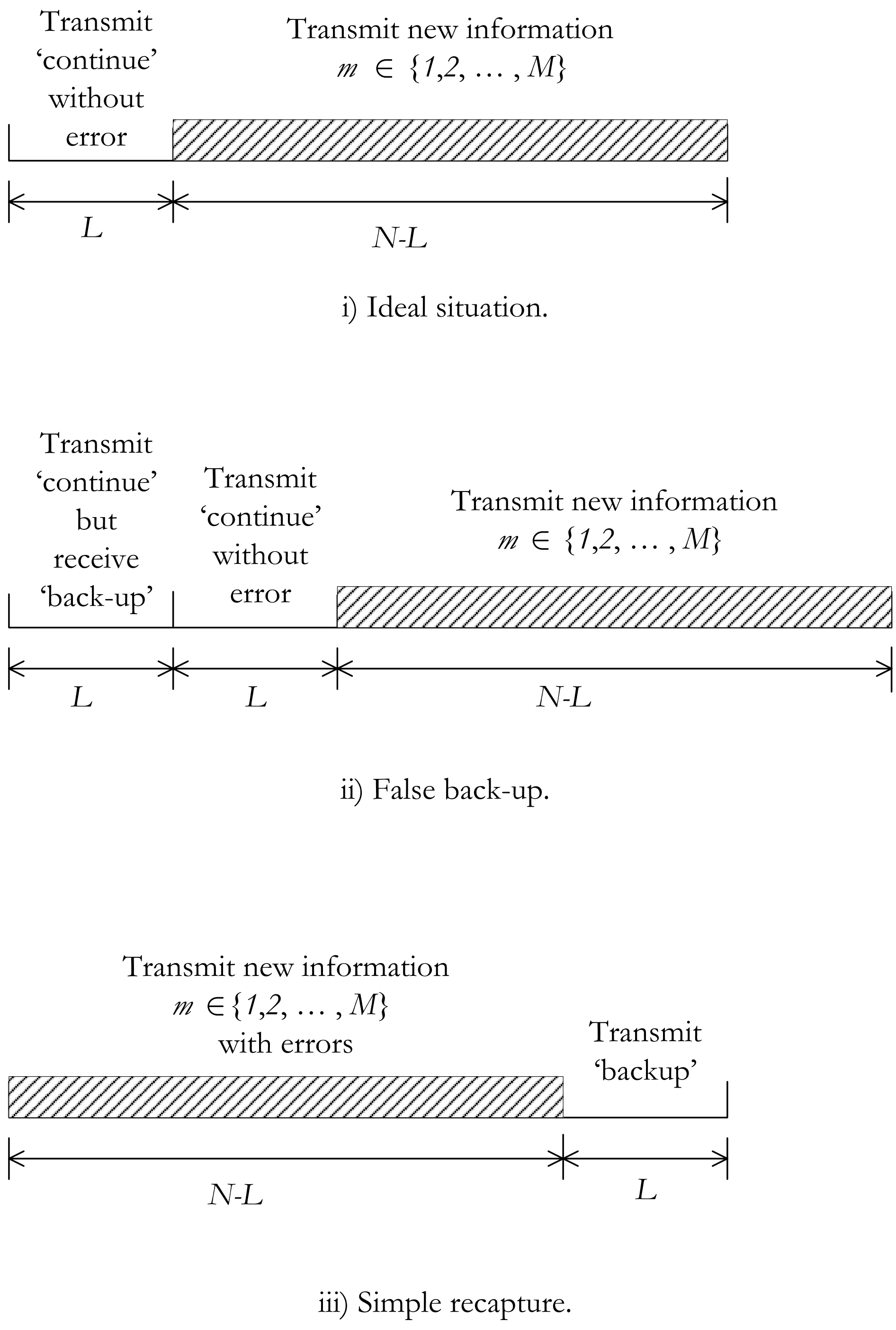

Figure 4.1: Illustration of three possible cases in the transmission of the control and information parts.

Write $\Upsilon(n)$ to denote the n-th entry in this process. Let i_n denote,

$$i_n = \min\{i \mid T^1 + \cdots + T^i \geq n\}. \tag{4.12}$$

Now consider the worst-case estimation error,

$$X_n = \sup_{w^n} |x_n - \hat{x}_n|, X_0 = 0, \; n \geq 0.$$

Using the notation developed above we have,

Lemma 4.3.1 *Consider the prescribed probabilities P_{e,h^0}, P_{e,h^1} and P_e. There exists a stochastic process Υ with random termination time T such that the worst-case estimation error is bounded by,*

$$X_{n+1} \leq \Upsilon(n) X_n + 2W, \; X_0 = 0, \; n \geq 0. \tag{4.13}$$

The distribution of Υ is given by,

$$\mathbb{P}\left(T = L, \prod_{i=1}^{T} \Upsilon_i = M\sigma^L\right) \;=\; P_{e,h^1} \tag{4.14}$$

$$\mathbb{P}\left(T = N, \prod_{i=1}^{T} \Upsilon_i = M^{-1}\sigma^N\right) \;\leq\; (1 - P_{e,h^1})(1 - P_e) \tag{4.15}$$

and for $K \geq 1$,

$$\begin{aligned} \mathbb{P}\left(T = K(N+L), \prod_{i=1}^{T} \Upsilon_i = \sigma^{K(N+L)}\right) & \\ \leq (1 - P_{e,h^1}) P_e & \\ \times \frac{1}{K}\binom{2(K-1)}{K-1} \left(P_{e,h^0}\right)^{K-1} \left(1 - P_{e,h^0}\right)^{K}. & \end{aligned} \tag{4.16}$$

Proof: The existence of such a renewal process Υ is obvious from the description of the encoder. What we need to verify is the specified distribution. The first two equations (4.14) and (4.15) are fairly clear. These correspond to the cases of false backup and the 'good' successful transmission. To verify the last inequality (4.16) we note a result known to combinatorial analysts. Consider the number of $\{-1, 1\}$-sequences $\{r_1, \cdots, r_{2K}\}$ such that $\sum_{i=1}^{j} r_i > 0$ for $j < 2K$ and $\sum_{i=1}^{2K} r_k = 0$. The number of such sequences is given by the Catalan number , [34],

$$\frac{1}{K}\binom{2(K-1)}{K-1}, K \geq 1.$$

Now consider the case scenario to which (4.16) corresponds. At time $n-1$, a renewal occurs. So the true state $x_{n-1} \in I_{n-1}$ and the decoder waits for a new codeword. The encoder delivers a cut but the cut is received in error. The encoder then keeps transmitting 'back-ups' until the smallest time at which the decoder has decided an equal number K of back-ups and false cuts. Associate the 'back-up' hypothesis with a -1 and the 'false cut' hypothesis with a $+1$. So add a point for each false cut and deduct a point for each back-up. This can occur in $\frac{1}{K}\binom{2(K-1)}{K-1}$ many ways and the probability of each occurrence is identical and given by $\big(1-P_{e,h^1}\big)P_e \times \big(P_{e,h^0}\big)^{K-1}\big(1-P_{e,h^0}\big)^K$. This proves the desired result. □

This lemma is useful for finding sufficient conditions for which the estimation error remains bounded. By the definition of X_n and recursive substitution in (4.13) we bound the estimation error as,

$$
\begin{aligned}
\mathbb{E}|x_n-\hat{x}_n|^\eta &\le \mathbb{E}X_n^\eta \\
&\le \mathbb{E}\Big(\sum_{n'=1}^{n-1} 2W \prod_{n''=n'}^{n-1} \Upsilon(n'')\Big)^\eta \qquad (4.17)\\
&= (2W)^\eta \mathbb{E}\Big(\sum_{n'=1}^{n-1}\prod_{n''=n'}^{n-1} \Upsilon(n'')\Big)^\eta \qquad (4.18)\\
& \qquad (4.19)
\end{aligned}
$$

However, for this bound to be useful we need to be able to calculate the expectation of X_n. For this we will make use of the following quantity,

$$
\gamma(\eta,\delta_1) \equiv \mathbb{E}\prod_{i=1}^{T}\big(\Upsilon_i(1+\delta_1)\big)^\eta \qquad (4.20)
$$

which denotes the average gain of the η-th moment of the estimation error in each renewal cycle of the system. Using the fact that Renewal cycles are independent we write out $\gamma(\eta,\delta_1)$ as,

$$
\begin{aligned}
\gamma(\eta,\delta_1) = &\Big(M\big(\sigma(1+\delta_1)\big)^L\Big)^\eta P_{e,h^1} \\
&+\Big(M^{-1}\big(\sigma(1+\delta_1)\big)^N\Big)^\eta (1-P_{e,h^1})(1-P_e) \\
&+(1-P_{e,h^1})P_e \qquad (4.21)\\
&\quad\times \sum_{K=1}^{\infty}\frac{1}{K}\binom{2(K-1)}{K-1}\big(\sigma(1+\delta_1)\big)^{K(N+L)\eta}\big(P_{e,h^0}\big)^{K-1}\big(1-P_{e,h^0}\big)^K.
\end{aligned}
$$

Fortunately, we know the generating function [34] for the series. For any $x \le 1/4$, $\sum_{K=1}^{\infty}(1/K)\binom{2(K-1)}{K-1}x^K = (1-\sqrt{1-4x})/2$. Using this we can simplify the expression in (4.21) into something more manageable,

$$
\begin{aligned}
\gamma(\eta,\delta_1) = &\Big(M\big(\sigma(1+\delta_1)\big)^L\Big)^\eta P_{e,h^1} \\
&+\Big(M^{-1}\big(\sigma(1+\delta_1)\big)^N\Big)^\eta (1-P_{e,h^1})(1-P_e) \\
&+(1-P_{e,h^1})\frac{P_e}{2P_{e,h^0}} \\
&\quad\times\Big[1-\sqrt{1-4\big(\sigma(1+\delta_1)\big)^{(N+L)\eta}P_{e,h^0}\big(1-P_{e,h^0}\big)}\Big]
\end{aligned}
$$

Coming back to the bound that we were after, leads to the next result.

Theorem 4.3.1 *Consider an arbitrary discrete-time memoryless channel. Design a time-sharing encoder-decoder by selecting L, N and M and two block coding schemes. The first scheme uses the channel L times to distinguish between a binary hypothesis and the second scheme uses the channel $N-L$ times to transmit an M-ary message. Calculate the probability of decoding error for these schemes and denote them as P_{e,h^0}, P_{e,h^1} and $P_e = \max_{m=1,\cdots,M} P_{e,m}$.*

Now consider system 4.1 with parameters $\sigma > 1$, $\eta > 0$ and $W > 0$. If you can find $\delta_1 > 0$ such that $\gamma(\eta, \delta_1) < 1$ then the design produces a bounded-error estimate for the system (4.1) with,

$$\sup_{n\geq 0} \mathbb{E}|x_n - \hat{x}_n|^\eta < \infty.$$

Proof: To prove this result it suffices to show that $\gamma(\eta, \delta_1) < 1$ implies that worst-case error X_n is η-stable. By (4.17) it suffices to show that,

$$\mathbb{E}\left[\sum_{n'=1}^{n-1} \prod_{n''=n'}^{n-1} \Upsilon(n'')\right]^\eta < \infty,\ n \geq 0. \tag{4.22}$$

where we have omitted the $2W$ term for notational simplicity. Let $\eta > 1$. By Hölder's inequality,

$$\begin{aligned}
&\left[\sum_{n'=1}^{n-1} \prod_{n''=n'}^{n-1} \Upsilon(n'')\right]^\eta \\
&= \left[\sum_{n'=1}^{n-1} \prod_{n''=n'}^{n-1} \left(\Upsilon(n'')(1+\delta_1)\right)(1+\delta_1)^{-(n-n')}\right]^\eta \\
&\overset{\frac{1}{p}+\frac{1}{q}=1}{\leq} \left[\sum_{n'=1}^{n-1} \left(\prod_{n''=n'}^{n-1} \left(\Upsilon(n'')(1+\delta_1)\right)\right)^p\right]^{\frac{\eta}{p}} \times \left[\sum_{n'=1}^{n-1} (1+\delta_1)^{-q(n-n')}\right]^{\frac{\eta}{q}} \\
&\overset{p=\eta}{=} \left[\sum_{n'=1}^{n-1} \left(\prod_{n''=n'}^{n-1} \left(\Upsilon(n'')(1+\delta_1)\right)\right)^\eta\right] \times \left[\sum_{n'=1}^{n-1} (1+\delta_1)^{-\frac{\eta}{\eta-1}(n-n')}\right]^{\eta-1} \\
&\equiv \left[\sum_{n'=1}^{n-1} \left(\prod_{n''=n'}^{n-1} \left(\Upsilon(n'')(1+\delta_1)\right)\right)^\eta\right] \times c_1.
\end{aligned}$$

For any $\eta > 1$ and $\delta_1 > 0$ note that c_1 is some finite constant that depends on η and δ_1. So we can bound (4.22) by,

$$\mathbb{E}\left[\sum_{n'=1}^{n-1} \prod_{n''=n'}^{n-1} \Upsilon(n'')\right]^\eta \leq c_1 \sum_{n'=1}^{n-1} \mathbb{E}\left(\prod_{n''=n'}^{n-1} \left(\Upsilon(n'')(1+\delta_1)\right)\right)^\eta. \tag{4.23}$$

Recalling the definition of $\Upsilon(n'')$ we write,

$$\begin{aligned}
&\prod_{n''=n'}^{n-1} \left(\Upsilon(n'')(1+\delta_1)\right) \\
&= \prod_{j=m_1}^{T^{i_{n'}}} \left(\Upsilon_j^{i_{n'}}(1+\delta_1)\right) \times \left(\prod_{i=i_{n'}}^{i_{n-1}-1} \prod_{j=1}^{T^i} \left(\Upsilon_j^i(1+\delta_1)\right)\right) \times \prod_{j=1}^{m_2} \left(\Upsilon_j^{i_{n-1}-1}(1+\delta_1)\right)
\end{aligned} \tag{4.24}$$

where i_n is the stopping time defined in (4.12) and $m_1 = n' - (T^1 + \cdots + T^{i_{n'}-1})$ and $m_2 = n-1-(T^1 + \cdots + T^{i_{n-1}-1})$. Now observe that $\mathbb{E}\prod_{i=1}^T \max(\Upsilon_i, 1)^\eta < \mathbb{E}\sigma^{\eta T}$. This is because the estimation error can never grow faster than σ. Furthermore, because $\gamma(\eta,\delta_1) < 1$ it follows by Lemma 4.3.1 that $\mathbb{E}\sigma^{\eta T} < \infty$. So $\mathbb{E}\prod_{i=1}^{T}\max(\Upsilon_i,1)^\eta = c_2$ for some constant that depends on η and σ. Now take expectations in (4.24) and keeping in mind that the $\Upsilon^1, \Upsilon^2, \cdots$ are independent,

$$\begin{aligned}
&\mathbb{E}\prod_{n''=n'}^{n-1}\big(\Upsilon(n'')(1+\delta_1)\big)^\eta \\
&= \mathbb{E}\prod_{j=m_1}^{T^{i_{n'}}}\Big(\Upsilon_j^{i_{n'}}(1+\delta_1)\Big)^\eta \times \left(\prod_{i=i_{n'}}^{i_{n-1}-1}\mathbb{E}\prod_{j=1}^{T^i}\Big(\Upsilon_j^i(1+\delta_1)\Big)^\eta\right)\times \mathbb{E}\prod_{j=1}^{m_2}\Big(\Upsilon_j^{i_{n-1}-1}(1+\delta_1)\Big)^\eta \\
&\le (c_2)^2\prod_{i=i_{n'}}^{i_{n-1}-1}\mathbb{E}\prod_{j=1}^{T^i}\Big(\Upsilon_j^i(1+\delta_1)\Big)^\eta \\
&= (c_2)^2\prod_{i=i_{n'}}^{i_{n-1}-1}\mathbb{E}\prod_{j=1}^{T}\Big(\Upsilon_j(1+\delta_1)\Big)^\eta \\
&= (c_2)^2\gamma(\eta,\delta_1)^{i_{n-1}-i_{n'}-1}.
\end{aligned}$$

So we can bound (4.23) and consequently (4.22) by,

$$\mathbb{E}\left[\sum_{n'=1}^{n-1}\prod_{n''=n'}^{n-1}\Upsilon(n'')2W\right]^\eta \le c_1c_2^2\sum_{n'=1}^{n-1}\gamma(\eta,\delta_1)^{i_{n-1}-i_{n'}-1} \tag{4.25}$$

