Chapter 4

The Capacity of Finite State Channels

4.1 Introduction

Determining the achievable rates at which information can be reliably transmitted across noisy channels has been one of the central pursuits in information theory since Shannon invented the subject in 1948. In this chapter, we consider these rates for the class of channels known as finite state channels (FSC). A FSC is a discrete-time channel where the distribution of the channel output depends on both the channel input and the underlying channel state. This allows the channel output to depend implicitly on previous inputs and outputs via the channel state.

In practice, there are three types of channel variation which FSCs are typically used to model. A *flat fading* channel is a time-varying channel whose state is independent of the channel inputs. An *intersymbol-interference* (ISI) channel is a time-varying channel whose state is a deterministic function of the previous channel inputs. Channels which exhibit both fading and ISI can also modeled, and their state is a stochastic function of the previous channel inputs.

A number of other authors have dealt with FSCs in the past, and we review some of their important contributions. Since it is easy to construct degenerate FSCs, most of these results are limited to a particular set of well behaved FSCs. A FSC in this particular set is referred to as an indecomposable FSC (IFSC). Blackwell, Breiman, and Thomasian introduced IFSCs in [7] and proved the natural analogue of the channel coding theorem for them. Birch discusses the achievable information rates of IFSCs in [5], and computes bounds for a few simple examples. In [14, p.100], Gallager gives an elegant derivation of the coding theorem and provides a method to explicitly compute the capacity when the receiver has perfect channel state information. This method cannot be applied, however, when the receiver only has imperfect state estimates computed from the previous channel outputs. Hirt considers linear filter channels with additive white Gaussian noise (AWGN) and equiprobable binary inputs in [16], and develops a Monte Carlo method for estimating achievable rates. In [15], Goldsmith and Varaiya take a different approach and provide an explicit method of estimating the capacity of flat fading IFSCs (i.e., where the state sequence is independent of the transmitted sequence). In this chapter, we provide a simple Monte Carlo method of estimating the achievable information rates of any IFSC and we focus on the problem of estimating the capacity of IFSCs with ISI (i.e., where the state sequence is a deterministic function of the transmitted sequence).

It is worth noting that this method, reported in [24], was discovered independently by Arnold and Loeliger in [1] and by Sharma and Singh[1]. It is quite surprising, in fact, that this method was not proposed earlier. It is simply an efficient application of the famous Shannon-McMillan-Breiman theorem. Nonetheless, [1], [27][1], and [24] represent the first publications where the achievable information rates of a general IFSC are computed to 3 or 4 digits of accuracy. Furthermore, these advances stimulated new interest in the subject which led Kavčić to formulate a very elegant generalization of the Arimoto-Blahut algorithm for finite state channels in [19].

The achievable information rate of an IFSC, for a given input process, is equal to the mutual information rate between the stochastic input process and the stochastic output process. This mutual information rate, $I(\mathcal{X};\mathcal{Y})$, is given by

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y}), \tag{4.1.1}$$

where $H(\mathcal{X})$, $H(\mathcal{Y})$, and $H(\mathcal{X},\mathcal{Y})$ are the respective entropy rates of the input process, the output process, and the joint input-output process. The symmetric information rate (SIR) of an IFSC is the maximum rate achievable by an input process which chooses each input independently and equiprobably from the source alphabet. The capacity of an IFSC is the largest rate achievable by any input process.

Our simple Monte Carlo method is based on estimating each of the entropy rates in (4.1.1). These entropy rates are estimated by simulating a long realization of the process and

[1] While the Monte Carlo method is introduced correctly in [27], it appears that most of the other results in their paper, based on regenerative theory, are actually incorrect. A correct analytical treatment can be found in Section 4.4.4.

Channel	Transfer Function	Normalized Response
Dicode	$(1-D)$	[1 -1]/$\sqrt{2}$
EPR4	$(1-D)(1+D)^2$	[1 1 -1 -1]/2
E^{2}PR4	$(1-D)(1+D)^3$	[1 2 0 -2 -1]/$\sqrt{10}$

Table 4.1: The transfer function and normalized response of a few partial response targets.

computing its probability using the forward recursion of the well known BCJR algorithm [2]. The fact that this probability can be used to estimate the entropy rate is a consequence of the Shannon-McMillan-Breiman theorem [10, p. 474]. Furthermore, this approach is general enough to allow the mutual information rate to be maximized over Markov input distributions of increasing length, and thus can be used to estimate a sequence of non-decreasing lower bounds on capacity.

This chapter is organized as follows. In Section 4.2, we introduce a few example finite state channels which are discussed throughout the chapter. Mathematical definitions and notation for the chapter are introduced in Section 4.3. In Section 4.4, we address the problem of estimating entropy rates. In particular, this section discusses our simple Monte Carlo method, a general analytical method, and an interesting connection with Lyapunov exponents. Section 4.5 uses the results of the previous section to discuss upper and lower bounds on the capacity of finite state channels. In Section 4.6, we give the numerical results of applying the Monte Carlo method to the example channels. Exact information rates are derived for the dicode erasure channel in Section 4.7. A pseudo-analytical method of estimating information rates based on density evolution, which is quite efficient for two state channels, is also described. Finally, in Section 4.8, we provide some concluding remarks.

4.2 Channel Models

4.2.1 Discrete-Time Linear Filter Channels with AWGN

A very common subset of IFSCs is the set of discrete-time linear filter channels with additive white Gaussian noise (AWGN), which are described by

$$y_k = \sum_{i=0}^{\nu} h_i x_{k-i} + n_k, \tag{4.2.1}$$

where ν is the channel memory, $\{x_k\}$ is the channel input (taken from a discrete alphabet), $\{y_k\}$ is the channel output, and $\{n_k\}$ is i.i.d. zero mean Gaussian noise with variance σ^2. Bounds on the capacity and SIR of this channel have been considered by many authors. In particular, we note the analytical results of Shamai *et al.* in [26] and the original Monte Carlo results of Hirt in [16]. Some examples of these channels are listed in Table 4.1, and were chosen from the class of of binary-input channels which are used to model equalized magnetic recording channels. The state diagram for the noiseless dicode channel (i.e., before the AWGN) is shown in Figure 4.2.1. A formal mathematical definition of these channels is given in Appendix 4A.1.

Computing the achievable information rates of these channels can also be simplified by writing the mutual information rate (4.1.1) as

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}).$$

This is because the second term is simply the entropy of the Gaussian noise sequence, $\{n_k\}$, which can be written in closed form [10, p. 225] as

$$H(\mathcal{Y}|\mathcal{X}) = \frac{1}{2}\log(2\pi e\sigma^2).$$

Therefore, estimating the SIR of these channels reduces to estimating $H(\mathcal{Y})$, and estimating the capacity of this channel reduces to estimating the supremum of $H(\mathcal{Y})$ over all input processes.