To finish the job we employ a large deviations bound. First write,

$$\gamma(\eta,\delta_1)^{i_{n-1}-i_{n'}-1} \le \gamma(\eta,\delta_1)^{-1}\Big(\mathbb{P}\big(i_{n-1}-i_{n'}<i\big)+\gamma(\eta,\delta_1)^i\Big) \tag{4.26}$$

Now, because i_n is a stopping time, by definition,

$$\mathbb{P}\big(i_{n-1}-i_{n'}<i\big) = \mathbb{P}\big(T^1+\cdots T^i \ge n-1-n'\big). \tag{4.27}$$

Letting $\mu = \mathbb{E}T$ and $n-1-n' = \lceil i\mu\rceil$ and invoking Markov's inequality we obtain,

$$\begin{aligned}
\mathbb{P}\big(T^1+\cdots T^i \ge n-1-n'\big) &\le \frac{\mathbb{E}(2^{\theta T})^i}{2^{\theta(n-1-n')}} \\
&= 2^{-(n-1-n')\left[\theta-\frac{\log_2\mathbb{E}2^{\theta T}}{\mu}\right]}
\end{aligned} \tag{4.28}$$

for any $\theta > 0$. But an old result (Durrett, Lemma 9.4) is that the exponent $\theta - \frac{\log_2\mathbb{E}2^{\theta T}}{\mu}$ is strictly positive for some small θ provided that there exists some $\theta > 0$ for which $\mathbb{E}2^{\theta T} < \infty$. This last condition is implied by our hypothesis $\gamma(\eta,\delta_1) < 1$. So using (4.28) in (4.27) in (4.26) we conclude that there exists some constant $c_3 < 1$ such that,

$$\gamma(\eta,\delta_1)^{i_{n-1}-i_{n'}-1} \le c_3^{n-1-n'}.$$

Plugging this back into (4.25) concludes the proof of Theorem 4.3.1 for the case $\eta > 1$. The proof follows for the case $\eta \le 1$ by replacing Hölder's inequality above with Jensen's inequality. □

Theorem 4.3.1 gives us a fairly general method of designing stable estimators and, consequently, anytime codes for arbitrary discrete-time memoryless channels. We first select a channel and a rate $0 \le R \le C$. These are the design specifications. Based on the choice of channel and rate we solve the following optimization problem,

$$
\begin{aligned}
&\text{maximize } \eta \qquad (4.29)\\
&\text{subject to:}\\
&1) \left(M\big(\sigma(1+\delta_1)\big)^L\right)^\eta P_{e,h^1}\\
&\quad + \left(M^{-1}\big(\sigma(1+\delta_1)\big)^N\right)^\eta (1-P_{e,h^1})(1-P_e)\\
&\quad + (1-P_{e,h^1})\frac{P_e}{2P_{e,h^0}} \times \left[1-\sqrt{1-4\left(\sigma(1+\delta_1)\right)^{(N+L)\eta}P_{e,h^0}\big(1-P_{e,h^0}\big)}\right] < 1
\end{aligned}
$$

2) $N, L, M, P_{e,h^0}, P_{e,h^1}, P_e$ are feasible: There exists a $(L, 2)$ block code whose probability of decoding error is P_{e,h^0} and P_{e,h^1} and there exists an $(N - L, M)$ block code whose worst case probability of decoding error is P_e.

3) $\delta_1 > 0$

4) $\sigma = 2^R + \delta_2,\ \delta_2 > 0$

Finally, having solved the optimization for a positive η we proceed to construct a rate R anytime encoder-decoder pair with anytime exponent $E_a = \eta R$ as described in Theorem 3.4.1.

Observe that δ_1 and δ_2 are slack parameters. The design space is therefore nothing but the space of all $(L, 2)$ and $(N - L, M)$ block codes. Since L, N and M are also unconstrained the design space is simply the space of all block codes that satisfies the first constraint. This is not a standard optimization problem. We do not know how to solve it in a generalized sense. It is also not obvious whether the exponent $E_a = \eta R$ predicted by (4.29) is truly optimal. For the Binary Erasure channel we will see that (4.29) is trivially solved and yields the optimal exponent. For the Binary Symmetric channel we are able to limit the design space to a small subset of block codes such that the exponent predicted by (4.29) is 'good'.

Remark 4.3.1 *The δ_1 is the slack parameter required by Theorem 4.3.1 and the δ_2 is the slack parameter required by Theorem 3.4.1. Both parameters may be selected arbitrarily small without any effect on the anytime exponent of the encoder-decoder pair. However, the coefficient β in Definition 3.1.1 does depend on δ_1 and δ_2.*

4.3.4 The formulas are tight for the Binary Erasure Channel

Consider a Binary Erasure Channel. Observe that the optimal encoder-decoder (cf. Section 3.3.2) coincides with the time-sharing encoder-decoder such that $L = 0$, $N = 1$ and $M = 2$. The parameter L is degenerate because the decoder knows when channel erasures occur. So there is no real need for 'back-up' or 'continue' messages. So the question is not whether the encoder-decoder is optimal. Rather we check whether the optimization problem in (4.29) yields the optimal anytime exponent.

We evaluate the anytime exponent $E_a = \eta R$ that is predicted by (4.29). For the optimal encoder-decoder we have $L = 0$, $N = 1$ and $M = 2$ which leads to the $P_{e,h^0} = P_{e,h^1} = 0$ and $P_e = \varepsilon$ where ε is the erasure probability of the channel. We need to evaluate the first constraint in (4.29), or equivalently, the expanded version in (4.21). By direct substitution the summation in (4.21) is simplified as,

$$(1-P_{e,h^1})P_e \times \sum_{K=1}^{\infty} \frac{1}{K}\binom{2(K-1)}{K-1}\big(\sigma(1+\delta_1)\big)^{(N+L)K\eta}\big(P_{e,h^0}\big)^{K-1}\big(1-P_{e,h^0}\big)^{K}$$
$$= \varepsilon\big(\sigma(1+\delta_1)\big)^{(N+L)\eta}.$$

Plugging this back into (4.21), we obtain,

$$\Big(2^{-1}\sigma(1+\delta_1)\Big)^{\eta}(1-\varepsilon)+\Big(\sigma(1+\delta_1)\Big)^{\eta}\varepsilon<1.$$

Rearranging terms and taking logs,

$$\begin{aligned}\log\sigma(1+\delta_1) &< -\frac{1}{\eta}\log\left[\frac{1}{2^{\eta}}(1-\varepsilon)+\varepsilon\right]\\ &= 1-\frac{1}{\eta}\log\Big[1+\varepsilon\big(2^{\eta}-1\big)\Big].\end{aligned} \tag{4.30}$$

Using Theorem 3.4.1 and observing that $\delta_1>0$ is arbitrary we see that our formulation predicts that the anytime exponent $E_a=\eta R$ is achievable by the rate $R=1-\frac{1}{\eta}\log\Big[1+\varepsilon\big(2^{\eta}-1\big)\Big]$ optimal encoder-decoder. Comparing this with the parametric representation in (3.13) reveals that the predicted values for the rate and exponent are the optimally achievable values for the Binary Erasure Channel.

4.3.5 The formulas are decent for a good Binary Symmetric Channel

In contrast with the situation for the Binary Erasure Channel there is no trivially optimal encoder-decoder for the case of the Binary Symmetric Channel. Unlike the case of the BEC, bits that were correctly decoded in the past can be corrupted by errors that occur in the present. Consequently, we are necessarily lead to think about the entire space of block codes to find a good solution to (4.29). In particular, we need to consider codes that provide a good trade-off amongst the rate at which new information is transmitted and the rate at which control information (ie. back-up or continue) is transmitted.

Let's begin with a good Binary Symmetric Channel. By this we mean that the channel cross-over probability ϵ is very small so that channel errors are very unlikely. Because errors are unlikely we consider those block codes that provide very little or no forward error protection. This intuition turns out to be correct at modest rates. For such rates we pick $L=1$ so that the binary control information ('backup' or 'continue') is transmitted using a single unprotected bit. We also pick $M=2^{N-L}$ ($N-L=1,2,\cdots$) so that the new true information is transmitted without any forward error protection. For these choices of N, L and M the optimization problem in (4.29) reduces to,

$$\text{maximize } \eta \tag{4.31}$$

subject to:

$$1)\ \Big(2^{N-1}\sigma(1+\delta_1)\Big)^{\eta}\epsilon$$
$$+\Big(2^{-(N-1)}\big(\sigma(1+\delta_1)\big)^{N}\Big)^{\eta}(1-\epsilon)^{N}$$

$$+ (1-\epsilon)\frac{(1-(1-\epsilon)^{N-1})}{2\epsilon} \times \left[1 - \sqrt{1 - 4\left(\sigma(1+\delta_1)\right)^{(N+1)\eta}\epsilon(1-\epsilon)}\right] < 1$$

2) $N = 2, 3, 4, \cdots$

3) $\delta_1 > 0$

4) $\sigma = 2^R + \delta_2,\ \delta_2 > 0$

To solve this optimization problem we need to determine the optimal $E_a(R) = \eta R$ as a function of the rate R. Observe that δ_1 and δ_2 are slack parameters. Consequently, for a fixed N we can calculate $E_a^{(N)} = \eta R$ as a function of R and then consider the optimal function E_a as the least upper bound over the entire family $\{E_a^{(N)}\}_{N=1,2,\cdots}$.

In Figure 4.2 we have numerically calculated $E_a^{(N)}$ $(N = 2, 3, \cdots, 11)$ for a Binary Symmetric Channel with cross-over probability $\epsilon = 0.01$. The blue curve is the upper bound E_a^+. The dashed curve is the sphere-packing bound and it is drawn only for reference. The gap between the anytime exponent $E_a^{(2)}$ and E_a^+ at $R = 1/4$ is fairly small and supports our intuition that forward error protection is not necessary for a good BSC channel. On the other hand, the gap is significant at very low and high rates.

High rates are not achievable because of the time-sharing property of our encoder-decoder. A fixed ratio, L/N, of the time is devoted to the transmission of control information, the 'back-up' or 'continue' signals. So unless L/N is very small rates close to capacity cannot be achieved. On the other hand, if we design L/N very small then we significantly increase the chances of forward transmission errors and consequently the transmission and retransmission of backups. There is very little or nothing that we can do with the proposed scheme to recover performance at high rates.

At low rates a more interesting situation occurs. At rates close to zero, true information bits are rare. In between the true bits, the encoder is transmitting an excess amount of control information (ie. 'back-up' or 'continue'). But this control information is unprotected so that 'false backups' occur frequently. Such backups are costly and lead to a performance loss. However, observe that there is a simple technique to recover this loss – do not transmit an excess amount of control information. To see how this can be done, consider the peak of the curve labeled $N = 2$ and denote this point by $(R^\star, E_a(R^\star))$. For rate $R < R^\star$ we can modify the encoder and decoder so that the encoder transmits true information at the optimal rate $R^\star$ and transmits 'dummy' bits during the excess time. The encoder knows that certain bits are dummy. Consequently, the reliability $E_a(R^\star)$ is achievable for all $R < R^\star$. This is illustrated in the figure by drawing a solid black line tangent to the peak.

4.4 Improving the estimator for a non-perfect channel

In the previous section we saw that a very simple time-sharing estimator performs well for a 'good' Binary Symmetric Channel. It turns out, however, that this estimator performs poorly if the channel is very noisy. In this section we modify the estimator for improved performance on a very noisy Binary Symmetric Channel. Our approach is to introduce an 'inner' encoder-decoder that converts the noisy channel into a good virtual channel. The simple estimator is then applied to the virtual channel.

4.4.1 The inner encoder-decoder: creating the virtual channel

The inner encoder-decoder pair is Schalkwijk's 1-up J-down (J is a parameter to be determined later) previously described in Section 2.4.1. We repeat the description here to avoid excessive page flipping.

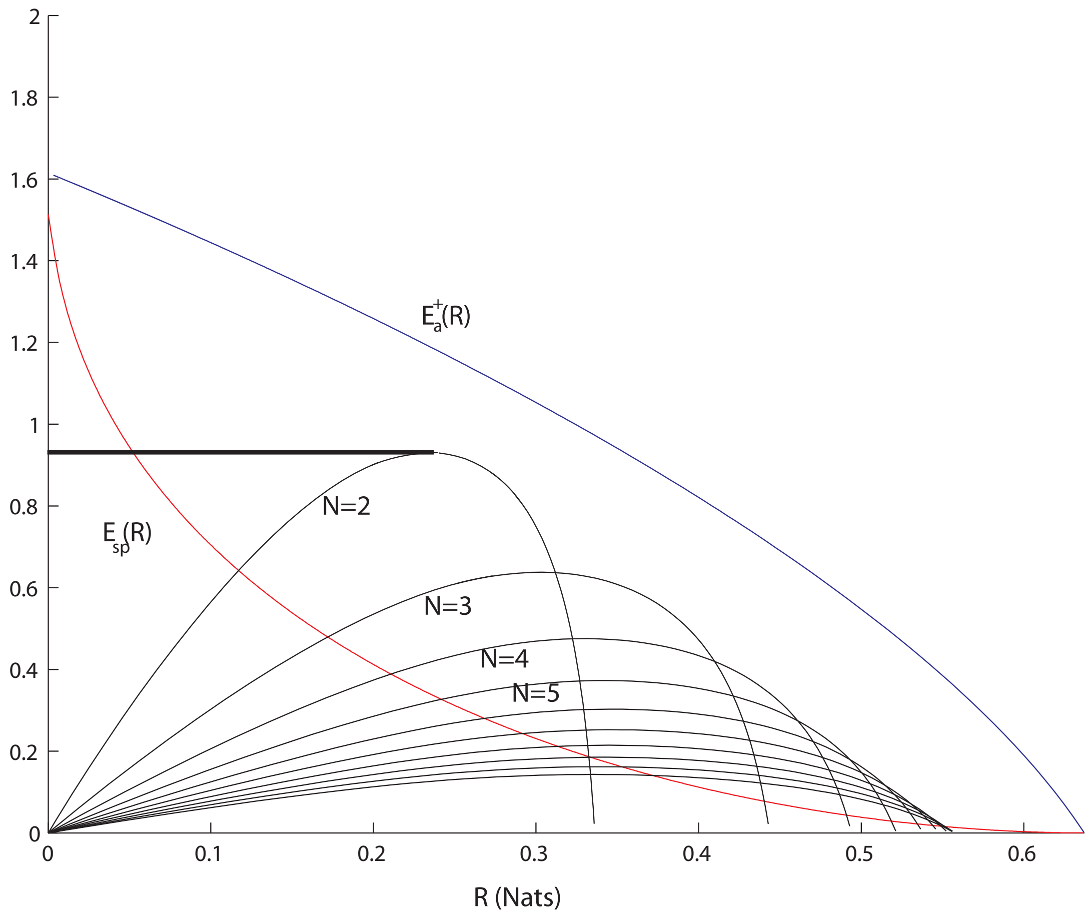

Figure 4.2: Anytime exponent $E_a^{(N)}$, $N = 2, 3, \cdots, 11$ of time-sharing code for a Binary Symmetric Channel with cross-over probability $\epsilon = 0.01$ and no forward error protection.

At time n a bit $m \in \{0, 1\}$ is handed off to the inner encoder. The encoder repeatedly emits m until the decoder makes a decision at a random time $n + T^I$, $T^I \geq J$.

As before the decoder is described by a Mealy machine consisting of $2J + 1$ discrete states as illustrated in Figure 2.4. At time n the decoder starts in state J. Suppose that the decoder is in state $z_{n'}$ at time $n \leq n' \leq n+T^I$. Upon receiving a one at time n' the decoder jumps to state $z_{n'+1} = z_{n'} + 1$. Upon receiving a zero at time n' the decoder jumps to state $z_{n'+1} = z_{n'} - 1$ at time $n' + 1$. The decoder makes a decision in favor of $m' = 0$ at the first time instant $n + T^I$ at which it makes a jump into state zero. Conversely, the decoder makes a decision in favor of $m' = 1$ at the first time instant $n + T^I$ at which it makes a jump into state $2J$.

We write,

$$q_m(t) = \mathbb{P}\big(T^I = J + 2t, m' \neq m \mid \text{ bit } m \text{ was transmitted}\big),$$
$$p_m(t) = \mathbb{P}\big(T^I = J + 2t, m' = m \mid \text{ bit } m \text{ was transmitted}\big).$$

By symmetry $p_1(t) = p_0(t)$ and $q_1(t) = q_0(t)$ so we will simply write $p(t)$ and $q(t)$ to denote the respective quantities. The statistics of $p(t)$ and $q(t)$ are given by the matrix expressions (2.19) and (2.20). For $J = 2$ and $J = 3$ these expressions reduce to,

$$p(t) = J^t(1-\epsilon)^{J+t}\epsilon^t,\ t \geq 0, \tag{4.32}$$

and

$$q(t) = J^t\epsilon^{J+t}(1-\epsilon)^t,\ t \geq 0. \tag{4.33}$$

We are not able to come up with such simple expressions for $J > 3$. The interested reader may wish to verify that for $J = 2$ and $J = 3$ that (4.32) and (4.33) indeed define a true probability distribution. In particular,

$$\sum_{t=0}^{\infty} p(t) + q(t) = 1.$$

Observe that the inner encoder-decoder pair described here is used for the reliable transmission of a single bit. The pair effectively converts a noisy Binary Symmetric Channel into something that resembles a less noisy binary channel. More precisely, we think of the transfer function from the input of the encoder to the output of the decoder to be a virtual channel. The transfer function is not quite a standard channel because different bits experience different delays from the time they enter the encoder until the time they leave the decoder.

4.4.2 The outer decoder

An outer decoder maintains an interval of confidence denoted as $I_n = [l_n, u_n],\ n \geq 0$. The decoder believes that the true state is contained in the interval and therefore produces the estimate $\hat{x}_n = (l_n + u_n)/2$ for each time $n \geq 0$. The description of the outer decoder then follows by describing the update rule for the interval of confidence.

At time n the outer decoder, just having received a code-word, adaptively listens for a new variable-length code-word. Let N be a parameter to be determined later. The outer decoder first receives a bit $b_{n+\tau}$ from the inner decoder. The time τ is a random variable that depends on the statistics of the inner encoder-decoder. The outer decoder uses $b_{n+\tau}$ to distinguish between a binary hypothesis: 'back-up'($b_{n+\tau} = 0$) or 'continue' ($b_{n+\tau} = 1$).

Say that a decision is made in favor of the 'continue' hypothesis. The decoder then receives N bits $(b_{n+\tau+t_1}, b_{n+\tau+t+1+t_2}, \cdots, b_{n+\tau+\sum_{i=1}^N t_i})$ from the inner decoder. The times $t_1, t_2, \cdots, t_N$ are randomly determined by the statistics of the inner encoder-decoder. These N bits are used to select one of 2^N possible messages according to the maximum a posterior probabilities. Denoting the selected message by m the decoder concludes that the true state x_n was in the m-th cut $[p_m, p_{m+1}]$ of a 2^N partition of $I_{n-1} = [l_{n-1}, u_{n-1}]$,

$$l_{n-1} = p_1 \leq p_2 \leq \cdots \leq p_{2^N} = u_{n-1}.$$

Consequently, at time $n + \tau + \sum_{i=1}^N t_i$ the decoder updates the interval of confidence as,

$$I_{n+\tau+\sum_{i=1}^N t_i} = \sigma^{\tau+\sum_{i=1}^N t_i}[p_m, p_{m+1}] + [-\Gamma^\sigma_{\tau+\sum_{i=1}^N t_i}(W), \Gamma^\sigma_{\tau+\sum_{i=1}^N t_i}(W)]. \tag{4.34}$$

The outer decoder saves the message m on a stack for future use and listens for a new code-word at time $n + \tau + \sum_{i=1}^N t_i$.

Suppose now that the decoder had decided in favor of the 'back-up' hypothesis at time $n+\tau$. The decoder concludes that the true state x_{n-1} was not in the interval I_{n-1} so the decoder pops a partition index m from the top of the stack and enlarges the interval as,

$$I_{n+\tau} = \sigma^{\tau}\left[\, l_{n-1} - m|I_{n-1}|,\, u_{n-1} + \big(2^N - (m+1)\big)|I_{n-1}|\,\right] + \big[-\Gamma_{\tau}^{\sigma}(W),\, \Gamma_{\tau}^{\sigma}(W)\big], \tag{4.35}$$

where Γ is the function defined in (4.6). In words, the interval is extended to the left by $m|I_{n-1}|$ and to the right by $(2^N - (m+1))$ and the resulting interval is propagated by the system equation (4.1). At time $n+\tau$ the decoder waits for a new variable-length codeword.

Observe that independent of the value of the partition index m the interval is updated such that,

$$\sigma^{\tau} I_{n-1} + \left[-\Gamma_{\tau}^{\sigma}(W),\, \Gamma_{\tau}^{\sigma}(W)\right] \subseteq I_{n+\tau}, \tag{4.36}$$

This concludes the description of the outer decoder.

Observe also that the outer decoder described here is almost identical to the description of the time-sharing decoder of Section 4.3.1. To see this take $L=1$ and $M=2^N$ in the description of the time-sharing decoder. The only notable difference is that the input to the time-sharing decoder comes from a true Binary Symmetric Channel whereas the input to the decoder described here comes from the virtual channel induced by the inner encoder-decoder.

4.4.3 The outer encoder

It should be no surprise the description of the outer encoder follows from the description of the time-sharing encoder of Section 4.3.2. The description begins by considering the behavior of the outer encoder at a time $n-1$ at which the true state x_{n-1} is contained in the decoder's interval of confidence, $x_{n-1} \in I_{n-1}$, and the decoder is waiting for a new code-word. The outer encoder advises the inner encoder to emit a 'one' indicating the continue hypothesis. A random τ time units later the decoder decides on either the continue or the back-up hypothesis. Suppose that — with probability $p(\tau)$ — the decoder decided on the continue hypothesis. In this case the encoder calculates the cut of the 2^N partition of I_{n-1} that contained x_{n-1} and hands off the N bits that represents the cut to the inner encoder. These bits are transmitted to the outer decoder at random times $n+\tau+t_1, n+\tau+t+1+t_2, \cdots, n+\tau+\sum_{i=1}^{N} t_i$ respectively. If there were no errors in the transmission of these N bits, at time $n+\tau+\sum_{i=1}^{N} t_i$ the decoder updates its interval of confidence according to (4.34) such that $x_{n+\tau+\sum_{i=1}^{N} t_i} \in I_{n+\tau+\sum_{i=1}^{N} t_i}$ and waits for a new code-word. This is the good case.

There are two more possible case scenarios. With probability $q(\tau)$ the decoder could have decided in favor of the 'back-up' hypothesis at time $n+\tau$. So at time $n+\tau$ the decoder mistakenly enlarges the interval according to (4.35) and waits for a new code-word. This case scenario is not so grim because at time $n+\tau$ we still have that $x_{n+\tau} \in I_{n+\tau}$.

Finally, the worst case scenario occurs if at time $n+\tau$ the decoder decides in favor of 'continue' but subsequently an error occurs in the transmission of the N bits that represent the cut of the partition of I_{n-1} that contained x_{n-1}. At time $n+\tau+\sum_{i=1}^{N} t_i$ the decoder mistakenly updates the interval according to (4.34) but the updated interval no longer contains the true state; $x_{n+\tau+\sum_{i=1}^{N} t_i} \notin I_{n+\tau+\sum_{i=1}^{N} t_i}$. Starting at the troublesome time $n+\tau+\sum_{i=1}^{N} t_i$ the encoder attempts to bias the decoder to always decide in favor of the 'back-up' hypothesis. The encoder goes back to normal operation at the smallest time n' such that the decoder decided in favor of an equal number of back-up and continue hypotheses in the interval $\{n, \cdots, n'\}$. It follows by (4.36) that,

$$\sigma^{n'-n} I_{n-1} + \big[-\Gamma^{\sigma}_{n'-n}(W), \Gamma^{\sigma}_{n'-n}(W) \big] \subseteq I_{n'}.$$

This concludes the description of the encoder.

4.4.4 Mathematical model of the modified system

The description of the outer encoder-decoder pair is almost identical to that of the time-sharing encoder-decoder pair of Sections 4.3.2 and 4.3.1. The information about the state is encoded in exactly the same way. The only thing that changes are the times at which the encodings occur. The latter is due to the virtual channel induced by the inner encoder-decoder. Consequently, the 'renewal' property is maintained and the model that was used to analyze the original system also applies here.

In conjunction with Lemma 4.3.1 we define a stochastic process $\Upsilon = \{\Upsilon_1, \cdots, \Upsilon_T\}$ with random termination time T such that the worst-case estimation error X_n is bounded by,

$$X_{n+1} \leq \Upsilon(n) X_n + 2W, \; X_0 = 0, \; n \geq 0. \tag{4.37}$$

We remind the reader that the worst-case estimation error is defined as $X_n = \sup_{w^n} |x_n - \hat{x}_n|$. We also remind the reader that the notation $\Upsilon(n)$ indicated the n-th entry of the process $\{\Upsilon^1_1, \Upsilon^1_2, \cdots, \Upsilon^1_{T^1}, \Upsilon^2_1, \Upsilon^2_2, \cdots, \Upsilon^2_{T^2}, \cdot$ in which the $\Upsilon^i = \{\Upsilon^i_1, \Upsilon^i_2, \cdots, \Upsilon^i_{T^i}\}$ are independent copies of Υ for all $i \geq 1$. Supposing that we are able to determine the distribution of Υ then we immediately get the following analog of Theorem 4.3.1, the proof of which follows from that of Theorem 4.3.1.

Theorem 4.4.1 *Consider a modified estimator with parameters J and N and consider system (4.1) with parameters $\sigma > 1$, $\eta > 0$ and $W > 0$. Let,*

$$\gamma(\delta_1, \eta, J, N) = \mathbb{E} \prod_{i=1}^{T} \big(\Upsilon_i (1 + \delta_1) \big)^{\eta}.$$

If you can find $\delta_1 > 0$ such that $\gamma(\delta_1, \eta, J, N) < 1$ then the modified estimator produces a stable estimate for system (4.1) with,

$$\sup_{n>0} \mathbb{E}|x_n - \hat{x}_n|^{\eta} < \infty.$$

To make this a useful result we determine the distribution of Υ and subsequently calculate $\gamma(\delta_1, \eta, J, N)$. Consider the three situations that are possible in between renewals. These are the cases of a false backup, a successful transmission and a recapturing sequence. The false backup is the easiest and follows by selecting,

$$\mathbb{P}\big(T = \tau, \prod_{i=1}^{T} \Upsilon_i = 2^N \sigma^{\tau}\big) = q(\tau), \; \tau \geq 0.$$

To see this observe that $q(\tau)$ is nothing but the probability that a 'continue' hypothesis is transmitted in error in τ time steps by the inner encoder-decoder. Then use $\Upsilon_i = 2^N \sigma^{\tau}$ in (4.37) and compare with the actual dynamics of the interval given by (4.35). The case of a successful transmission is also straightforward and follows from (4.34) by selecting,

$$\mathbb{P}\big(T = \tau + \sum_{i=1}^{N} t_i,\ \prod_{i=1}^{T} \Upsilon_i = 2^{-N}\sigma^T\big) \\ = p(\tau)p(t_1)p(t_2)p(t_3)\cdots p(t_N), \\ \tau, t_1, t_2, \cdots, t_N \geq 0.$$

The quantity $p(\tau)$ represents the probability of successfully transmitting the continue hypothesis in τ time steps and the quantity $p(t_1)p(t_2)\cdots p(t_N)$ represent the probabilities of the subsequent successful transmission of the partition cut.

To complete the model we need to describe the distribution of Υ that corresponds to a recapturing sequence. Analogous to the situation in Lemma 4.3.1 we consider a recapturing sequence consisting of K errors and K true backups,

$$\begin{aligned}\mathbb{P}\big(T = \sum_{k=1}^{K}\sum_{i=1}^{N}\big(\tau^{(k)} + t_i^{(k)}\big) + \sum_{k=1}^{K}\varsigma^{(k)},\ \prod_{i=1}^{T}\Upsilon_i = \sigma^T\big) \\ = p(\tau^{(1)})P \\ \times \frac{1}{K}\binom{2(K-1)}{K-1} q(\tau^{(2)})q(\tau^{(3)})\cdots q(\tau^{(K)}) \\ \times p(\varsigma^{(1)})p(\varsigma^{(2)})p(\varsigma^{(3)})\cdots p(\varsigma^{(K)}) \\ \times \prod_{k=2}^{K}\Big(\big(p(t_1^{(k)}) + q(t_1^{(k)})\big)\big(p(t_2^{(k)}) + q(t_2^{(k)})\big)\cdots\big(p(t_N^{(k)}) + q(t_N^{(k)})\big) \\ \tau^{(k)}, \varsigma^{(k)}, t_1^{(k)}, t_2^{(k)}, \cdots, t_N^{(k)} \geq 0,\ k = 1, 2, \cdots, K.\end{aligned}$$

in which P denotes the overall probability of error in transmitting the partition cut,

$$P = 1 - p(t_1)p(t_2)\cdots p(t_N).$$

To verify these expressions it suffices to note that $p(\tau^{(1)})$ is the probability of successfully transmitting the first continue hypothesis in $\tau^{(1)}$ time steps and P is the probability that an error occurs in one or more of the N bits in the subsequent transmission of the partition cut. The remaining quantities describe the transmission of the $K-1$ false continues and K true backups. Keep in mind that a false continue occurs if a 'back-up' is understood as a 'continue' and the decoder expects to receive N bits after deciding in favor of a continue. Each of the $K-1$ false continues is captured by the product $q(\tau^{(k)})\big(p(t_1^{(k)}) + q(t_1^{(k)})\big)\big(p(t_2^{(k)}) + q(t_2^{(k)})\big)\cdots\big(p(t_N^{(k)}) + q(t_N^{(k)})\big)$ for $k = 2, \cdots, K$. Observe that $p(t) + q(t)$ is nothing but the probability that a bit is transmitted in t time steps. Finally, each of the K true backups is captured by the terms $p(\varsigma^{(k)})$, $k = 1, 2, \cdots, K$, since this quantity describes the successful transmission of the backup hypothesis in $\varsigma^{(k)}$ time steps.