4.2.2 The Dicode Erasure Channel

Since it is difficult, if not impossible, to derive a closed form expression for the entropy rate of the dicode channel with AWGN, we also consider the somewhat artificial dicode erasure channel (DEC). This is a simple channel based on the $1 - D$ linear ISI channel whose noiseless state diagram is shown in Figure 4.2.1. The DEC corresponds to taking the output of the dicode

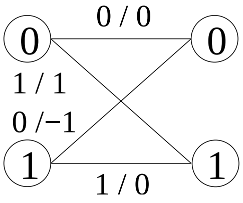

Figure 4.2.1: The state transition diagram of the dicode channel.

channel, $(+1, 0, -1)$, and either erasing it with probability ϵ or transmitting it perfectly with probability $1 - \epsilon$. The state diagram for the noiseless dicode channel is shown in Figure 4.2.1. A formal mathematical definition of the DEC is channel is given in Appendix 4A.2.

The properties of this channel are similar to the dicode channel with AWGN, and again the mutual information rate can be simplified to

$$I(\mathcal{X};\mathcal{Y}) = H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}).$$

In this case, the second term is simply the entropy of the erasure position sequence which can be written in closed form as $H(\mathcal{Y}|\mathcal{X}) = -\epsilon \log \epsilon - (1-\epsilon)\log(1-\epsilon)$. Therefore, the SIR and capacity of this channel can also be determined by considering only $H(\mathcal{Y})$.

4.2.3 The Finite State Z-Channel

The Z-channel is a well-known discrete memoryless channel (DMC) which models a communications system with one “good” symbol and “bad” symbol. The “good” symbol is transmitted perfectly by the channel and the “bad” symbol is either transmitted correctly (with probability $1 - p$) or swapped with the “good” symbol (with probability p). Consider a finite state analogue of this channel in which the “good” and “bad” symbols are not fixed, but depend on the previous input symbol. One trellis section for such a channel, which we call the finite state Z-channel is shown in Fig. 4.2.2. The edges are labeled with the input bit and the output bits, where $B(p)$ stands for the Bernoulli distribution which produces a one with probability p. A formal mathematical definition of this channel is given in Appendix 4A.3.

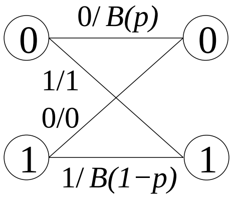

Figure 4.2.2: The state transition diagram of the finite state Z-channel. The symbol $B(p)$ refers to a binary random variable which equals 1 with probability p and 0 with probability $1 - p$.

4.3 Definitions

4.3.1 The Indecomposable Finite State Channel

A finite state channel (FSC) is a stochastic mapping from a sequence of inputs, $\{X_t\}_{t\geq 1}$, chosen from the finite input alphabet, $\mathbb{X}$, to a sequence of outputs, $\{Y_t\}_{t\geq 1}$, chosen from the (possibly infinite) output alphabet, $\mathbb{Y}$. Let $\{S_t\}_{t\geq 1}$ be the state sequence of the channel, which takes values in the finite set $\mathcal{S} = \{0, 1, \dots, N_S - 1\}$. When the output alphabet is countable, the channel statistics are completely defined by the time-invariant conditional probability, $f_{ij}(x, y) \triangleq Pr(Y_t = y, S_{t+1} = j | X_t = x, S_t = i)$. For uncountable $\mathbb{Y}$, we abuse this notation slightly and let, for each j, $f_{ij}(x, y)$ be a continuous density function of the output, y, given starting state i and input x. In this way, we formally define a finite state channel by the triple, $(\mathbb{X}, \mathbb{Y}, \mathbf{F}(\cdot, \cdot))$, where $[\mathbf{F}(x, y)]_{ij} = f_{ij}(x, y)$. Each example channel in Section 4.2 is defined formally using this notation in Appendix 4A.

Since many properties of a FSC can be related to the properties of a finite state Markov chain (FSMC), we start by reviewing some terminology from the theory of FSMCs. A FSMC is *irreducible* if there is a directed path from any state to any other state. If the greatest common divisor of the lengths of all cycles (i.e., paths from a state back to itself) is one, then it is *aperiodic*. A FSMC is ergodic or *primitive* if it is both irreducible and aperiodic. These ideas can also be applied to a non-negative square matrix, $\mathbf{A}$, by associating the matrix with a FSMC which has a path from state i to state j if and only if $[\mathbf{A}]_{ij} > 0$. Using this, we say that a FSC is *indecomposable* if its zero-one connectivity matrix, defined by

$$[\mathbf{F}(*, *)]_{ij} = \begin{cases} 1 & \exists\, x \in \mathbb{X}, y \in \mathbb{Y} \text{ s.t. } f_{ij}(x, y) > 0 \\ 0 & \text{otherwise} \end{cases},$$

is primitive.

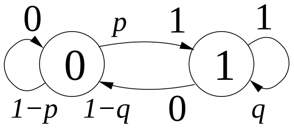

Figure 4.3.1: The state diagram of a two state input process which sends a 1 with probability p from the 0 state and with probability q from the 1 state.

4.3.2 The Markov Input Process

When computing achievable information rates, it is typical to treat the input sequence as a stochastic process as well. Let $\{T_t\}_{t\geq 1}$ be the state sequence of an ergodic FSMC taking values in the finite set $\mathcal{T} = \{0, 1, \ldots, N_T - 1\}$. The statistics of the input process, $\{X_t\}_{t\geq 1}$, are defined by the transition probabilities of the chain, $\theta_{ij} \triangleq Pr(T_{t+1} = j | T_t = i)$, and the edge labels, ϕ_{ij}, with $X_t = \phi_{T_t,T_{t+1}}$. We refer to this type of input process as a Markov input process, and denote it by the pair $(\mathbf{\Theta}, \mathbf{\Phi})$, where $[\mathbf{\Theta}]_{ij} = \theta_{ij}$ and $[\mathbf{\Phi}]_{ij} = \phi_{ij}$.

For example, the state diagram of a general two state Markov input process in shown in Figure 4.3.1. The formal definition of this same process is given by $(\mathbf{\Theta}, \mathbf{\Phi})$ where $\theta_{0,1} = 1 - \theta_{0,0} = p$, $\theta_{1,1} = 1 - \theta_{1,0} = q$, $\phi_{0,0} = \phi_{1,0} = 0$, and $\phi_{0,1} = \phi_{1,1} = 1$.

4.3.3 Combining the Input Process and the Finite State Channel

When the channel inputs are generated by a Markov input process, the channel output, $\{Y_t\}_{t\geq 1}$, can be viewed as coming from stochastic process. In this case, the distribution of Y_t depends only on state transitions in the combined state space of the channel and input. Let $\mathcal{Q} = \{0, 1, \ldots, N_T N_S - 1\}$ and, for any $q \in \mathcal{Q}$, let the $\mathcal{T}$-state of q be $r(q) = \lfloor q/N_S \rfloor$ and the $\mathcal{S}$-state of q by $s(q) = q \mod N_S$. Using this, we can write the state transition probabilities of the combined process as

$$p_{ij} \triangleq Pr(Q_{t+1} = j | Q_t = i) = \theta_{r(i),r(j)} \int_{\mathbb{Y}} f_{s(i),s(j)}(\phi_{r(i),r(j)}, y) dy,$$

where the integral is taken to be a sum if $\mathbb{Y}$ is countable. We also define the conditional observation probability of y, given the transition, to be

$$g_{ij}(y) \triangleq Pr(Y_t = y | Q_{t+1} = j, Q_t = i) = f_{s(i),s(j)}(\phi_{r(i),r(j)}, y).$$

We refer to the stochastic output sequence, $\{Y_t\}_{t\geq 1}$, as a finite state process (FSP) and define it formally in the next section.

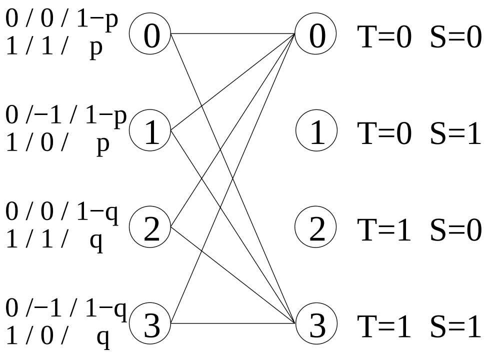

Figure 4.3.2: The combined state diagram for a general two state Markov input process and the dicode channel.

As an example, we show in Figure 4.3.2 the state diagram formed by combining a general two state Markov input process with the dicode channel. The edge labels on the left side of the figure give the input symbol, the output symbol, and the transition probability for each edge. Each state is labeled by its Q-value, and the corresponding S and T values are also shown on the right side of the figure.

Remark 4.3.1. One problem with joining the state spaces of the input and channel processes is that the resulting Markov chain may no longer be primitive. Suppose that the input process and the channel keep the same state variable (e.g., the input process remembers its last output and the channel remembers its last input). The state diagram for the combined process of this type is shown in Figure 4.3.2. The resulting Markov chain is reducible, but it still has a unique ergodic component. Taking only the ergodic component, consisting of states 0 and 3, results in an ergodic finite state process. In other cases, the state diagram for the combined process may actually be disconnected. In general, we will require that the Markov chain associated with the combined process is primitive. Therefore, some care must taken in choosing the input process and/or reducing the combined process. Another example of this problem is given in Appendix 4A.4.

4.3.4 The Finite State Process

Let $\{Q_t\}_{t\geq 1}$ be an ergodic FSMC taking values from the set $\mathcal{Q} = \{0, 1, \ldots, N_Q - 1\}$. The finite state process (FSP), $\{Y_t\}_{t\geq 1}$, is an ergodic stochastic process controlled by $\{Q_t\}_{t\geq 1}$ which takes values from the alphabet $\mathbb{Y}$. The transition probabilities for $\{Q_t\}_{t\geq 1}$ are given by $Pr(Q_{t+1} = j|Q_t = i) = p_{ij}$, and the dependence of $\{Y_t\}_{t\geq 1}$ on $\{Q_t\}_{t\geq 1}$ is given by $Pr(Y_t = y|Q_{t+1} = j, Q_t = i) = g_{ij}(y)$. While this notation is precise for countable $\mathbb{Y}$, we abuse it slightly for uncountable $\mathbb{Y}$ and let $g_{ij}(y)$ be the continuous density function of the output, y, associated with the state transition from state i to state j. The FSP, $\{Y_t\}_{t\geq 1}$, is defined formally by the triple $(\mathbb{Y}, \mathbf{P}, \mathbf{G}(\cdot))$, where $p_{ij} = [\mathbf{P}]_{ij}$ and $g_{ij}(\cdot) = [\mathbf{G}(\cdot)]_{ij}$.

We note that any FSP can be stationary if the initial state if chosen properly. Let $\boldsymbol{\pi} = [\ \pi_1 \quad \pi_2 \quad \ldots \quad \pi_r\ ]$ be the unique stationary distribution of $\{Q_t\}_{t\geq 1}$ which satisfies $\pi_j = \sum_i \pi_i p_{ij}$. If the initial state, Q_1, of the underlying Markov chain is chosen such that $Pr(Q_1 = j) = \pi_j$, then $\{Y_t\}_{t\geq 1}$ is stationary in the sense that $Pr(\mathbf{Y}_1^k = \mathbf{y}_1^k) = Pr(\mathbf{Y}_{t+1}^{t+k} = \mathbf{y}_1^k)$ for all $k \geq 0$ and all $t \geq 1$. This initialization is assumed throughout the discussion of FSPs.

If $\mathbb{Y}$ is a finite set, then an identical process can also be generated as a function of a FSMC. More precisely, this means that there exists a FSMC, $\{X_t\}_{t\geq 1}$, and a mapping ξ, such that $Y_t = \xi(X_t)$. The process $\{Y_t\}_{t\geq 1}$ can also described as the output of a hidden Markov model. We present $\{Y_t\}_{t\geq 1}$ as a FSP because it is the most natural represetation when considering the entropy rate of the process.

4.4 The Entropy Rate of a Finite State Process

Since a number of authors have considered the entropy rate of a FSP in the past, we review some of the key results. Blackwell appears to have been the first to consider the entropy rate of a function of a FSMC. In [6], he gives an explicit formula for the entropy rate, in terms of the solution to an integral equation, and he notes that this result suggests that the entropy rate is “intrinsically a complicated function” of the underlying Markov chain, $\{X_t\}_{t\geq 1}$, and the mapping, ξ. Birch [5] derives a sequence of Markov upper and lower bounds for the entropy rate and shows, under fairly restrictive conditions, that the gap between them converges to zero exponentially fast. The complexity of computing his bounds also grows exponentially, however, making them less useful in practice.

Here, we attack the problem first by introducing an efficient Monte Carlo method of estimating the entropy rate based on the Shannon-McMillan-Breiman theorem. Then, we work towards analytical approaches of computing the entropy rate (via Blackwell’s integral equation). We also discuss conditions under which a central limit theorem (CLT) holds for the entropy rate. Under these same conditions, we prove that the gap between sequences of Markov upper and lower bounds on the entropy rate converges to zero exponentially fast. Finally, we describe a connection between the entropy rate of a FSP and the largest Lyapunov exponent of an associated sequence of random matrices. It is worth noting that the natural logarithm is denoted $\ln$ while the base 2 logarithm is denoted $\log$.

4.4.1 A Simple Monte Carlo Method

Let $\{Y_t\}_{t\geq 1}$ be an ergodic FSP defined by $(\mathbb{Y}, \mathbf{P}, \mathbf{F}(\cdot))$. We start by using the definition of the entropy rate for a stationary process [10, Chap. 4],

$$H(\mathcal{Y}) \triangleq - \lim_{n\to\infty} \frac{1}{n} E\left[\log Pr(\mathbf{Y}_1^n)\right],$$

to define the sample entropy rate as

$$\hat{H}_n(\mathbf{Y}_1^n) = -\frac{1}{n} \log Pr(\mathbf{Y}_1^n). \tag{4.4.1}$$

It is worth noting that $\hat{H}_n(\mathbf{Y}_1^n)$ is a random variable, and the asymptotic convergence of that random variable to the true entropy rate is guaranteed by the Shannon-McMillan-Breiman theorem [10, p. 474]. Mathematically speaking, this theorem states that

$$\lim_{n\to\infty} -\frac{1}{n} \log Pr(\mathbf{Y}_1^n) = H(\mathcal{Y})$$

for almost all realizations of $\mathbf{Y}_1^n$ (i.e., almost surely). While the original proof only holds for finite alphabet processes, it was extended to more general processes by Barron [3].