Having determined the distribution of Υ we revert our attention to the calculation of the quantity $\gamma(\delta_1, \eta, J, N)$ which is central to Theorem 4.4.1. By direct calculation,

$$\gamma(\delta_1, \eta, J, N) = \mathbb{E}\prod_{i=1}^{T}\big(\Upsilon_i(1+\delta_1)\big)^{\eta}$$

$$= \sum_{\tau \geq 0} \Big(2^N \big(\sigma(1+\delta_1)\big)^{(J+2\tau)} \Big)^\eta q(\tau)$$

$$+ \sum_{\substack{\tau \geq 0 \\ t_1, \cdots, t_N \geq 0}}^{\infty} \Big(2^{-N} \big(\sigma(1+\delta_1)\big)^{(J(N+1)+2\tau+2t_1+\cdots+2t_N)} \Big)^\eta p(\tau)p(t_1)\cdots p(t_N)$$

$$+ \sum_{K=1}^{\infty} \sum_{\substack{\tau^{(k)}, \varsigma^{(k)} \geq 0 \\ t_1^{(k)}, t_2^{(k)}, \cdots, t_N^{(k)} \geq 0 \\ k=1,2,\cdots,K}} \Big(\big(\sigma(1+\delta_1)\big)^{\sum_{k=1}^K \sum_{i=1}^N \left(\tau^{(k)}+t_i^{(k)}\right) + \sum_{k=1}^K \varsigma^{(k)}} \Big)^\eta$$

$$\times p(\tau^{(1)}) P \frac{1}{K} \binom{2(K-1)}{(K-1)}$$

$$\times \prod_{k=1}^K p(\varsigma^{(k)})$$

$$\times \prod_{k=2}^K \prod_{n=1}^N q(\tau^{(k)}) \big(p(t_n^{(k)}) + q(t_n^{(k)})\big)$$

For $J = 2$ and $J = 3$ this seemingly long expression reduces to something quite simple. This is because the sums are geometric for these very special cases. To make the simplifications first observe that probabilities add up to one so that,

$$\sum_{t_1,\cdots,t_N \geq 0} \Big(\big(\sigma(1+\delta_1)\big)^{2t_1+2t_2+\cdots+2t_N} \Big)^\eta P$$

$$= 1 - \sum_{t_1,\cdots,t_N \geq 0} \Big(\big(\sigma(1+\delta_1)\big)^{2t_1+2t_2+\cdots+2t_N} \Big)^\eta p(t_1)p(t_2)\cdots p(t_N)$$

$$= 1 - \Big(\sum_{t \geq 0} \big(\sigma(1+\delta_1)\big)^{2t\eta} p(t) \Big)^N .$$

Using this and (4.32) and (4.33) in the lengthy expression for $\gamma(\delta_1, \eta, J, N)$ we obtain,

$$\gamma(\delta_1, \eta, J, N) = \Big(2^N \big(\sigma(1+\delta_1)\big)^J \Big)^\eta Q_J$$

$$+ \Big(2^{-N} \big(\sigma(1+\delta_1)\big)^{J(N+1)} \Big)^\eta (P_J)^{N+1}$$

$$+ P_J \frac{\big(1-(P_J)^N\big)}{2Q_J\big(P_J+Q_J\big)}$$

$$\times \sum_{K=1}^{\infty} \frac{1}{K} \binom{2(K-1)}{(K-1)} \Big(\big(\sigma(1+\delta_1)\big)^{J(N+2)} \Big)^\eta P_J^K Q_J^K (P_J+Q_J)^K,$$

where P_J and Q_J are shorthand for,

$$P_J = \sum_{t \geq 0} \big(\sigma(1+\delta_1)\big)^{2t\eta} p(t)$$

$$= \frac{(1-\epsilon)^J}{1 - J\big(\sigma(1+\delta_1)\big)^{2\eta} \epsilon(1-\epsilon)},$$

and

$$\begin{aligned} Q_J &= \sum_{t\geq 0} \big(\sigma(1+\delta_1)\big)^{2t\eta} q(t) \\ &= \frac{\epsilon^J}{1 - J\big(\sigma(1+\delta_1)\big)^{2\eta}\epsilon(1-\epsilon)}. \end{aligned}$$

To conclude we make use of the following generating function,

$$(1-\sqrt{1-4x})/2 = \sum_{K=1}^{\infty} (1/K) \binom{2(K-1)}{K-1} x^K$$

which holds for $x < 1/4$. Applying this to the series in the last expression for $\gamma(\delta_1, \eta, J, N)$ we write,

$$\begin{aligned} \gamma(\delta_1, \eta, J, N) &= \left(2^N \big(\sigma(1+\delta_1)\big)^J\right)^\eta Q_J \\ &+ \left(2^{-N}\big(\sigma(1+\delta_1)\big)^{J(N+1)}\right)^\eta (P_J)^{N+1} \\ &+ P_J \frac{\big(1-(P_J)^N\big)}{2Q_J\big(P_J+Q_J\big)} \\ &\quad \times \left[1 - \sqrt{1 - 4\big(\sigma(1+\delta_1)\big)^{J(N+2)\eta} P_J Q_J (P_J+Q_J)}\right]. \end{aligned} \tag{4.38}$$

4.4.5 Performance for a noisy Binary Symmetric Channel

We now evaluate the anytime exponent of the modified estimator for a very noisy Binary Symmetric Channel with cross-over probability $\epsilon = 1/4$. To evaluate the exponent we take out usual approach and we use (4.38) in Theorem 4.4.1 to obtain the following optimization problem,

$$\begin{aligned} &\text{maximize } \eta \\ &\text{subject to:} \\ &1)\ \left(2^N \big(\sigma(1+\delta_1)\big)^J\right)^\eta Q_J \\ &\quad + \left(2^{-N}\big(\sigma(1+\delta_1)\big)^{J(N+1)}\right)^\eta (P_J)^{N+1} \\ &\quad + P_J \frac{(1-(P_J)^N)}{2Q_J\big(P_J+Q_J\big)} \\ &\quad\quad \times \left[1 - \sqrt{1 - 4\left(\sigma(1+\delta_1)\right)^{J(N+2)\eta} P_J Q_J \big(P_J+Q_J\big)}\right] < 1 \\ &2)\ N = 1, 2, 3, \cdots \\ &3)\ \delta_1 > 0 \\ &4)\ \sigma = 2^R + \delta_2,\ \delta_2 > 0 \end{aligned} \tag{4.39}$$

Then, by Theorem 4.4.1, if (η, R) is achievable, then the exponent $E_a(R) = \eta R$ is achievable.

We compare (4.39) with (4.31). By similarity in the construction we can solve for (4.39) using the same technique that was used to solve for (4.31). The basic idea is to observe that δ_1 and δ_2 are slack parameters and, consequently, for a fixed (J, N) we can calculate a new function $E_a^{(J,N)}(R) = \eta R$ as a function of R. Having computed $E_a^{(J,N)}$ for $N = 1, 2, \cdots$ and $J = 2, 3$ we then consider the optimal function E_a as the least upper bound over the entire family $\{E_a^{(J,N)}\}_{N=1,2,\cdots,\ J=2,3}$. These curves evaluated for the BSC for $J = 3$, $N = 1, 2, \cdots, 6$ and $\epsilon = 1/4$ are illustrated in Figure 4.3.

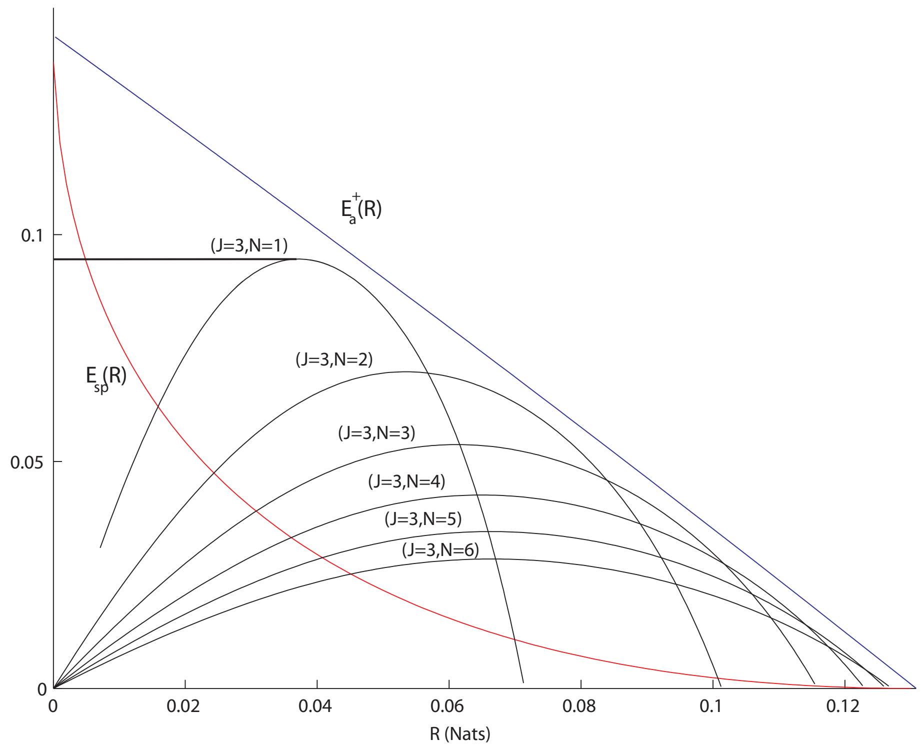

Figure 4.3: The improved anytime exponent for a noisy Binary Symmetric Channel with cross-over probability $\epsilon = 0.25$.

4.5 Discussion

If we compare the contour of the achievable regions in Figures 4.2 and Figure 4.3 we see that the aspect ratio of the contour is 'tighter' to the upper bound for the noisy channel. From this perspective, the code performs better for the noisy Binary Symmetric Channel (BSC) than it does for the good BSC. It turns out that this surprising result has a simple explanation.

To see what's going on we first keep in mind that the method of encoding is time-sharing so that a fixed proportion of the time is spent on transmitting control information (cf. the 'continue' or 'back-up' hypothesis). For a good channel errors rarely occur. In the absence of errors, however, the control information carries no 'useful' information and therefore may be seen as a waste of time that would have otherwise been used to transmit new information. This fixed overhead explains why rates close to capacity are not achievable for the good channel. In comparison, errors occur frequently in a very noisy channel. So the control information is useful most of the time and, consequently, the fixed overhead incurs a smaller hit.

Chapter 5

Regenerative relaying

In this brief chapter we consider the problem of the reliable transmission of real-time data over two serially interconnected channels as illustrated in Figure 5.1. We give a precise characterization of the source-to-destination anytime reliability. To avoid trivialities we assume that the interconnected channels are not symbol-by-symbol compatible. By this we mean that the relaying node must decode, interpret and re-encode the incoming data in a streaming fashion. Generalizations to more elaborate graph topologies are possible, however, these will not be addressed here.

Formally, consider the relaying system is illustrated in Figure 5.1. A source produces one bit of new information every time step. Denote the n-th bit by s_n. At time n, an encoder observes s_n. Based on its observations it emits $a_n \in \{1, 2, \cdots, 2^L\}$ through an L-bit erasure channel. The channel is described by the memoryless probability distribution,

$$\begin{aligned}&\mathbb{P}(b_n = a \mid a_n = a) = 1 - \varepsilon,\\ &\mathbb{P}(b_n = \emptyset \mid a_k = a) = \varepsilon,\ n \geq 0,\ a \in \{1, 2, \cdots, 2^L\}\end{aligned}$$

where $b_n \in \{1, 2, \cdots, 2^L\} \cup \{\emptyset\}$ denotes the channel output at time n. At time n a relay receives b_n and based on what it has received up to time n it produces an estimate $r(n) = \big(r_1(n), r_2(n), \cdots, r_n(n)\big)$ of the entire history of the source bits s_1^n. Based on its estimate $r(n)$ the relay emits a channel symbol c_n through a discrete-time memoryless, but otherwise arbitrary, channel (DMC). We will not need the transfer function of the DMC, except that the channel output at time n will be denoted by d_n.

Finally, the true destination receives d_n at time n. Based on what it has received up to time n it produces an estimate $\tilde{s}(n) = \big(\tilde{s}_1(n), \tilde{s}_2(n), \cdots, \tilde{s}_n(n)\big)$ of the entire history of the true source bits s_1^n. The basic problem is to find parameters $\alpha > 0$ and β such that,

$$\mathbb{P}\big(\tilde{s}_n(n + d) \neq s_n\big) \leq \beta\, 2^{-\alpha d},\ n \geq 0, d \geq 0, \tag{5.1}$$

or to determine when no such parameters exist. In particular, we are interested in the largest such parameter α which will be called the anytime exponent of the source-to-destination system.

Before we proceed to bury ourselves in equations it is instructive to check whether there is a simpler way to relay the information from the source to the destination. Suppose that the second DMC had exactly $2^L + 1$ inputs. In this case we could take the relay to simply forward the channel symbols b_n to the channel symbols c_n,

$$c_n = b_n,\ n \geq 0.$$

In this way we obtain a 'concatenated' discrete-time memoryless channel from the source to the destination. So the basic problem reduces to a point-to-point anytime coding problem. This would mean that there exists $\alpha, \beta > 0$ in (5.1) if and only if

$$C^{\mathcal{A}}(\alpha) > 1$$

where $C^{\mathcal{A}}(\alpha)$ is the α-anytime capacity of the concatenated channel.

On the other hand, if the input symbol set of the second DMC is not compatible with the output symbols of the L-bit erasure channel then there is no channel-to-channel mapping of symbols without reconstructing the true source bits at the relay. This is because the channel symbol c_n depends explicitly on each of the n bits in $s_1^n = (s_1, s_2, \cdots, s_n)$.

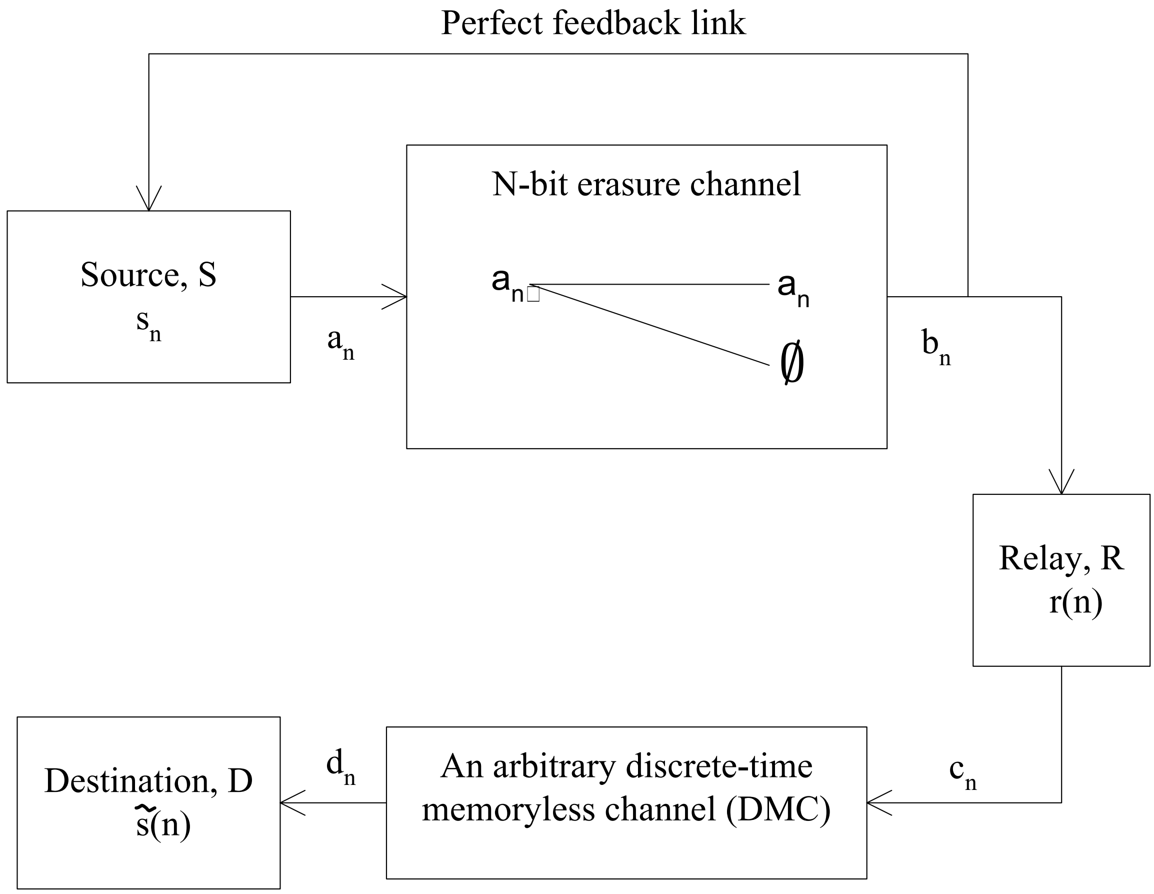

Figure 5.1: An abstract relaying system. The L-bit erasure channel is a discrete-time memoryless channel described by $\mathbb{P}(b_n = a \mid a_n = a) = 1 - \varepsilon$ and $\mathbb{P}(b_n = \emptyset \mid a_n = a) = \varepsilon$ for all $a \in \{1, 2, \cdots, 2^L\}$.

5.1 Producing an estimate at the relay

The system consisting of the source, L-bit Erasure Channel and the relay is modeled as an infinite first-in first-out (FIFO) binary queue. At each time n the source inserts a new bit at the end of the queue. Immediately after inserting a new entry, the source selects the first $l \leq L$ non-empty entries in the queue and emits these bits in a packet of length L over the L-bit Erasure Channel. If no erasure occurs then these l entries are removed from the queue. If an erasure occurs then the entries are left in the queue to be re-transmitted at time $n+1$. When n is strictly less than L the source inserts $L-l$ dummy bits to form an L-bit packet.

Suppose that by time n the relay has received $f(n)$ true information bits,

$$f(n) = n - \text{source's queue length},$$

where the queue length is a random variable depending on the number of erasures that have occurred up to time n. Because the relay knows when erasures occur it can determine the source's queue length and consequently $f(n)$. So at time n the relay produces its best estimate as,

$$r(n) = (s_1, s_2, \cdots, s_{f(n)}, 1, 1, \cdots, 1),\ n \geq 0, f(n) \leq n.$$

Clearly no better estimate is possible. So the estimate is completely determined by the dynamics of $f(n)$, or equivalently by the dynamics of the queue length.