Efficiently applying the Shannon-McMillan-Breiman theorem to our FSP is equivalent to efficiently computing $\log Pr(\mathbf{Y}_1^n)$ for large n. This quantity has a natural decomposition of the form

$$\log Pr(\mathbf{Y}_1^n) = \sum_{t=1}^{n} \log Pr(Y_t | \mathbf{Y}_1^{t-1}), \tag{4.4.2}$$

and it turns out that the forward recursion of the BCJR algorithm [2] is ideal for computing this quantity. We note that random realizations, $\mathbf{y}_1^n$, of the process, $\mathbf{Y}_1^n$, are generated as a byproduct of any channel simulation. Let us define the forward state probability vector at time t, $\boldsymbol{\alpha}^{(t)}$, in terms of its components,

$$\alpha_i^{(t)} = Pr(Q_t = i | \mathbf{Y}_1^{t-1} = \mathbf{y}_1^{t-1}), \tag{4.4.3}$$

for $i \in \mathcal{Q}$. Using this, the forward recursion of the BCJR algorithm can be written as

$$\alpha_j^{(t+1)} = \frac{1}{A_t} \sum_{i=0}^{N_Q-1} \alpha_i^{(t)} Pr(Y_t = y_t, Q_{t+1} = j | Q_t = i), \tag{4.4.4}$$

where A_t is the standard normalization factor chosen to ensure that $\sum_{j=0}^{N_Q-1} \alpha_j^{(t+1)} = 1$. We note that the probability, $Pr(Y_t = y, Q_{t+1} = j | Q_t = i)$, required by (4.4.4) depends on the FSP and can be written as

$$\begin{aligned} Pr(Y_t = y, Q_{t+1} = j | Q_t = i) &= Pr(Y_t = y | Q_{t+1} = j, Q_t = i) Pr(Q_{t+1} = j | Q_t = i) \\ &= g_{ij}(y) p_{ij}. \end{aligned}$$

Proposition 4.4.1. *The sample entropy rate of a realization,* $\mathbf{y}_1^n$*, of the FSP,* $\{Y_t\}_{t \geq 1}$*, is given by*

$$\hat{H}_n(\mathbf{y}_1^n) = -\frac{1}{n} \sum_{t=1}^{n} \log A_t.$$

Proof. From (4.4.4), we see that

$$\begin{aligned} A_t &= \sum_{j=0}^{N_Q-1} \alpha_j^{(t+1)} \\ &= \sum_{j=0}^{N_Q-1} \left(\sum_{i=0}^{N_Q-1} \alpha_i^{(t)} Pr(Y_t = y_t, Q_{t+1} = j | Q_t = i) \right) \\ &= Pr(Y_t = y_t | \mathbf{Y}_1^{t-1} = \mathbf{y}_1^{t-1}), \end{aligned}$$

which means that $\log Pr(\mathbf{Y}_1^n)$ can be computed using (4.4.2). Combining this with (4.4.1) completes the proof. □

Remark 4.4.2. The complexity of this method is linear in the number of states, N_Q, and linear in the length of the realization, n. Furthermore, if a central limit theorem holds for the entropy rate, then the variance of the estimate will decay like $O\left(n^{-1/2}\right)$.

We believe that the rapid mixing of the underlying Markov chain and the form of (4.4.2) leads naturally to a central limit theorem for the sample entropy rate. The following conjecture makes this notion precise. We note that the conclusion of this conjecture is proven, under more restrictive conditions, in Section 4.4.6.

Conjecture 4.4.3. *Let $\{Y_t\}_{t\geq 1}$ be an ergodic FSP which gives rise to the conditional probability sequence, $\{A_t\}_{t\geq 1}$, where $A_t = Pr(Y_t|\mathbf{Y}_1^{t-1})$. If (i) $\lim_{t\to\infty} E\left[(-\log A_t)^{2+\epsilon}\right] < \infty$, then the sample entropy rate obeys a central limit theorem of the form*

$$\sqrt{n}\left[\hat{H}_n(\mathcal{Y}) - H(\mathcal{Y})\right] \xrightarrow{d} N(0,\sigma^2).$$

The variance, σ^2, of the estimate is given by

$$\sigma^2 = R(0) + 2\sum_{\tau=1}^{\infty} R(\tau), \tag{4.4.5}$$

where $R(\tau) = \lim_{t\to\infty} E\left[(\log A_t + H(\mathcal{Y}))\left(\log A_{t-\tau} + H(\mathcal{Y})\right)\right]$. If we also have that (ii) $\lim_{t\to\infty} E\left[(-\log A_t)^{4+\epsilon}\right] < \infty$, then we can estimate the variance using finite truncations of (4.4.5) with $R(\tau)$ set to the sample autocorrelation,

$$\hat{R}_n(\tau) = \frac{1}{n-\tau}\sum_{t=\tau+1}^{n}\left(\log A_t + \hat{H}_n(\mathcal{Y})\right)\left(\log A_{t-\tau} + \hat{H}_n(\mathcal{Y})\right).$$

Motivation. This conjecture is based on the fact that $\{A_t\}_{t\geq 1}$ is asymptotically stationary and our belief that the autocorrelation, $R(\tau)$, decays exponentially with τ. These conditions are generally sufficient to imply a central limit theorem for sums like (4.4.2). □

4.4.2 The Statistical Moments of Entropy

While the entropy of a random variable is usually defined to be $E\left[-\log Pr(Y)\right]$, one might also consider the random variable $Z = -\log Pr(Y)$. We refer to the kth moment of the random variable, Z, as the kth moment of the entropy. One reason for examining these quantities is that most CLTs require that the increments have finite second moments. Here, we show, under mild conditions, that the kth moment of the entropy is bounded, for all finite k.

Let $p(y)$ be the probability density of any absolutely continuous random variable. Since the function $p(y)$ must integrate to one, we know the tails must decay faster than $1/|y|$. If we assume the slightly stronger condition that $p(|y|) = O(|y|^{-1-\epsilon})$, for some $\epsilon > 0$, then we find that all finite moments of the entropy are bounded. Recall that all finite moments of a random variable are finite if the exponential moments, $E\left[e^{sZ}\right] = E\left[Pr(Y)^{-s}\right]$, are finite for some $s > 0$. We can upper bound this expectation with

$$\begin{aligned} E\left[Pr(Y)^{-s}\right] &= \int_{-\infty}^{\infty} p(y)p(y)^{-s}dy \\ &\leq \int_{-a}^{a} p(y)^{1-s}dy + 2C\int_{a}^{\infty} |y|^{(s-1)(1+\epsilon)}dy, \end{aligned}$$

where a is chosen large enough that $p(y) \leq C|y|^{-1-\epsilon}$ for all $|y| > a$. Using the fact that $p(y) < p(y)^{1-s}$ whenever $p(y) > 1$, it is easy to verify that the first term is less than $2a$. The second term will also be finite as long as $(s-1)(1+\epsilon) < 1$ which is equivalent to $s < \epsilon/(1+\epsilon)$. Since this expectation is finite for $s \in [0, \epsilon/(1+\epsilon))$, all finite moments of Z are bounded.