On the other hand, the queue is a simple $D/M^{[L]}/1$ system (deterministic arrivals, bulk Markov departures). So we can determine the dynamics of the relay's estimate. Let $z_n \in \{0, 1, \cdots\}$ denote the queue length at time n. and $P_{i,j}$ denote the one-step transition probability,

$$P_{i,j} = \mathbb{P}(z_{n+1} = j \mid z_n = i),$$

and let $\pi_z(i)$ denote the stationary distribution,

$$\pi_z(i) = \lim_{n\to\infty} \mathbb{P}(z_n = i),$$

if one exists. Using this notation we have,

$$\begin{aligned} \mathbb{P}\big(r_n(n+d) \neq s_n\big) &= \mathbb{P}\{\text{finding more than } d \text{ elements in queue at time } n+d\} \\ &= \sum_{i=d}^{\infty} \pi_z(i). \end{aligned} \tag{5.2}$$

Consequently, to conclude the analysis we need to calculate $\pi_z(i)$. From the description of the queue above, we determine the one-step transition probabilities,

$$\begin{aligned} P_{00} &= 1-\varepsilon \\ P_{i,0} &= 1-\varepsilon,\ i = 1, \cdots, L-1 \\ P_{i,i-L+1} &= 1-\varepsilon,\ i \geq L \\ P_{i,i+1} &= \varepsilon,\ i \geq 0, \end{aligned} \tag{5.3}$$

and the 'generic' balance equation,

$$\begin{aligned} \pi_z(i) &= P_{i+L-1,i}\pi_z(i+L) + P_{i-1,i}\pi_z(i-1) \\ &= (1-\varepsilon)\pi_z(i+L) + \varepsilon\pi_z(i-1),\ i \geq L. \end{aligned} \tag{5.4}$$

It is clear now that $\pi_z(i)$ will have a solution of the form γ^i up to a coefficient to be determined by the balance equations for $\pi_z(i)$, $i = 1, 2, \cdots, L-1$. Using (5.2) and (5.4) we obtain,

Theorem 5.1.1 *Let the L-bit erasure channel have erasure probability ε such that*

$$(1-\varepsilon)\gamma^{L+1} - \gamma + \varepsilon = 0 \tag{5.5}$$

has a solution $|\gamma'| < 1$. Then for any time $n \geq 0$ delay $d \geq 0$, the probability of estimation error for the relay is upper bounded as,

$$\mathbb{P}\big(r_n(n+d) \neq s_n\big) \leq \beta' 2^{-d\log_2 \frac{1}{\gamma'}} \tag{5.6}$$

for some strictly positive constant $\beta' > 0$ that depends on ε.

Remark 5.1.1 *If equation (5.5) has a solution γ' in the unit circle then it has at most one such solution in the unit circle. For a fixed ε, we have that $\gamma' \to \varepsilon$ as $L \to \infty$.*

Remark 5.1.2 *For $L = 2$ we have a unique solution $|\gamma'| < 1$ when $\varepsilon < 1/2$. This solution is given by*

$$\gamma' = \frac{\varepsilon}{1-\varepsilon}.$$

5.2 Relaying the estimate

We now have a simple model for the first part of the system consisting of the source, the L-bit Erasure Channel and the estimator at the relay. The question that remains to be answered is how the intermediate estimate will be transmitted to the true destination. Many solutions are possible. In this section we describe a general solution that is simple and performs well.

Let p' and q be two integers to be determined later. We define $p = \log_2 \sum_{i=0}^{p'} 2^i$ and pick an 'off-the-shelf' rate-p/q anytime code for the discrete-time memoryless channel that lies between the relay and the true destination. We will not worry about the specifics of the code; only that the code can emit one of $\{1, 2, \cdots, 2^{p'}\}$ symbols every q time steps.

The relay maintains an infinite binary queue in conjunction with the source's queue. Every q time steps the relay selects and removes up to the first $0 \leq i \leq p'$ entries of its own queue and emits i true information bits plus $p' - i$ 'dummy' symbols. So every q time steps the relay prepares a packet that can take up to $\sum_{i=0}^{p'} 2^i$ values and hands it to the 'off-the-shelf' anytime encoder, which then ships it off to the true destination. On the other hand, at each time step, the relay receives up to L true source bits. It simply places these bits at the end of the queue.

The source's and relay's queues form a series queuing system. In the language of queuing systems, the source's queue is a $D/M^{[L]}/1$ (deterministic arrival, bulk departure) and the relay's queue is one of deterministic departures. The arrivals to the relay's queue are exactly equal to the departures of the source's queue. We have illustrated the situation in Figure 5.2.

Let $\pi_{z,y}(i,j)$ denote the stationary distribution of the series queuing system. Suppose for the moment that we have calculated $\pi_{z,y}(i,j)$. Would this help us in achieving our ultimate objective, to determine $\alpha, \beta > 0$ in (5.1)? We attempt to answer this question in the following theorem where the notation $C^{\mathcal{A}}(\alpha)$ denotes the α-anytime capacity of the link between the relay and the true destination.

Remark 5.2.1 *Suppose $C^{\mathcal{A}}(\alpha) \leq 1$ and the L-bit erasure channel has a non-zero probability $\varepsilon > 0$ of erasures. Then, by definition of the Anytime capacity, there is no $\alpha, \beta > 0$ for which,*

$$\mathbb{P}\big(\tilde{s}_n(n+d) \neq s_n\big) \leq \beta\, 2^{-\alpha d},\; n \geq 0, d \geq 0.$$

Theorem 5.2.1 *Pick p' and q and pick an off-the-shelf rate $R = \frac{\log_2 \sum_{i=0}^{p'} 2^i}{q}$ anytime encoder such that $R > 1$. Then, if,*

1. *there exists some α' for which $C^{\mathcal{A}}(\alpha') > R$,*
2. *and there exists a stationary distribution $\pi_{z,y}(i,j)$ for the joint statistics of the source's and relay's queues,*

we can find an $\delta > 0$ for which

$$\mathbb{P}\big(\tilde{s}_n(n+d) \neq s_n\big) \leq \beta(\delta) \inf_{\alpha: C^{\mathcal{A}}(\alpha) \geq R+\delta} \inf_{0 \leq \rho \leq 1} \big[2^{-\alpha(1-\rho)d} + \sum_{i+j \geq \rho d} \pi_{z,y}(i,j)\big].$$

Consequently, we can find some $\alpha, \beta > 0$ that satisfy (5.1).

Proof: Fix $0 \leq \rho \leq 1$. Then the probability of the error event $\chi = \{\tilde{s}_n(n+d) \neq s_n\}$ is upper bounded by,

$$\begin{aligned} \mathbb{P}\big(\chi\big) \quad \leq \quad & \mathbb{P}\big(\chi \mid \text{relay has emitted bit } s_n \text{ by time } n+\rho d\big) \\ & + \mathbb{P}\big(\text{relay has not emitted bit } s_n \text{ by time } n+\rho d\big) \end{aligned} \tag{5.7}$$

By the definition of anytime capacity,

$$\mathbb{P}\big(\chi \mid \text{relay has emitted bit } s_n \text{ by time } n+\rho d\big) \leq \beta' 2^{-\alpha d}$$

for any α such that $C^{\mathcal{A}}(\alpha) > R$. The coefficient β' depends on how close $C^{\mathcal{A}}(\alpha)$ is to the transmission rate R. To lose the dependence we insist on the condition $C^{\mathcal{A}}(\alpha) \geq R + \delta$ so that

$$\mathbb{P}\big(\chi \mid \text{relay has emitted bit } s_n \text{ by time } n+\rho d\big) \leq \beta(\delta) \inf_{\alpha: C^{\mathcal{A}}(\alpha) \geq R+\delta} 2^{-\alpha d} \tag{5.8}$$

for some finite constant $\beta(\delta) < \infty$.

Finally, the relay has not emitted bit s_n by time $n + \rho d$ if we find more than a total of ρd entries in the source's queue plus the relay's queue. This is nothing but,

$$\mathbb{P}\big(\text{relay has not emitted bit } s_n \text{ by time } kn+\rho d\big) = \sum_{i+j \geq \lfloor \rho d \rfloor} \pi_{z,y}(i,j). \tag{5.9}$$

Using (5.9) and (5.8) in (5.7) and taking the infremum over all $0 \leq \rho \leq 1$ gives the desired result. □

This result has an appealing aspect. The stationary distribution π is independent of the second discrete-time memoryless channel. It is a function only of the erasure probability ε of the erasure channel and the rate R at which the relay emits bits. On the other hand, the anytime exponent is independent of the stationary distribution. It depends only on the rate R. So the relaying problem is reduced to picking the optimum rate R. Picking a small rate increases the performance of the anytime encoder but degrades the performance of the queues. Picking a larger rate improves the performance of the queues but degrades the performance of the anytime encoder.

5.3 Analysis of the queuing system

Before we stated our theorems we made the assumption that we could calculate π. It turns out that the problem has enough structure that we can find a closed form solution up to the calculation of coefficients. Here are the calculations.

Recall that z_n denotes the state of the source's queue at time n. Denote the relay's state by y_n. Recall also that $\pi_{z,y}(i,j)$ is the stationary distribution of $P_n(z_n = i, y_n = j)$. As we did in the previous section, we determine π by solving for the balance equations. By direct observation of the one-step transition probabilities (5.3) and for sufficiently large i, j,

$$\begin{aligned}
\pi_{z,y}(i,j) &= \pi_{z,y}\big(i-p+qL,\ j+q(L-1)\big)P^{(q)}_{j+q(L-1),\,j} \qquad (5.10)\\
&= \pi_{z,y}\big(i-p+(q-1)L,\ j+(q-1)(L-1)-1\big)P^{(q)}_{j+(q-1)(L-1)-1,\,j}\\
&= \pi_{z,y}\big(i-p+(q-2)L,\ j+(q-2)(L-1)-2\big)P^{(q)}_{j+(q-2)(L-1)-2,\,j}\\
&\ \vdots\\
&= \pi_{z,y}\big(i-p,\ j-q\big)P^{(q)}_{j-q,\,j}\\
&= \sum_{k=0}^{q} \pi_{z,y}\big(i-p+(q-k)L,\ j+(q-k)(L-1)-k\big)P^{(q)}_{j+(q-k)(L-1)-k,\,j}
\end{aligned}$$

where $P^{(q)}_{j_1,j_2}$ denotes the q-step transition probability of the state of the source's queue:

$$\mathbb{P}(z_{n+q} = j_2, x_n = j_1).$$

That there is such a compact representation for the balance equations is not a surprise. This is consequence of the fact that the series queuing system consists of one queue driving the other. This is why the balance equations are defined only in terms of the q-step transition probability of the driving queue.

To conclude the calculations we need two more things; the q-step transition probabilities and a guess for the solution. Again by direct observation of (5.3) we see that for large j there are exactly $\binom{q}{k}$ ways to go from state $z_n = j + (q-k)(L-1) - k$ to state $z_{n+q} = j$ in q steps. Each of these choices involve k 'up' transitions and $q-k$ 'down' transitions. So we have,

$$P^{(q)}_{j+(q-k)(L-1)-k,\,j} = \binom{q}{k}(1-\varepsilon)^k\varepsilon^{q-k}.$$

Plug this back into (5.10) and write,

$$\pi_{z,y}(i,j) = \sum_{k=0}^{q}\binom{q}{k}\pi_{z,y}\big(i-p+(q-k)L,\ j+(q-k)(L-1)-k\big)\,(1-\varepsilon)^k\varepsilon^{q-k}. \qquad (5.11)$$

We guess that the solution to this last recurrence equation is of the form,

$$\begin{aligned}
\pi_{z,y}(i,j) &= \pi_y(i)\pi_z(j)\\
&= \eta^i\gamma^j
\end{aligned}$$

where π_y is the stationary distribution of $\{y_n\}$ and π_z is the stationary distribution of $\{z_n\}$. Plug this guess into (5.11) and simplify,

$$\begin{aligned} 1 &= \eta^{qL-p}\gamma^{q(L-1)}\left(\sum_{k=0}^{q}\binom{q}{k}\eta^{-kL}\gamma^{-k(L-1)}(1-\varepsilon)^k\varepsilon^{q-k}\right) \\ &= \eta^{qL-p}\gamma^{q(L-1)}\left(\eta^{-L}\gamma^{-(L-1)}(1-\varepsilon)+\varepsilon\right)^q \\ &= \eta^{\frac{qL-p}{q}}\gamma^{(L-1)}\left(\eta^{-L}\gamma^{-(L-1)}(1-\varepsilon)+\varepsilon\right) \end{aligned}$$

and with one final step,

$$\eta^R = (1-\varepsilon) + \left(\gamma^{L-1}\varepsilon\right)\eta^L, \tag{5.12}$$

where $R = p/q$ is the rate at which the relay emits bits to the final destination. Finally, for appropriate values of L, R, ε if we can find $|\gamma| < 1$ in (5.5) and $|\eta| < 1$ in (5.12) then these must be the unique solutions of (5.5) and (5.12) respectively. This concludes our analysis and treatment of the subject. The reader is encouraged to compare the structure of the characteristic equation (5.12) with the representation in (3.13) of a Binary Erasure Channel.

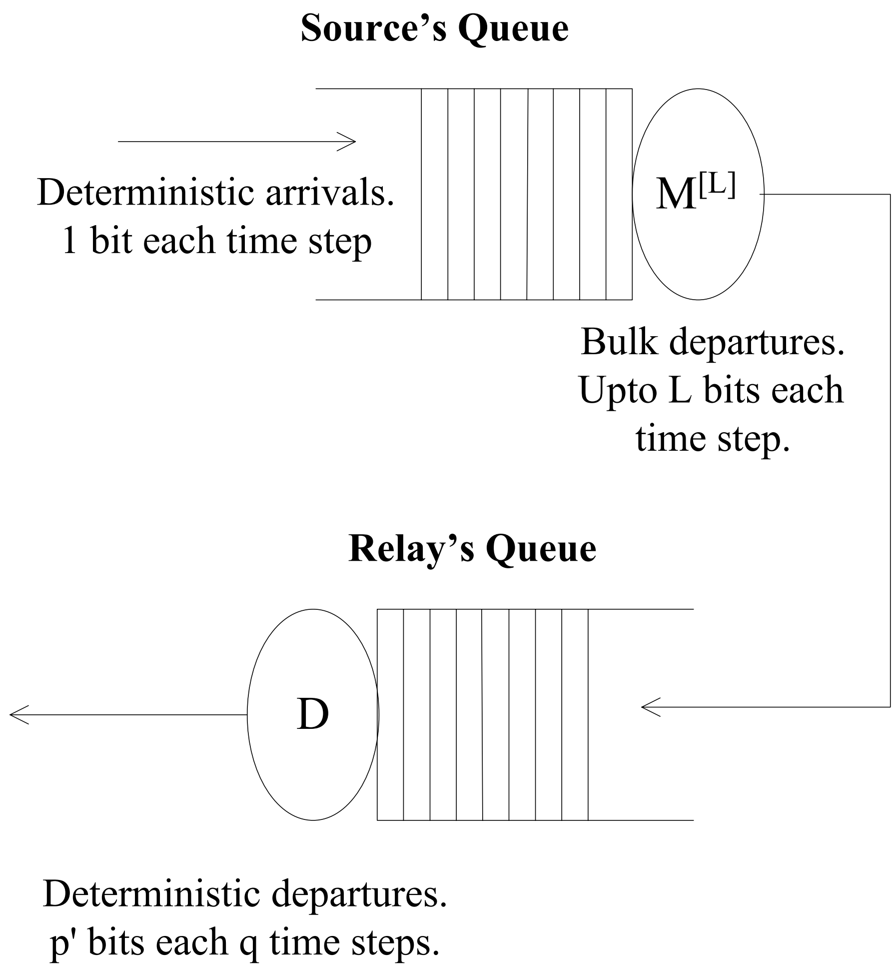

Figure 5.2: Illustration of the series queuing system for the source and relay.

Chapter 6

Recurrence and stability in infinite Markovian jump systems

In one of our many initial attempts to model anytime codes we considered a systems approach based on jump systems. Ultimately we discovered that there were better methodologies but we nevertheless pursued our investigation for the sake of completeness. In this chapter we present what we've learned about jump systems and a set of stability criteria that may be viewed as a simpler and finite alternative to the traditional criteria which requires the solution of an infinite set of balance equations. The material is presented in a self-contained fashion and may be read independently of the other chapters.

Discrete-time jump linear systems are useful models for many systems that exhibit abrupt changes in their dynamics. In the physical world, the change in dynamics is due to the occurrence of discrete events at random times. For example, the time and number of customers that arrive at a checkout-stand or the time and amount of change in the price of stock. Intuitively, there are two sources of randomness; the time at which an event occurs and which of several possible events have occurred.

The system is said to be Markovian if the times between events are independent (memoryless) and the event which will occur next depends only on the last occurred event. Formally, let $\theta(k)$, $k \geq 1$, be a Markov chain that takes values in the infinite set $S = \{0, 1, 2, \cdots\}$. To each $i \in S$ assign an n-by-n matrix A_i. Based on this Markov chain and assignment, the Markovian jump linear system is,

$$x_{k+1} = A_{\theta(k)} x_k + v_k,\ k \geq 0, \tag{6.1}$$

in which $x_k \in \mathbb{R}^n$ is the state variable and $v_k \in \mathbb{R}^n$ is a disturbance term.

We are interested in determining the asymptotic stability of the state x_k as k tends to ∞. Because the system can exhibit an infinite number of dynamics, the general problem is difficult. An exact characterization of stability requires solving an infinite number of coupled balance equations.

In contrast, there are many physical systems that are typically simple even though the system may potentially exhibit an infinite number of complex dynamics at random times. This is because a lot of events are rare. In this article, we attempt to find a simple characterization of such systems.

To illustrate the idea, let's solve the problem for a simple example. Consider the Markov chain, $\theta(k)$, illustrated in Figure 6.1. For simplicity there is no disturbance, $v_k = 0$, $k \geq 0$. Take $A_{(0,0)} = 1/2$ and for all $1 \leq j + 1 \leq i$, $A_{(i,j)} = 2$. So we have a system that is seemingly unstable. However, we will show that for $q_i = C\lambda^{-i}$, $i \geq 0$, $\lambda > 16/3$, the Markov chain visits state $(0, 0)$ so frequently that the system 'typically' behaves like $x_{k+1} = 1/2\, x_k$ and is consequently stable.

Here are the calculations. We define the quantities,

$$\begin{aligned} T^n &= \min\{k > 0 \mid \theta(T^1 + \cdots + T^{n-1} + k) = (0,0)\} \\ n_k &= \min\{n \mid T^1 + \cdots + T^n \geq k\} \end{aligned}$$

The random variable (r.v.) T^n is the time between the $(n-1)$st and nth return to state $(0,0)$. The r.v. n_k is the number of times $(0,0)$ has been visited by time $k \geq 1$. By the Markov property, T^n are independent and identically distributed (i.i.d.). By symmetry in the problem, $P(T^1 = i+1) = q_i$, $i \geq 0$. Using this notation, we write

$$x_k = 2^m \times 2^{T^{n_k-1}-2} \times \cdots \times 2^{T^1-2} \times x_0,$$

in which $0 \leq m < T^{n_k}$. By taking the square, expectations and using the independence of the return times,

$$\begin{aligned} &E\big(\|x_k\|^2 \,\big|\, \theta(0) = (0,0), x_0\,\big) \\ &\qquad = E2^{2m} \times E\big(E2^{2(T^1-2)}\big)^{n_k-1} \times \|x_0\|^2 \\ &\qquad \leq E4^{T^1} \times E\big(E4^{T^1-2}\big)^{n_k-1} \times \|x_0\|^2 \\ &\qquad = \frac{4\lambda}{\lambda-4} \times E\bigg(\frac{\lambda}{4\lambda-16}\bigg)^{n_k-1} \times \|x_0\|^2. \end{aligned}$$

Furthermore, by definition of n_k and by Markov's inequality,

$$P(n_k < n) = P(T^1 + \cdots + T^n \geq k) \leq \frac{E\big(4^{T^1}\big)^n}{4^k}.$$

So, taking $n = \lfloor k/\mu \rfloor$, for an appropriately chosen $\mu > 0$, it is easy to see that $P(n_k < \lfloor k/\mu \rfloor) \to 0$ exponentially with k. On the other hand, $\frac{\lambda}{4\lambda-16} < 1$ because $\lambda > 16/3$. So we have the following simple inequality,

$$E\bigg(\frac{\lambda}{4\lambda-16}\bigg)^{n_k} \leq P\big(n_k < \lfloor k/\mu \rfloor\big) + \bigg(\frac{\lambda}{4\lambda-16}\bigg)^{\lfloor k/\mu \rfloor} P\big(n_k \geq \lfloor k/\mu \rfloor\big).$$

Evidently, this expectation, and hence, $E\big(\|x_k\|^2 \,\big|\, \theta(0) = (0,0), x_0\,\big)$ tends to zero exponentially as k tends to infinity.

So what have our calculations revealed? We started with a system that was unstable all the time except when the underlying Markov chain visited state $(0,0)$. Regardless, we showed that, by recurrence, the system was stable. We believe that the preceding example illustrates a basic fact about many real-world jump systems. That is, the system may potentially exhibit an infinite number of dynamics. But the system typically exhibits only a finite set of 'good' dynamics. Deviations to 'bad' dynamics for long periods are unlikely. Consequently, the system, although seemingly complex, behaves like a simple and finite system. The main results of this chapter (Theorems 6.2.1 and 6.2.2) models this effect.

The general approach for determining stability of (6.1) is to look at the Lyapunov exponents of the system. Costa and Fragoso [7] showed that stochastic stability (c.f. Definition 6.1.1) is equivalent to the existence of a solution to an infinite set of coupled Lyapunov equations. The ideas presented here should be seen as an alternative way to determine stability when the system (6.1) is sufficiently recurrent. Our idea is similar to the notion of 'regularity' used by Fang and Loparo in [11] to find computationally efficient methods for determining stability of discrete-time jump linear systems. A similar idea, under a different name, can also be found in Bucklew [4].

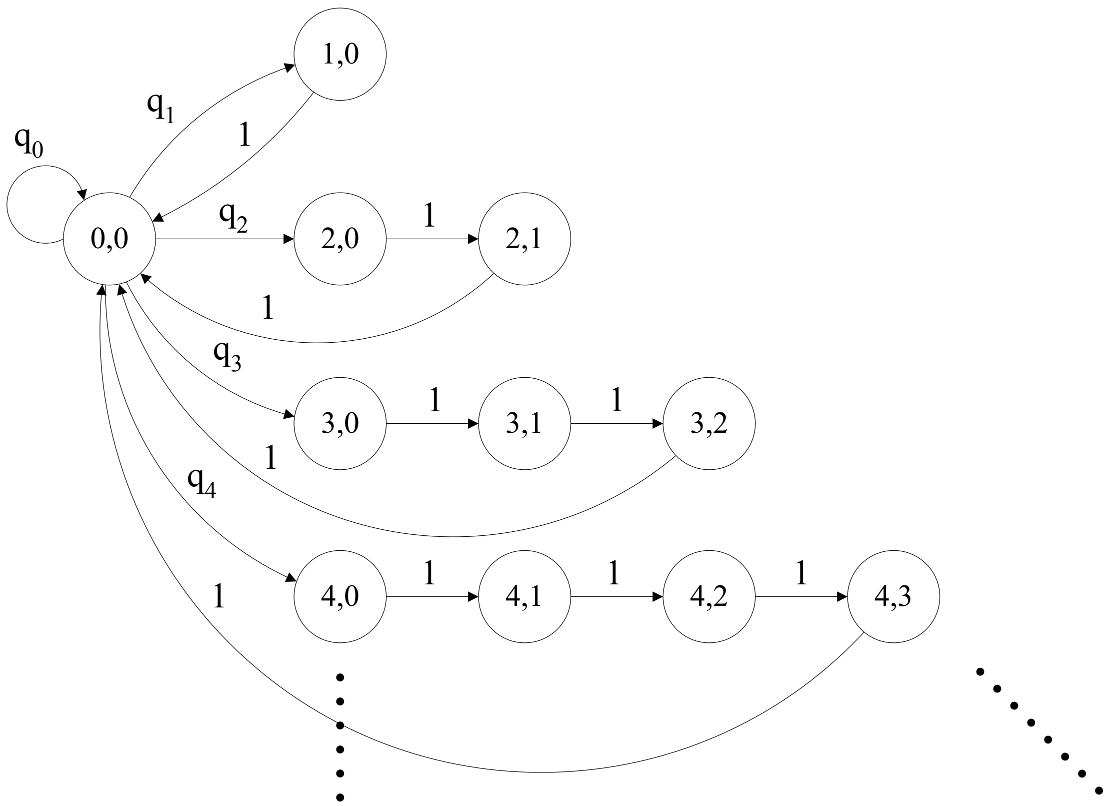

Figure 6.1: A simple infinite Markov jump linear system. The state $\theta(k)$ of the chain takes values in $S = \{(0,0),(1,0),(2,0),(2,1),(3,0),(3,1),(3,2),\cdots\}$. The transition probabilities are $q_i = C\lambda^{-i}$, $i \geq 0$, $\lambda > 16/3$ and $C = 1/\sum_{i=0}^{\infty} \lambda^{-i}$. The coefficient process is $A_{(0,0)} = 1/2$ and for all $1 \leq j+1 \leq i$, $A_{(i,j)} = 2$.

6.1 Preliminaries

The concepts of stability for infinite jump systems is somewhat complex and fragile. In this section, we present some notation and preliminary results to familiarize the reader with the intricacies.

Recall that the infinite Markov jump linear system is,

$$x_{k+1} = A_{\theta(k)}x_k + v_k,\ k \geq 0, \tag{6.2}$$

where $x_k \in \mathbb{R}^n$ is the state variable, $\|v_k\| \leq V$ is an unknown but bounded disturbance term, A_i, $i \in S = \{0, 1, 2, \cdots\}$ is a collection of n-by-n matrices and $\theta(k) \in S$ is a countable state Markov chain. The Markov chain is assumed to be time-homogeneous, irreducible and aperiodic.

The uncertainty in the system is due to (i) the disturbance v_k, (ii) the initial state x_0 and (iii) the Markov chain $\theta(k)$. We do not assume a probabilistic structure on the disturbance. For simplicity we also take x_0 as a fixed non-random vector. Then the only source of randomness is the Markov chain: the distribution $p = (p_0, p_1, \cdots)$ of the initial state $\theta(0)$ and the transition probabilities $P = (p_{ij})_{S\times S}$. Denote the underlying probability space by $(\Omega, \mathcal{F}, \mathcal{P})$. The solution process $x_k = x(k, x_0, \omega, v^k)$, where $v^k = (v_1, \cdots, v_k)$, is then a random process on the probability space Ω. We will use the explicit notation $x(k, x_0, \omega, v^k)$ only when necessary, otherwise we will simply write x_k.

The preceding structure leads us to the following definitions of stability which are similar to those found in [12].

Definition 6.1.1 *Let Φ be a set of probability distributions (for the initial state $\theta(0)$) on S. Then, for $\ell > 0$, system (6.2) is said to be,*

a. *Asymptotically ℓ-stable w.r.t. Φ if for any $x_0 \in \mathbb{R}^n$ and distribution $p \in \Phi$,*

$$\lim_{k\to\infty} E\|x(k, x_0, \omega, 0)\|^\ell = 0.$$

b. *Stochastically ℓ-stable w.r.t. Φ if for any $x_0 \in \mathbb{R}^n$ and distribution $p \in \Phi$,*

$$\sum_{k=0}^{\infty} E\|x(k, x_0, \omega, 0)\|^\ell < \infty.$$

c. *Exponentially ℓ-stable w.r.t. Φ if for any $x_0 \in \mathbb{R}^n$ and distribution $p \in \Phi$ of $\theta(0)$, there exists $K, \tau > 0$ such that,*

$$E\|x(k, x_0, \omega, 0)\|^\ell \leq K\|x_0\|^\ell 2^{-\tau k}.$$

d. *Asymptotically ℓ-bounded w.r.t. Φ if for any $x_0 \in \mathbb{R}^n$ and distribution $p \in \Phi$,*

$$\lim_{k\to\infty} \sup_{v^k} E\|x(k, x_0, \omega, v^k)\|^\ell < \infty.$$

The cases for $\ell = 1$ and $\ell = 2$ are also called mean stable and mean-square stable respectively.

One might expect that some of the concepts of stability are related. Indeed, in Definition 6.1.1, (b) implies (a) and (c) implies both (a) and (b). Furthermore, if the Markov chain has only a finite number of states, then (a) also implies (b), so that (a) and (b) are equivalent.

One might also expect that exponential stability implies asymptotic boundedness. However, this conjecture is false in the absence of additional assumptions. To see why, consider a simple example with $p_{i(i+1)} = 1$, $A_0 = 0$ and for all $i > 0$, $A_i = 2$. Thus, if we take $\Phi = \{(1, 0, 0, \cdots)\}$ then the system is asymptotically, exponentially and stochastically stable, but not asymptotically bounded. On the other hand, suppose that the Markov chain admits a stationary distribution $\pi = (\pi_0, \pi_1, \cdots)$ and the chain is stationary, then exponential stability implies asymptotic boundedness. To show this, write,

$$x(k, x_0 = 0, \omega, v^k) = \sum_{j=1}^{k} A_{\theta(k,\omega)} \cdots A_{\theta(j+1,\omega)} v_j.$$

Then,

$$\begin{aligned} E\|x(k, x_0 = 0, \omega, v^k)\| &\leq \textstyle\sum_{j=0}^{k} E\|A_{\theta(k,\omega)} \cdots A_{\theta(j+1,\omega)}\|V \\ &= \textstyle\sum_{j=0}^{k} E\|A_{\theta(j,\omega)} \cdots A_{\theta(1,\omega)}\|V \\ &= \textstyle\sum_{j=0}^{k} E\|x(j, \|x_0\| = V, \omega, 0)\| \end{aligned} \tag{6.3}$$

in which the second equality follows from the stationarity of $\theta(k)$. But the system is exponentially stable, so there exists $\lambda, \beta > 0$ such that $E\|x(k, x_0 = 0, \omega, 0)\| \leq \lambda V 2^{-\beta k}$. Thus, the last series in (6.3) is convergent so that the system is asymptotically bounded. This simple derivation leads to the following result.

Corollary 6.1.1 *Let Φ be a set of probability distributions on $S = \{0, 1, 2, \cdots\}$ that are P-invariant. That is, if $p = (p_0, p_1, \cdots) \in \Phi$ then $p' = ({p'}_0, {p'}_1, \cdots) \in \Phi$ where ${p'}_i = \sum_{j=0}^{\infty} p_{ji} p_j$. If there exists constants $K, \lambda > 0$ such that for all x_0 and $p \in \Phi$,*

$$E\|x(k, x_0, \omega, v^k = (0, \cdots, 0)\| \leq K\|x_0\| 2^{-\lambda k},$$

the system is asymptotically bounded.

Let's conclude our discussion of stability with the following question. Consider system (6.2) and the largest set of distributions Φ for which the system is stable with respect to Φ. Then, how big is Φ? In our experience, we've seen that it can be quite small. For example, recall the simple system we looked at in the introduction. If the distribution $P\big(\theta(0) = (i, 0)\big)$, $i \geq 1$ is 'heavy-tailed', then the system is unstable. For example, let,

$$P\big(\theta(0) = (i, 0)\big) = Di^{-2},\ D = 1/\sum_{i=1}^{\infty} i^{-2},$$

Then, the instability of the system follows by calculating the following expectation,

$$\begin{aligned} E\big(\|x_k\|^2\big) &\geq 2^{2k}\|x_0\|^2 P\{\text{the chain was in a long branch at time } k = 0\} \\ &\geq 2^{2k}\|x_0\|^2 \sum_{i=k}^{\infty} Di^{-2} \\ &\geq 2^{2k}\|x_0\|^2 \int_{k+1}^{\infty} Dy^{-2}\mathrm{d}y \\ &= \frac{2^{2k}\|x_0\|^2 D}{(k+1)} \end{aligned}$$

and letting k tend to infinity.

6.2 Main results

Our objective is to characterize the stability of jump systems that have the following property: the system may potentially exhibit an infinite number of dynamics but the system typically exhibits only a finite set of 'good' dynamics and deviations to other 'bad' dynamics for long periods are not likely. Evidently, our first task is to describe more precisely what we mean by this.