There are also distributions that are poorly behaved with respect to entropy, however. Consider the probability distribution on the integers given by

$$Pr(Y = n) = \frac{1}{Cn(\log n)^{\rho}},$$

for $n \geq 3$. As long as $\rho > 1$, we can compute a finite

$$C = \sum_{n=3}^{\infty} \frac{1}{n(\log n)^{\rho}}$$

which normalizes this distribution. The kth moment of the entropy for this distribution is given by

$$\sum_{n=3}^{\infty} \frac{1}{Cn(\log n)^{\rho}} \left(-\log\left(Cn(\log n)^{\rho}\right)\right)^{k},$$

which can be lower bounded by

$$\sum_{n=n_0}^{\infty} \frac{2^{-k}}{Cn(\log n)^{\rho-k}}$$

if n_0 is chosen large enough that $\log\left(Cn(\log n)^{\rho}\right) \geq (\log n)/2$. This lower bound will be finite only if $\rho - k > 1$. So for $1 < \rho \leq 2$, the distribution is well-defined but the entropy and all higher moments are infinite. Likewise, the finite variance condition necessary for a CLT requires that $\rho > 3$.

Now, let us focus on the value of the entropy increment, $-\log A_t$, during a transition from state i to state j. In this case, the true distribution of Y_t is given by $g_{ij}(y)$, but the simulation method computes A_t based on the assumed distribution,

$$P_{\boldsymbol{\alpha}}(y) = \sum_{i,j} \alpha_i^{(t)} h_i(y),$$

where $h_i(y) = Pr(Y_t = y | Q_t = i) = \sum_j p_{ij} g_{ij}(y)$. For a particular transition and forward state probabilities, the expectation of $-\log A_t$ can now be written as

$$E\left[-\log A_t | Q_t = i, Q_{t+1} = j, \boldsymbol{\alpha}\right] = E_{g_{ij}(Y)} - \left[\log P_{\boldsymbol{\alpha}}(Y)\right].$$

This general approach can also be used to upper bound the higher moments, $E[-\log A_t]^k$, required by Conjecture 4.4.3. In particular, we consider the bound

$$E_{g_{ij}(Y)}\left[(-\log P_{\boldsymbol{\alpha}}(Y))^k\right] \leq E_{g_{ij}(Y)}\left[\left(-\log\left(\min_i h_i(Y)\right)\right)^k\right],$$

which is based on maximizing the LHS over $\boldsymbol{\alpha}$.

Suppose that the output alphabet is finite (or a bounded continuous set) and there is an $\epsilon > 0$ such that $\min_i \inf_y h_i(y) \geq \epsilon$. In that case, the magnitude of the kth moment can be upper bounded with $E[-\log A_t]^k \leq (-\log\epsilon)^k$. If the output alphabet is countably infinite (or an unbounded continuous set) and $h_i(y) > 0$ for all bounded y, then the magnitude of the kth moment will depend only on the tails of the $h_i(y)$. Let $g(y)$ be the $g_{ij}(y)$ whose tail decays most slowly and $h(y)$ be the $h_i(y)$ whose tail decays most quickly. The magnitude of the kth moment will be finite if

$$E_{g(Y)}\left[(-\log h(Y))^k\right] < \infty.$$

Example 4.4.4. Suppose all of the $g_{ij}(y)$ are Gaussian densities with finite mean and variance, We assume that the particular mean and variance depends on the transition $i \to j$. In this case, the tails of each density decay like $O(e^{-ay^2})$ where a depends on the variance. The magnitude of the kth cross-moment for any two Gaussians is upper bounded by

$$\begin{aligned}\int_{-\infty}^{\infty} C_1 e^{-ay^2}\left(-\log(C_2 e^{-by^2})\right)^k dy &= \int_{-\infty}^{\infty} C_1 e^{-ay^2}\left(-\log C_2 + \frac{by^2}{\ln 2}\right)^k dy \\ &= \sum_{i=0}^{k}(-\log C_2)^i\left(\frac{b}{\ln 2}\right)^{k-i}\int_{-\infty}^{\infty} C_1 e^{-ay^2} y^{2(k-i)} dy.\end{aligned}$$

Since the integral really just computes the $2(k-i)$th moment of a Gaussian, the expression is bounded for all finite k.

4.4.3 A Matrix Perspective

In this section, we introduce a natural connection between the product of random matrices and the entropy rate of a FSP. This connection is interesting in its own right, but will also be very helpful in understanding the results of the next few sections.

Definition 4.4.5. For any $y \in \mathbb{Y}$, the *transition-observation probability matrix,* $\mathbf{M}(y)$, is an $N_Q \times N_Q$ matrix defined by

$$[\mathbf{M}(y)]_{ij} \triangleq Pr(Y_t = y, Q_{t+1} = j | Q_t = i) = p_{ij} f_{ij}(y).$$

These matrices behave similarly to transition probability matrices because their sequential products compute the n-step transition observation probabilities of the form,

$$[\mathbf{M}(y_k)\mathbf{M}(y_{k+1}) \ldots \mathbf{M}(y_{k+n})]_{ij} = Pr(\mathbf{Y}_k^{k+n} = \mathbf{y}_k^{k+n}, Q_{k+n+1} = j | Q_k = i).$$

This means that we can write $Pr(\mathbf{Y}_1^n)$ as the matrix product

$$Pr(\mathbf{Y}_1^n) = \boldsymbol{\pi}\mathbf{M}(y_1)\mathbf{M}(y_2) \ldots \mathbf{M}(y_n)\mathbf{1}, \tag{4.4.6}$$

where $\boldsymbol{\pi}$ is the row vector associated with the unique stationary distribution of $\{Q_t\}_{t \geq 1}$ and $\mathbf{1}$ is a column vector of all ones.

The forward recursion of the BCJR algorithm can also be written in matrix form with

$$\boldsymbol{\alpha}^{(t+1)} = \frac{\boldsymbol{\alpha}^{(t)}\mathbf{M}(y_t)}{\left\|\boldsymbol{\alpha}^{(t)}\mathbf{M}(y_t)\right\|_1}, \tag{4.4.7}$$

where $\boldsymbol{\alpha}^{(t)} = [\ \alpha_1^{(t)} \quad \alpha_2^{(t)} \quad \ldots \quad \alpha_{N_Q}^{(t)}\ ]$ and $\|\mathbf{x}\|_1 = \sum_i |x_i|$. This update formula is referred to as the *projective product*, and its properties are discussed at some length in [20]. We note that the order of the matrix-vector product in (4.4.7) is reversed with respect to [20]. The two most important properties of the projective product given by Lemma 2.2 of [20] are: (i) it is Lipschitz continuous if the smallest row sum is strictly greater than zero and (ii) it is a strict contraction if the matrix is positive. We note that these are really the only properties required for a self-contained proof of Theorem 4.4.9 which is stated in the next section.