For any state $i \in S$ of the underlying Markov chain, we define the following random variables,

$$\begin{aligned} R(i) &= \min\{k \geq 1 \mid \theta(k) = i\}, \\ T^1(i) &= \min\{k \geq 1 \mid \theta(Y(i) + k) = i\}, \end{aligned}$$

and for $n > 1$, recursively,

$$T^n(i) = \min\{k \geq 1 \mid \theta(R(i) + T^1(i) + \cdots T^{n-1}(i) + k) = i\}.$$

So $R(i)$ is the time to reach state i for the first time and $T^n(i)$ is the time between the $(n-1)$st and nth return to i. Based on the return-times write the coefficient process $\{A_{\theta(k)}\}$ as follows,

$$\begin{aligned} A_k^R(i) &= A_{\theta(k)}, \;\; k = 1, \cdots, R(i), \\ A_k^1(i) &= A_{\theta(k+R(i))}, \; k = 1, \cdots, T^1(i), \end{aligned}$$

and for $n > 1$,

$$A_k^n(i) = A_{\theta(k+R(i)+T^1(i)+T^2(i)+\cdots+T^{n-1}(i))}, \; k = 1, \cdots, T^n(i).$$

Finally, denote,

$$\begin{aligned} T(i) &= T^1(i), \\ \Gamma(i) &= \prod_{k=1}^{T^1(i)} A_k^1(i), \\ \Delta(i) &= \prod_{k=1}^{T^1(i)} \max\left(\|A_k^1(i)\|, 1\right) \end{aligned} \tag{6.4}$$

Lemma 6.2.1 *Consider system (6.2) with $v_k = 0$. Let $\ell > 0$ and $i \in S$ be a state of the underlying Markov chain, $\theta(k)$, for which $E\big(\Delta(i)^\ell\big)$ is finite. Then, there exists a finite random variable $\gamma\big(\theta(0)\big)$ such that,*

$$E\big(\|x_k\|^\ell \mid \theta(0)\big) \leq \gamma\big(\theta(0)\big)\, E\Big(E\|\Gamma(i)\|^{\ell n_k(i)}\Big)$$

in which $n_k(i) = \min\{n \geq 1 \mid T^1(i) + \cdots + T^n(i) \geq k\}$.

Remark 6.2.1 *The random variable $\gamma(\theta(0))$ is a function of two things: $\Delta(i)$ and the time $R(i)$ at which the Markov chain reaches i for the first time. The details are shown in the proof of the Lemma in the appendix.*

In this lemma, we have attempted to capture a notion of typicality in infinite jump systems. We pick a state $i \in S$ of the underlying Markov chain. Each time the chain visits i, the expected magnitude of the state $E\|x_k\|^\ell$ grows (if $E\|\Gamma(i)\|^\ell > 1$) or shrinks (if $E\|\Gamma(i)\|^\ell < 1$) by a factor of $E\|\Gamma(i)\|^\ell$. If the chain is positive recurrent, then the chain visits i infinitely often. So, the number $E\|\Gamma(i)\|^\ell$ characterizes the typical behavior of the system. Intuitively, the behavior is good when $E\|\Gamma(i)\|^\ell < 1$ and the chain visits i frequently.

Let's come back to stability. With respect to the preceding characterization, are there any general conditions for stability?

Theorem 6.2.1 *Consider system (6.2) with $v_k = 0$. Let $\ell > 0$ and $i \in S$ be a state of the underlying Markov chain, $\theta(k)$, for which $E\big(\Delta(i)^\ell\big)$ is finite and $E\|\Gamma(i)\|^\ell < 1$. Consider the set $\Phi = \{(1,0,0,\cdots),(0,1,0,\cdots),\cdots\}$ of distributions of $\theta(0)$. Then,*

a. *If $ET(i) < \infty$ the system is asymptotically ℓ-stable w.r.t. Φ,*

b. *If $E\big(T(i)^2\big) < \infty$ the system is stochastically ℓ-stable w.r.t. Φ,*

c. *If $\exists \tau > 0$ s.t. $Ee^{\tau T(i)} < \infty$, the system is exponentially ℓ-stable w.r.t. Φ.*

Proof: The system is asymptotically ℓ-stable w.r.t. the given Φ if and only if $E\|x_k\|^\ell \to 0$ for each $\phi \in \Phi$. Due to the special structure of Φ, this is equivalent to $E\big(\|x_k\|^\ell \,\big|\, \theta(0)\big) \to 0$. Similarly for stochastic and exponential stability.

By definition of $n_k(i)$ (Lemma 6.2.1) we have,

$$P(n_k(i) < n) = P(T^1(i) + \cdots + T^n(i) \geq k).$$

On the other hand, for any $\lambda < 1$, we have,

$$\begin{aligned} E(\lambda^{n_k(i)}) &\leq P(n_k(i) < n) + \lambda^n P(n_k(i) \geq n) \\ &\leq P(T^1(i) + \cdots + T^n(i) \geq k) + \lambda^n. \end{aligned} \tag{6.5}$$

Using this inequality we will now show the desired results using classical techniques from the theory of large deviations.

Case a. Let $n = \lfloor \sqrt{k}/\mu \rfloor$ for some $\mu > ET(i)$. By Markov's inequality,

$$P(T^1(i) + \cdots + T^n(i) \geq k) \leq \frac{nET(i)}{k} \leq \frac{1}{\sqrt{k}}.$$

So $P(T^1(i) + \cdots + T^n(i) \geq k) \to 0$ as $k \to \infty$. We also have that $\lambda^{\lfloor \sqrt{k}/\mu \rfloor} \to 0$ as $k \to \infty$. Thus, by (6.5), $E(\lambda^{n_k(i)}) \to 0$ as $k \to \infty$. So by taking $\lambda = E\|\Gamma(i)\|^\ell$ in Lemma 6.2.1, $E\big(\|x_k\|^\ell \,\big|\, \theta(0)\big) \to 0$ as $k \to \infty$. In conclusion, the system is asymptotically ℓ-stable w.r.t. $\Phi = \{(1,0,0,\cdots),(0,1,0,\cdots),\cdots\}$.

Case b. Let $n = \lfloor k^{1/4}/\mu \rfloor$ for some $\mu > ET(i)$. Again by Markov's inequality,

$$P(T^1(i) + \cdots + T^n(i) \geq k) \leq \frac{n^2 ET(i)^2}{k} \leq \frac{1}{k^{3/2}}.$$

So the series $\sum_{t=1}^{k} P(T^1(i) + \cdots + T^n(i) \geq t)$ is convergent. On the other hand, the series $\sum_{t=1}^{\infty} \lambda^{\lfloor k^{1/4}/\mu \rfloor}$ is also convergent. This follows by the integral test of convergence; that is,

$$\sum_{t=1}^{\infty} \lambda^{k^{1/4}/\mu} < \infty \Leftrightarrow \int_1^\infty \lambda^{t^{1/4}} dt < \infty.$$

But, by the change of variables $y = t^{1/4}$, $\int_1^\infty \lambda^{t^{1/4}} dt = \int_1^\infty 4y^3 \lambda^y dy < \infty$. Using the convergence of the preceding two series, we have by (6.5),

$$\lim_{k\to\infty} \sum_{t=1}^{k} E(\lambda^{n_t(i)}) < \infty.$$

The desired result follows once again by taking $\lambda = E\|\Gamma(i)\|^\ell$ in Lemma 6.2.1 and observing the definition of stochastic stability.

Case c. Let $n = \lfloor k/\mu \rfloor$ for some $\mu > ET^1(i)$. For the last time, by Markov's inequality,

$$\begin{aligned} P(T^1(i) + \cdots + T^n(i) \geq k) &\leq E(e^{\tau T(i)})^n / \mathrm{e}^{\tau k} \\ &= \mathrm{e}^{-k(\tau - \ln E(e^{\tau T(i)})/\mu)},\ \tau > 0. \end{aligned}$$

A known result (Durrett, Chapter 1, Lemma 9.4) is that the exponent $\tau - \ln E(e^{\tau T(i)})/\mu > 0$ for some small τ provided that there exists $\tau > 0$ for which $Ee^{\tau T(i)} < \infty$. In conclusion, by (6.5), $E(\lambda^{n_t(i)}) \to 0$ exponentially as $k \to \infty$ and the desired result follows again by taking $\lambda = E\|\Gamma(i)\|^\ell$ in Lemma 6.2.1.□

This shows that there is a close connection between stability and recurrence. Intuitively, if the system is recurrent, it exhibits typical behavior. The preceding Theorem justifies this line of thinking. Deviations from the typical behavior are characterized by examining the frequency at which the underlying Markov chain visits state i.

We now present the final result of this chapter. It says that, by recurrence, exponential stability implies mean boundedness. This may seem obvious, but recall (c.f. Section 6.1) that the implication is not in general true.

Theorem 6.2.2 *Consider the system (6.2). Suppose one can find a state $i \in S$ of the underlying Markov chain for which $E\Delta(i)$ is finite and $E\|\Gamma(i)\| < 1$. If for some $\tau > 0$, $Ee^{\tau T(i)} < \infty$,*

$$\lim_{k\to\infty} \sup_{v_1,\ldots,v_k} E\big(\|x_k\| \mid \theta(0)\big) < \infty.$$

In other words, system (6.2) is asymptotically mean-bounded with respect to $\Phi = \{(1,0,0,\cdots),(0,1,0,\cdots),\cdots\}$.

Proof: At time k, the state of the system (6.2) is given by the random sum,

$$x_k = A_{\theta(k)} \cdots A_{\theta(1)}\, x_0 + \sum_{i=1}^{k} A_{\theta(k)} \cdots A_{\theta(i+1)}\, v_i. \tag{6.6}$$

Based on the state i of the underlying Markov chain, define,

$$z_{n,m}(i,z) = \prod_{j=1}^{T^n} A_j^n(i) \cdots \prod_{j=1}^{T^m} A_j^m(i) \times z,\ m \leq n.$$

Using this, we may decompose (6.6) into two parts: the 'zero-state' and the 'zero-input'. We write,

$$\begin{aligned} x_k =& A_p^{n_k(i)}(i) \cdots A_1^{n_k(i)}(i)\, z_{n_k(i)-1,0}(i, z_0) \\ &+ \sum_{j=1}^{k} A_p^{n_k(i)}(i) \cdots A_1^{n_k(i)}(i)\, z_{n_k(i)-1,n_j(i)}\big(i, A^{n_j(i)}_{T^{n_j(i)}}(i) \cdots A_q^{n_j(i)}(i)\, v_j\big) \end{aligned} \tag{6.7}$$

in which $n_k(i) = \min\{n \mid T^1(i) + \cdots + T^n(i) \geq k\}$ and $0 \leq p < T^{n_k(i)}$ and $0 \leq q < T^{n_j(i)}$.

By Theorem 6.2.1 we already know that the zero-input part of (6.7) tends to zero exponentially with k. So we only need to worry about the series in (6.7). Taking norms and expectation of the series and observing that $A_k^n(i)$ is independent of $z_{n,m}(i,z)$, $m \leq n$, we obtain,

$$\begin{aligned} & E\|\sum_{j=1}^{k} A_p^{n_k(i)}(i) \cdots A_1^{n_k(i)}(i)\, z_{n_k(i)-1,n_j(i)}\big(A_{T^{n_j(i)}}^{n_j(i)} \cdots A_q^{n_j(i)}\, v_j\big)\| \\ & \leq \sum_{j=1} E\big\|A_p^{n_k(i)}(i) \cdots A_1^{n_k(i)}(i)\big\| E\big\|z_{n_k(i)-1,n_j(i)}\big(A_{T^{n_j(i)}}^{n_j(i)} \cdots A_q^{n_j(i)}\, v_j\big)\big\| \\ & \leq \sum_{j=1} E\Delta(i) E\big\|z_{n_k(i)-1,n_j(i)}\big(A_{T^{n_j(i)}}^{n_j(i)} \cdots A_q^{n_j(i)}\, v_j\big)\big\|. \end{aligned} \tag{6.8}$$

On the other hand, by definition of $z_{n,m}(i,z)$, we have,

$$E\|z_{n_k(i)-1,n_j(i)}(z)\| \leq E\Delta(i)\, V\, E\big(E\|\Gamma^1\|\big)^{n_k(i)-1-n_j(i)}. \tag{6.9}$$

Now observe that $P(n_k(i) - 1 - n_j(i) < n) = P(T^1(i) + \cdots + T^{n+1}(i) \geq k - j)$ and for any $0 \leq \lambda < 1$

$$E\lambda^{n_k(i)-1-n_j(i)} \leq P(n_k(i) - 1 - n_j(i) < n) + \lambda^n P(n_k(i) - 1 - n_j(i) \geq n).$$

Similar to what we did in Theorem 6.2.1, let $n = \lfloor (k-j)/\mu \rfloor$ for some $\mu > ET(i)$ and use Markov's inequality to obtain,

$$P(T^1(i) + \cdots + T^{n+1}(i) \geq k - j) \leq e^{-(k-j)(\tau - \ln E(e^{\tau T(i)})/\mu)}, \tau > 0.$$

Finally, invoking the large deviations result (Durrett, Chapter 1, Lemma 9.4) once again, we have that the exponent $\tau - \ln E(e^{\tau T(i)})/\mu > 0$ for some small $\tau > 0$ and $\mu > ET(i)$. Consequently, $E\lambda^{n_k(i)-1-n_j(i)} \leq e^{\epsilon(k-j)}$ for some small $\epsilon > 0$. This means, however, that the series $\sum_{j=1}^{k} \big(E\Delta(i)\big)^2 V\, E\lambda^{n_k(i)-1-n_j}$ is convergent. The desired result follows by taking $\lambda = E\|\Gamma^1(i)\|$ and substituting (6.9) into (6.8). □

Proof of Lemma 6.2.1

Define a process, $z(i) = \{z_n(i)\}$, by sampling x at the return-times $T^n(i)$ according to $z_0(i) = x_{R(i)}$ and for $n \geq 1$, $z_n(i) = x_{R(i)+T^1(i)+\cdots+T^n(i)}$. This new process may be written recursively as,

$$z_n(i) = \left(\prod_{k=1}^{T^n(i)} A_k^n(i)\right) z_{n-1}(i). \tag{6.10}$$

The original system, x, is related to the sampled system, $z(i)$, by the relation,

$$x_k = \prod_{j=1}^{m} A_j^{n_k(i)}(i) \times z_{n_k(i)-1}(i),\ k \geq 0, \tag{6.11}$$

in which $0 \leq m < T^{n_k}(i)$.

By the Markov property, the blocks $\{A_1^n(i), \cdots, A_{T^n(i)}^n(i)\}$ are i.i.d. So, for $z_{n-1}(i) \neq 0$, the random variable $\|z_n(i)\|/\|z_{n-1}(i)\|$ is independent of $z_{n-1}(i)$. Denote $Z = \{z_n(i) \neq 0,\ 0 \leq n \leq n_k(i)-1\}$.

Using the independence and (6.10),

$$
\begin{aligned}
E\big(\|z_{n_k(i)-1}(i)\|^\ell\big) &= E\big(\|z_{n_k(i)-1}(i)\|^\ell \,\big|\, Z\big)P(Z) + E\big(\|z_{n_k(i)-1}(i)\|^\ell \,\big|\, Z^c\big)P(Z^c) \\
&= E\big(\|z_{n_k(i)-1}(i)\|^\ell \,\big|\, Z\big)P(Z) \\
&= EE\Bigg(\Big(\frac{\|z_{n_k(i)-1}(i)\|}{\|z_{n_k(i)-2}(i)\|}\cdots\frac{\|z_1(i)\|}{\|z_0(i)\|}\|z_0(i)\|\Big)^\ell \,\Bigg|\, n_k(i), Z\Bigg)P(Z) \\
&= E\Bigg(E\Big(\frac{\|z_{n_k(i)-1}(i)\|}{\|z_{n_k(i)-2}(i)\|} \,\Big|\, Z\Big)^\ell \cdots E\Big(\frac{\|z_1(i)\|}{\|z_0(i)\|} \,\Big|\, Z\Big)^\ell E\Big(\|z_0\|\Big)^\ell\Bigg)P(Z) \\
&\leq E\big(E\|\Gamma(i)\|^\ell\big)^{n_k(i)-1}E\big(\|z_0(i)\|^\ell\big)P(Z). \qquad (6.12)
\end{aligned}
$$

Finally, taking norms and expectation of (6.11) and using the independence of $A_m^{n_k(i)}$, $0 \leq m < T^{n_k}(i)$ and $z_{n_k(i)-1}$, we obtain,

$$
\begin{aligned}
E\big(\|x_k\|^\ell \,\big|\, \theta(0)\big) &= E\Big(\prod_{j=1}^{m} A_j^{n_k(i)}(i)\Big)^\ell \times E\big(\|z_{n_k(i)-1}(i)\|^\ell \,\big|\, \theta(0)\big) \\
&\leq E\big(\Delta(i)\big)^\ell \times E\big(\|z_{n_k(i)-1}(i)\|^\ell \,\big|\, \theta(0)\big). \qquad (6.13)
\end{aligned}
$$

The desired result follows by substituting (6.12) into (6.13) and observing that,

$$
E\big(\|z_0(i)\|^\ell \,\big|\, \theta(0)\big) = E\big(\|x_{R(i)}\|^\ell \,\big|\, \theta(0)\big),
$$

which, by the following claim, is a finite random variable. □

Claim: $E\big(\|x_{R(i)}\|^\ell \,\big|\, \theta(0)\big) < \infty$.

Proof of claim: Recall that we assume the Markov chain is irreducible. So for any state $j \in S$, there is a finite integer $t > 0$ for which,

$$
q_t(i,j) = P\big(\theta(t) = j,\, T^1(i) \geq t \,\big|\, \theta(0) = i\big) > 0.
$$

Using this,

$$
\begin{aligned}
E(\ \Delta\ (i)^\ell) &= E\big(\textstyle\prod_{k=1}^{T^1(i)} \max(\|A_{\theta(k)}\|^\ell, 1) \,\big|\, \theta(0) = i\big) \\
&\geq E\big(\textstyle\prod_{k=1}^{T^1(i)} \max(\|A_{\theta(k)}\|^\ell, 1)\| \,\big|\, \theta(0) = i, \theta(t) = j, T^1(i) \geq t\big) \times q_t(i,j) \\
&\geq E\big(\textstyle\prod_{k=t+1}^{T^1(i)} \|A_{\theta(k)}\|^\ell \,\big|\, \theta(t) = j\big) \times q_t(i,j) \\
&= E\big(\textstyle\prod_{k=1}^{R(i)} \|A_k^{R(i)}\|^\ell \,\big|\, \theta(0) = j\big) \times q_t(i,j).
\end{aligned}
$$

Dividing both sides by $q_t(i,j)$ gives the desired result. □

6.3 Conclusion

Infinite Markov jump linear systems are good models for many real-world systems. The stability analysis of such systems is difficult because the system may potentially exhibit an infinite number of different dynamics at random times. In general, one needs to solve an infinite number of balance equations.

In this chapter, we first presented a simple example to demonstrate a basic property of many real-world jump systems: even though there are an infinite number of different dynamics, only a finite set of these are typical. Consequently, the system, although seemingly complex, behaves like a simple and finite system. The main results of this paper (Theorems 6.2.1 and 6.2.2) was our attempt to capture this effect.

We are unaware of any other literature that has addressed this issue. In comparison to known results, we provide an alternative view of jump system that may be useful to determine stability for systems with the mentioned property.

Chapter 7

Future work: Noisy feedback

Consider a discrete memoryless channel and the basic problem of digital communication: the reliable transmission of an infinite stream of bits such that each bit experiences a finite end-to-end delay. In almost all practical situations, this problem is solved by grouping the bits into blocks and by the use of 'good' block codes. So the problem is not why we use such blocks but how to group the bits. If there is no feedback channel or if the feedback is perfect, there is no ambiguity among the encoder and decoder as to which block corresponds to which group of bits. So the overall reliability of the entire stream is determined by the reliability of the block code. Unfortunately, this construction cannot be taken for granted if the feedback channel is noisy. To see this, keep in mind that feedback is used to determine whether a particular block will be retransmitted at a future time. However, a long sequence of errors is always possible, so that the encoder and decoder are in disagreement about which block of bits is being transmitted. Observe that such a situation leads to a very large probability of decoding error which may effectively be the dominating term in the overall reliability of the bit stream.

In a practical situation we overcome the difficulty by the use of sequence numbers. Just think of the way we achieve reliability in TCP/IP. However, from a theoretical point of view, for any given block length, the sequence numbers would grow unboundedly so that asymptotically the transmission rate tends to zero. The infinite problem is therefore fundamentally interesting and different from the case of no feedback or even perfect feedback.

In the following paragraphs we describe a coding scheme that was described by A. Sahai and the author to address the infinite problem. The idea is to encode the retransmission requests using a low-rate Anytime code. Because of the Anytime property, the encoder and decoder are eventually guaranteed to be in agreement about which block is being transmitted. Therefore, the undetected feedback error can never dominate the overall reliability of the infinite bit-stream. The basic idea is to mimic Yamamoto and Itoh's [37, 38] variable-length block coding scheme for the infinite setting:

- Each block is broken up into sub-blocks which are transmitted using a good block-code. The decoder feeds back its estimate of each of the sub-blocks using an expurgated block code,
- After the estimate of the final sub-block is transmitted, the encoder emits a 'confirm' or 'deny' control signal. The decoder feeds back its estimate of the control signal using an anytime code. If denied, the decoder expects to receive a retransmission of the same block at a future time. If confirmed, the transmission of the block is finalized.
- Finally, blocks are interleaved so that a retransmission, if necessary, occurs at a suitable time in the future. This gives the encoder enough time to interpret the Anytime decision feedback with a very small probability of error.

It is sufficient for the encoder and decoder to agree on the history of the control signal (i.e. ‘confirm’ or ‘deny’) for them to agree on the current block being transmitted. The Anytime code on the feedback channel ensures this condition with an exponentially decaying probability of error. Therefore, disagreement cannot contribute significantly to a decoding error. An error occurs if the encoder transmitted a confirmation even though the preceding sub-blocks were not decoded correctly. This happens if there is an error in the expurgated feedback for the sub-blocks, leading to the main result of [23]:

$$E_f(\bar{R}) = \left(\frac{1}{C_1} + \frac{1}{E_{ex}}\right)^{-1}\left(1 - \frac{\bar{R}}{C}\right),\ 0 \le \bar{R} \le C,$$

where $\bar{R}$ is the average rate of transmission and C_1 is the relative entropy of the channel,

$$C_1 = \max_{i,k} \sum_{\ell} p_{i\ell} \log_2 \frac{p_{i\ell}}{p_{k\ell}}.$$

The interesting thing about this result is the limiting condition as the feedback noise tends to zero. Observe that for any rate $0 \le \bar{R} \le C$, the expurgated exponent $E_{ex}(\bar{R})$ tends to infinity as the feedback noise tends to zero. Thus, by comparison to (2.2.1) this shows that the reliability of the proposed methods tends to Burnashev’s variable-length block coding reliability as the feedback noise tends to zero. Because Burnashev’s reliability is tight, the proposed method is close to optimal if the feedback noise is relatively small.

The preliminary results described above and in detail in [23] show that Anytime coding can be a useful tool for using feedback in a noisy feedback setting. In Figure 7.1 we’ve plotted the exponent E_f for the case of the forward Binary Symmetric Channel (BSC) with cross-over probability .1 and the reverse BSC with cross-over probability 10^{-6}. In the same figure we plot the Sphere-packing bound for the forward channel. By comparison, noisy feedback can be used to significantly improve the reliability at higher rates. It turns out that our formulation does not lead to improvements for very noisy feedback. However, we believe extensions are possible: this is the subject of future research.

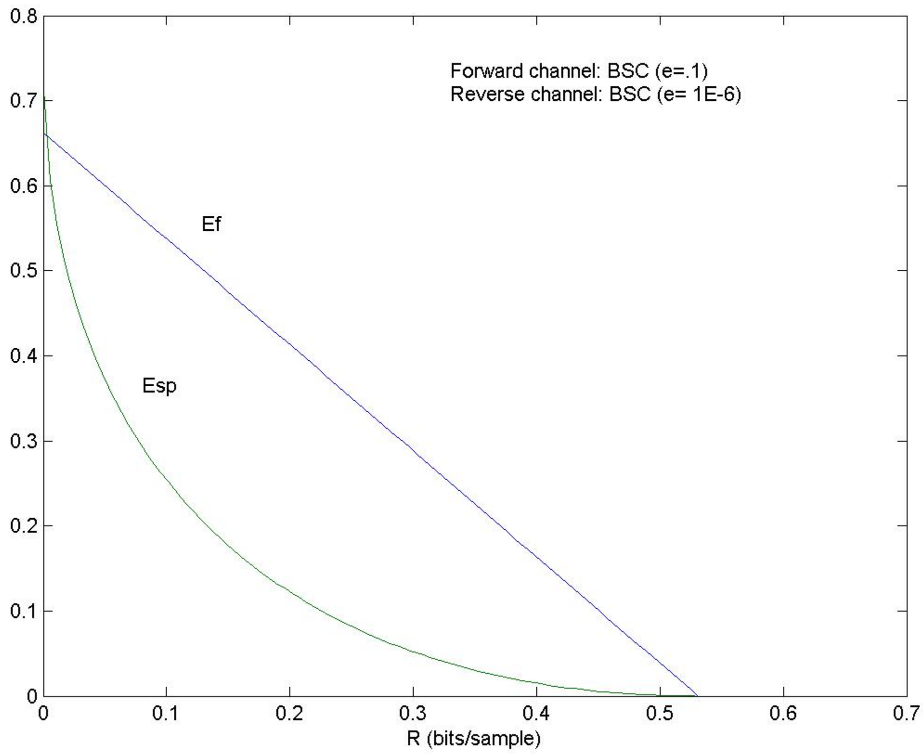

Figure 7.1: The noisy feedback exponent E_f and the Sphere-packing bound E_{sp}. The forward channel is a Binary Symmetric Channel (BSC) with cross-over probability .1 and the reverse channel is a BSC with cross-over probability 10^{-6}.

Bibliography

[1] E. A. Arutyunyan. Lower Bound for Error Probability in Channels with Feedback. *Problemy Peredachi Informatsii*, 13(2):36–44, 1977.

[2] A. Barg and G. D. Forney. Random Codes: Minimum Distances and Error Exponents. *IEEE Transactions on Information Theory*, 48(9):2568–2573, 2002.

[3] E. R. Berlekamp. Block Coding for the Binary Symmetric Channel with Noiseless, Delayless Feedback. *In H.B. Mann (ed.), Error-Correcting Codes*, John Wiley, New York, pp. 61–88, 1968.

[4] J. Bucklew. The Blind Simulation Problem and Regenerative Processes. *IEEE Transactions on Information Theory*, 44(7):2877–2891, 1998.

[5] M. V. Burnashev. Data Transmission over a Discrete Channel with Feedback, Random Transmission Time. *Problemy Peredachi Informatsii*, 12(4):10–30, 1976.

[6] S. S. L. Chang. Theory of Information Feedback Systems. *IEEE Transactions on Information Theory*, 2(3):29–40, 1956.

[7] O. Costa and M. D. Fragoso. Discrete-time LQ-Optimal Control Problems for Infinite Markov Jump Parameter Systems. *IEEE Transactions on Automatic Control*, 40(12):2076–2088, 1995.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley, New York, 1991.

[9] R. L. Dobrushin. An Asymptotic Bound for the Probability Error of Information Transmission Through a Channel Without Memory using the Feedback. *Problemy kibernetiki*, 8:161–168, 1962.

[10] R. Durrett. *Probability: Theory and Examples.* Second Edition, Duxbury Press, 1996.

[11] Y. Fang and K. A. Loparo. On the Relationship Between the Sample Path and Moment Lyapunov Exponents for Jump Linear Systems. *IEEE Transactions on Automatic Control*, 47(9):1556–1560, 2002.

[12] X. Feng, K. A. Loparo, Y. Ji, and H. J. Chizeck. Stochastic Stability Properties of Jump Linear Systems. *IEEE Transactions on Automatic Control*, 37(1):38–53, 1992.

[13] G. Forney. Exponential Error Bounds for Erasure, List, and Decision Feedback Schemes. *IEEE Transactions on Information Theory*, 14(2):206–220, 1968.

[14] R. Gallager. *Information Theory and Reliable Communication.* John Wiley, New York, 1968.

[15] A. Goldsmith and P. Varaiya. Capacity, Mutual Information, and Coding for Finite-State Markov Channels. *IEEE Transactions on Information Theory*, 42(3):868–886, 1996.

[16] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory.* Third Edition, John Wiley, New York, 1998.

[17] M. Horstein. Sequential Transmission using Noiseless Feedback. *IEEE Transactions on Information Theory*, 9(3):136–143, 1963.

[18] G. F. Lawler. *Introduction to Stochastic Processes.* Chapman & Hall, New York, 1995.

[19] V. I. Levenshtein. Straight-Line Bound for the Undetected Error Exponent. *Problemy Peredachi Informatsii*, 25(1):33–37, 1987.

[20] M. Mushkin and I. Bar-David. Capacity and Coding for the Gilbert-Elliott Channels. *IEEE Transactions on Information Theory*, 35(6):1277–1290, 1989.

[21] M. Pinsker. Bounds on the Probability and of the Number of Correctable Errors for Nonblock Codes. *Problemy Perdachi Informatsii*, 3(4):58–71, 1967.

[22] A. Sahai. *Anytime Information Theory.* PhD thesis, Massachusetts Institute of Technology, 2001.

[23] A. Sahai and T. Simsek. On the Variable-Delay Reliability Function of Discrete Memoryless Channels with access to Noisy Feedback. *IEEE Information Theory Workshop*, San Antonio, Texas, October 24–29, 2004.

[24] J. P. M. Schalkwijk. A Coding Scheme for Additive Noise Channels with Feedback – Part II: Bandlimited Channels. *IEEE Transactions on Information Theory*, 12(2):183–192, 1966.

[25] J. P. M. Schalkwijk. A Class of Simple and Optimal Strategies for Block Coding on the Binary Symmetric Channel with Noiseless Feedback. *IEEE Transactions on Information Theory*, 17(3):283–287, 1971.

[26] J. P. M. Schalkwijk and T. Kailath. A Coding Scheme for Additive Noise Channels with Feedback – Part I: No Bandwidth Constraint. *IEEE Transactions on Information Theory*, 12(2):172–182, 1966.

[27] J. P. M. Schalkwijk and K. A. Post. On the Error Probability for a Class of Binary Recursive Feedback Strategies. *IEEE Transactions on Information Theory*, 19(4):498–511, 1973.

[28] T. Simsek, R. Jain, and P. Varaiya. Control under Communication Constraints. *Proceedings of the 41st IEEE Conference on Decision and Control*, pages 3209–3216, 2002.

[29] S. Tatikonda. *Control under Communication Constraints.* PhD thesis, Massachusetts Institute of Technology, 2000.

[30] A. Tchamkerten and E. Telatar. A Feedback Strategy for Binary Symmetric Channels. *Proceedings of the IEEE International Symposium on Information Theory*, page 362, 2002.

[31] T. Veugen. A Simple Class of Capacity-Achieving Strategies for Discrete Memoryless Channels with Feedback. *IEEE Transactions on Information Theory*, 42(6):2221–2228, 1996.

[32] A. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[33] J. C. Walrand and P. Varaiya. Optimal Causal Coding-Decoding Problems. *IEEE Transactions on Information Theory*, 29(6):814–820, 1983.

[34] E. W. Weisstein. *Catalan Number.* From MathWorld–A Wolfram Web Resource, http://mathworld.wolfram.com/CatalanNumber.html.

[35] H. S. Witsenhausen. On the Structure of Real-time Source Coders. *Bell Systems Technical Journal*, 58:1437–1451, 1979.

[36] A. Wyner. On the Schalkwijk-Kailath Coding Scheme with a Peak Energy Constraint. *IEEE Transactions on Information Theory*, 14(1):129–134, 1968.

[37] H. Yamamoto and K. Itoh. Asymptotic Performance of a Modified Schalkwijk-Barron Scheme for Channels with Noiseless Feedback. *IEEE Transactions on Information Theory*, 25(6):729–733, 1979.

[38] H. Yamamoto and K. Itoh. Correction to "Asymptotic Performance of a Modified Schalkwijk-Barron Scheme for Channels with Noiseless Feedback". *IEEE Transactions on Information Theory*, 26(5):628, 1980.

[39] K. Zigangirov. Upper Bounds on the Error Probability for Channels with Feedback. *Probl. Per. Inf.*, 6(2):87–92, 1970.