4.4.4 The Analytical Approach

It appears that the most straightforward analytical approach to the entropy rate problem is the original method proposed by Blackwell [6]. Applying the same approach to this setup gives an integral equation whose solutions are the stationary distributions of the joint Markov chain formed by joining the true state and the forward state probability vector, $\left\{Q_t, \boldsymbol{\alpha}^{(t)}\right\}_{t \geq 1}$. The entropy rate is then computed with

$$\lim_{t \to \infty} E\left[\log Pr(Y_t | \mathbf{Y}_1^{t-1})\right] = \lim_{t \to \infty} E\left[\log \sum_{i=0}^{N_Q - 1} Pr(Y_t | Q_t = i) Pr(Q_t = i | \mathbf{Y}_1^{t-1})\right],$$

where $Pr(Y_t = y|Q_t = i) = \sum_j p_{ij} f_{ij}(y)$ is independent of t and the limit distribution, $\lim_{t\to\infty} Pr(Q_t = i|\mathbf{Y}_1^{t-1})$, depends on the true state, q_t, and is given by a stationary distribution of the joint Markov chain (cf., a solution of Blackwell's integral equation). One problem with this method, besides its general intractability, is the fact that the stationary distribution may not be unique. This is equivalent to saying that the integral equation may not have a unique solution.

Since many of the probability distributions in this section can be rather badly behaved, rigorous treatment requires that we use some measure theory. The following analysis is based on general state space Markov chains as described in [22]. Let $\Omega = \mathcal{Q} \times \mathfrak{D}(\mathcal{Q})$ be the sample space of the joint Markov chain, where $\mathcal{Q} = \{0, 1, \ldots, N_Q - 1\}$ and $\mathfrak{D}(\mathcal{Q})$ is the set of probability distributions (i.e., the set of non-negative vectors of length N_Q which sum to one). Let $\{\mu_t(q, A)\}_{t\geq 1}$ be the probability measure defined by $\mu_t(q, A) = Pr(\boldsymbol{\alpha}^{(t)} \in A, Q_t = q)$ for any $A \in \Sigma$, where Σ is the sigma field of Borel subsets of $\mathfrak{D}(\mathcal{Q})$. The transitions of this Markov chain are described by

$$\mu_{t+1}(j, A) = \sum_{i=0}^{N_Q-1} \int_{\mathfrak{D}(\mathcal{Q})} \mu_t(i, dx) P_{ij}(x, A),$$

where the transition kernel, $P_{ij}(x, A) = Pr(\boldsymbol{\alpha}^{(t+1)} \in A, Q_{t+1} = j|\boldsymbol{\alpha}^{(t)} = x, Q_t = i)$, is a probability measure defined on $A \in \Sigma$. The kernel can be written explicitly as

$$P_{ij}(x, A) = \int_{\{z\in\mathbb{Y}|L(x,y)\in A\}} p_{ij} g_{ij}(dz),$$

where $L(\boldsymbol{\alpha}, y) = \boldsymbol{\alpha}\mathbf{M}(y)/\left\|\boldsymbol{\alpha}\mathbf{M}(y)\right\|_1$ is the forward recursion update.

Before we continue, it is worth discussing some of the standard definitions and notation associated with Markov chains on general state spaces. Our notation, $P_{ij}(x, A)$, for the transition kernel is natural, albeit somewhat non-standard, considering the decomposition of our state space into discrete and continuous components. The n-step transition kernel is denoted $P_{ij}^{(n)}(x, A)$, and the unique stationary distribution is denoted $\pi(i, A)$ if it exists. The transition kernel can also be treated as an operator which maps the set of bounded measurable functions back to itself. The operator notation is given by

$$P^{(n)} r(i, x) = \sum_{j=0}^{N_Q-1} \int_{\mathfrak{D}(\mathcal{Q})} P_{ij}^{(n)}(x, dz) r(j, z),$$

and is useful for discussing the convergence of a Markov chain to a stationary distribution.

A general state space Markov chain is *uniformly ergodic* if it converges in total variation to a unique stationary distribution at a geometric rate which is independent of the starting state [22, p. 382]. This is equivalent to saying that there exists some $\rho < 1$ such that

$$\sup_{i,x} \left| P^{(n)} r(i,x) - \sum_{j=0}^{N_Q - 1} \int_{\mathfrak{D}(\mathcal{Q})} r(j,z)\pi(j,dz) \right| \leq C\rho^n \tag{4.4.8}$$

for all bounded measurable functions, $r(i,A)$, which satisfy $\sup_{i,A} |r(i,A)| \leq 1$. This type of convergence is generally too strong for our problem, however. If (4.4.8) holds only for all bounded continuous functions (in some topology), then the Markov chain converges weakly[2] to a unique stationary distribution. While this behavior is referred to as *geometric ergodicity* in [21], we say instead that the Markov chain is *weakly uniform ergodic* to avoid confusion with the geometric ergodicity defined in [22, p. 354].

Now, we consider the first condition under which the limit distribution, $\pi(s,A) = \lim_{t\to\infty} \mu_t(s,A)$, exists and is unique. This is based on a comment by Blackwell describing when the support of $\pi(s,A)$ is at most countably infinite [7]. Under this condition, Theorem 4.4.7 shows that the joint Markov chain is uniformly ergodic.

Condition 4.4.6. The output alphabet, $\mathbb{Y}$, is countable and there exists a finite output sequence which gives the observer perfect state knowledge (i.e., the joint Markov chain is in true state q with $\alpha_q = 1$). Using the DEC for an example, we see that the output $y_t = 1$ satisfies this condition because it implies with certainty that $s_{t+1} = 1$.

Theorem 4.4.7. *If Condition 4.4.6 holds, then $\pi(s,A)$ exists, is unique, and is supported on a countable set. Furthermore, the joint Markov chain is uniformly ergodic.*

Proof. Let z be state of the joint Markov chain after the output sequence which provides perfect state knowledge. Since this state is reachable from any other state, the ψ*-irreducibility* of this Markov chain is given by Theorem 4.0.1 of [22]. The state, z, also satisfies the conditions of an *atom* as defined in [22, p. 100]. Since any finite output sequence will occur infinitely often with probability 1, the point z is also *Harris recurrent* as defined in [22, p. 200]. Applying Theorem 10.2.2 of [22] shows that $\pi(s,A)$ exists and is unique.

[2] This set of functions provides a metric for the weak convergence of probability measures on a separable metric space [29].

Next, we show that $\pi(s, A)$ is supported on a countable set. Since the return time to state z is finite with probability 1, we assume the joint Markov chain is in state z at time τ and index any state in the support set by its output sequence, $\{y_t\}_{t \geq \tau}$, starting from state z. Therefore, the support set of $\pi(s, A)$ is at most the set of finite strings generated by the alphabet $\mathbb{Y}$, which is countably infinite. In particular, for any $\epsilon > 0$, there is a finite set of strings with total probability greater than $1 - \epsilon$.

Since the underlying Markov chain, $\{Q_t\}_{t \geq 1}$, is primitive, the path to perfect knowledge can start at any time. So, without loss of generality, we assume the output sequence which provides perfect state knowledge starts in any state, takes n steps, ends in state q, and occurs with probability δ. This means that $P_{iq}^{(n)}(x, z) \geq \delta$ for all $i \in \mathcal{Q}$ and all $x \in \mathfrak{D}(\mathcal{Q})$, which is also known as *Doeblin's Condition* [22, p. 391]. Applying Theorem 16.2.3 of [22], we find that the joint Markov chain is uniformly ergodic. □

This leads to the second condition under which the limit distribution, $\lim_{t \to \infty} \mu_t(s, A) = \pi(s, A)$, exists and is unique. This condition is essentially identical to the condition used by Le Gland and Mevel to prove weakly uniform ergodicity in [21].

Condition 4.4.8. Every output has positive probability during every transition. Mathematically, this means that $g_{ij}(y) > 0$ for all $y \in \mathbb{Y}$ and every i, j such that $p_{ij} > 0$. For example, any real output channel with AWGN satisfies this condition.

Since the joint Markov chain implied by Condition 4.4.8 does not, in general, satisfy a minorization condition [22, p. 102], we must turn to methods which exploit the continuity of $P_{ij}(x, A)$. We say that a general state space Markov chain is *(weak) Feller* if its transition kernel maps the set of bounded continuous functions (in some topology) to itself [22, p. 128]. Based on the properties of (4.4.7), one can verify that the joint Markov chain will be weak Feller as long as the minimum row sum of $\mathbf{M}(y)$ is strictly positive for all $y \in \mathbb{Y}$. Unfortunately, the methods of [22] still cannot be used to prove that the joint Markov chain is weakly uniform ergodic because its stationary distribution may not be absolutely continuous. In many cases, it will be singular continuous and concentrated on a set of dimension smaller than that of Ω. For simplicity, we simply adapt the results of [21] to our case. We note, however, that the results of iterated function systems (or iterated random functions) may also be applied to prove this result [12][29].

Theorem 4.4.9 (Le Gland-Mevel). *If Condition 4.4.8, then $\mu_\infty(s, A)$ exists and is unique. Furthermore, the joint Markov chain is weakly uniform ergodic.*

Proof. The analysis in [21] is applied to finite state processes whose output distribution is only a function of the initial state (i.e., $g_{ij}(y) = g_{ik}(y)$ for all j, k). There is a one-to-one correspondence between these two models, however. For example, one can map every transition in our model to a state in their model and represent the same process. Since $g_{ij}(y) > 0$ for all $y \in \mathbb{Y}$ and every i, j, we find that the output distribution of each state in their model will also be positive. Along with the ergodicity of the underlying FSMC, this gives the conditions necessary for Theorem 3.5 of [21]. Therefore, the joint Markov chain is weakly uniform ergodic. □

Now, we address the issue of CLTs for the entropy rate. For uniformly ergodic Markov chains, we use the CLT given by Chen in Theorem II-4.3 of [8]. This CLT is both very general and has the most easily verifiable conditions. For FSPs which satisfy Condition 4.4.8, we use the CLT given by Corollary 4.4.14. One could also prove this directly using the exponential decay of correlation implied by weakly uniform ergodicity, or alternatively, by using the theory of iterated function systems [4]. Unfortunately, all of these methods break down simultaneously if the product $\mathbf{M}(y_t)\mathbf{M}(y_{t+1})\cdots\mathbf{M}(y_{t+n})$ does not become strictly positive for some n.

Theorem 4.4.10 (Chen). *Let $\{X_t\}_{t\geq 1}$ be a uniformly ergodic Markov chain with unique stationary distribution $\pi(x)$. Let $f(x)$ be a measurable function and $S_n = \sum_{t=1}^{n} f(X_t)$. If we assume that (i) $E_\pi\left[f(X)\right] = 0$ and (ii) $E_\pi\left[f^2(X)\right] < \infty$, then*

$$S_n/\sqrt{n} \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 = R(0) + 2\sum_{\tau=1}^{\infty} R(\tau) < \infty$ and

$$R(\tau) = \lim_{t\to\infty} E\left[f(X_t)f(X_{t-\tau})\right].$$

Corollary 4.4.11. *Consider the FSP $\{Y_t\}_{t\geq 1}$ and its joint Markov chain $\left\{Q_t, \boldsymbol{\alpha}^{(t)}\right\}_{t\geq 1}$. Suppose (i) the process satisfies the finite variance condition $\lim_{t\to\infty} E\left[\left(\ln Pr(Y_t|\mathbf{Y}_1^{t-1})\right)^2\right]$ and (ii) the joint Markov chain satisfies Condition 4.4.6. In this case, the sample entropy rate, $\hat{H}_n(\mathcal{Y})$, obeys*

$$\sqrt{n}\left[\hat{H}_n(\mathcal{Y}) - H(\mathcal{Y})\right] \xrightarrow{d} N(0, \sigma^2),$$

where σ^2 is finite and given by (4.4.5).

Proof. Using (4.4.4), it is easy to see that $Pr(Y_t|Y_1^{t-1}) = f(Y_t, \boldsymbol{\alpha}^{(t)})$ for some measurable function, f. Now, we introduce the extended Markov chain, $\left\{Q_t, Y_t, \boldsymbol{\alpha}^{(t)}\right\}_{t\geq 1}$, since the function requires the Y_t value. Since the random variable Y_t is conditionally independent of all other quantities given Q_t and Q_{t+1}, it follows that the extended Markov chain inherits the ergodicity properties of the joint Markov chain. Since (i) implies that the joint Markov chain is uniformly ergodic and (ii) implies the finite variance condition of Theorem 4.4.10, we simply apply Theorem 4.4.10 to complete the proof. □

4.4.5 Entropy Rate Bounds

It is well known [10, Chap. 4] that the entropy rate of an ergodic FSP, $\{Y_t\}_{t\geq 1}$, is sandwiched between the Markov upper and lower bounds given by

$$H(Y_k|Y_{k-1}, Y_{k-2}, \ldots, Y_1, Q_1) \leq H(\mathcal{Y}) \leq H(Y_k|Y_{k-1}, Y_{k-2}, \ldots, Y_1). \tag{4.4.9}$$

In fact, Birch proves that the gap between these bounds decays exponentially in k for functions of a FSMC whose transition matrices are strictly positive [5]. The mixing properties of the underlying FSMC make it easy to believe that this gap actually decays exponentially for all FSPs.

Since all three of the quantities in (4.4.9) can be written as integrals over a state distribution of the joint Markov chain, we show that the gap decays to zero exponentially if the joint Markov chain is weakly uniform ergodic. Let $\mu_k(i, A)$ be the state distribution of the joint Markov chain. The entropy of Y_k can be written as a function of $\mu_k(i, A)$ with

$$H(Y_k|\mu_k) = \sum_{i=0}^{N_Q-1} \int_{\mathfrak{D}(\mathcal{Q})} \mu_k(i, dx) V(i, x),$$

where

$$V(i, \boldsymbol{\alpha}) = \int_{\mathbb{Y}} \sum_{j=0}^{N_Q-1} p_{ij} f_{ij}(y) \log \left(\sum_{m=0}^{N_Q-1} \sum_{l=0}^{N_Q-1} \alpha_l p_{lm} f_{lm}(y) \right) dy.$$

The function $V(i, \boldsymbol{\alpha})$ gives the entropy rate of Y conditioned on the true state being i and the state probability vector being $\boldsymbol{\alpha}$. While $V(i, \boldsymbol{\alpha})$ is unbounded as $\alpha_i \to 0$, it is a continuous function of $\boldsymbol{\alpha}$ as long as $\alpha_i > 0$. Fortunately, the probability, $Pr(\alpha_i^{(t)} = 0|Q_t = i)$, must be zero because events with probability zero cannot occur.

Let $\boldsymbol{\pi} = [\ \pi_1 \quad \pi_2 \quad \ldots \quad \pi_r\ ]$ be the unique stationary distribution of $\{Q_t\}_{t\geq 1}$ which satisfies $\pi_j = \sum_i \pi_i p_{ij}$. The lower bound, $H(Y_k|Y_{k-1}, Y_{k-2}, \ldots, Y_1, Q_1)$, is obtained by starting the chain with the distribution

$$\mu_1(i, \boldsymbol{\alpha}) = \begin{cases} \pi_i & \text{if } \alpha_i = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.4.10}$$

and taking k steps, because this initial condition corresponds to stationary Q-state probabilities and perfect state knowledge. The upper bound is obtained by starting the chain with the distribution

$$\mu_1(i, \boldsymbol{\alpha}) = \begin{cases} \pi_i & \text{if } \boldsymbol{\alpha} = \boldsymbol{\pi} \\ 0 & \text{otherwise} \end{cases} \tag{4.4.11}$$

and taking k steps because this initial condition corresponds to stationary Q-state probabilities and no state knowledge. The true entropy rate can be computed by using either initialization and letting $k \to \infty$, because all initial conditions eventually converge to unique stationary distribution $\mu_\infty(i, A)$.

If the joint Markov chain is (weakly) uniform ergodic, then the state distribution converges to $\mu_\infty(i, A)$ exponentially fast in k regardless of the initial conditions. Since the upper and lower bounds are only functions of the state distribution, we find that both of these bounds converge to the true entropy rate exponentially fast in k.

4.4.6 Connections with Lyapunov Exponents

Consider any stationary stochastic process, $\{Y_t\}_{t\geq 1}$, equipped with a function, $\mathbf{M}(y)$, that maps each $y \in \mathbb{Y}$ to an $r \times r$ matrix. Let $\mathbf{Z}(\mathbf{Y}_1^n) = \mathbf{M}(Y_1)\mathbf{M}(Y_2)\ldots\mathbf{M}(Y_n)$ be the cumulative product of random matrices generated by this process and let $\{\mathbf{y}_1^n\}_{n\geq 1}$ be a sequence of realizations with increasing length. Now, consider the limit

$$\lim_{n\to\infty} \frac{1}{n} \log \|\mathbf{x}\mathbf{Z}(\mathbf{y}_1^n)\|,$$

where $\mathbf{x}$ is any non-zero row vector and $\|\cdot\|$ is any vector norm. Oseledec's multiplicative ergodic theorem says that this limit is deterministic for almost all realizations [23]. While the proof takes a very different approach and is quite difficult, one way of thinking about this is that the matrix sequence, $\{\mathbf{Z}(\mathbf{y}_1^n)\}_{n\geq 1}$, can be associated with r eigenvalue sequences which grow (or decay) exponentially in n. The normalized exponential growth rate of each eigenvalue sequence almost surely has a deterministic limit known as the Lyapunov exponent. The Lyapunov spectrum is the ordered set of Lyapunov exponents, $\gamma_1 > \gamma_2 > \ldots > \gamma_s$, along with their multiplicities, $d_1, d_2, \ldots, d_s$. An earlier ergodic theorem due to Furstenberg and Kesten [13] gives a simple proof for the top Lyapunov exponent, and says that the limit

$$\lim_{n\to\infty} \frac{1}{n} \log \|\mathbf{Z}(\mathbf{Y}_1^n)\| = \gamma_1$$

convergences almost surely, where $\|\cdot\|$ is now taken to the matrix norm induced by the previous vector norm (see [18, p. 303]).

The connection between Lyapunov exponents and the entropy rate of a FSP is given by the following proposition.

Proposition 4.4.12. *The largest Lyapunov exponent, γ_1, of the product of the transition-observation matrices, $\mathbf{M}(y_1)\mathbf{M}(y_2)\ldots\mathbf{M}(y_n)$, is almost surely equal to $-H(\mathcal{Y})$, where $H(\mathcal{Y})$ is the entropy rate of the FSP, $\{Y_t\}_{t\geq 1}$. Mathematically, we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \|\mathbf{M}(Y_1)\mathbf{M}(Y_2)\ldots\mathbf{M}(Y_n)\| = \gamma_1 = -H(\mathcal{Y})$$

for almost all $\mathbf{Y}_1^n$.

Proof. Using (4.4.6), the probability $Pr(\mathbf{Y}_1^n)$ can be written in the form

$$Pr(\mathbf{Y}_1^n) = \sum_{i=1}^{r} \pi_i \sum_{j=1}^{r} [\mathbf{Z}(\mathbf{Y}_1^n)]_{ij} \,. \tag{4.4.12}$$

Applying the matrix norm induced (see [18, p. 303]) by the vector norm, $\|\cdot\|_\infty$, to $\mathbf{Z}(\mathbf{Y}_1^n)$ gives

$$\|\mathbf{Z}(\mathbf{Y}_1^n)\|_\infty = \max_i \sum_{j=1}^{r} [\mathbf{Z}(\mathbf{Y}_1^n)]_{ij} \,,$$

because our matrix is non-negative. Now, we can sandwich $Pr(\mathbf{Y}_1^n)$ with

$$\min_i \pi_i \|\mathbf{Z}(\mathbf{Y}_1^n)\|_\infty \leq Pr(\mathbf{Y}_1^n) \leq \|\mathbf{Z}(\mathbf{Y}_1^n)\|_\infty \tag{4.4.13}$$

by replacing the second sum in (4.4.12) by its maximum value to get an upper bound, and then applying the smallest π_i to that upper bound to get a lower bound. The ergodicity of the Markov chain $\{Q_t\}_{t\geq 1}$ implies that $\min_i \pi_i > 0$, and therefore the inequality (4.4.13) can be rewritten as

$$\frac{1}{n}\log\|\mathbf{Z}(\mathbf{Y}_1^n)\|_\infty \leq \frac{1}{n}\log Pr(\mathbf{Y}_1^n) \leq \frac{1}{n}\log\|\mathbf{Z}(\mathbf{Y}_1^n)\|_\infty + \frac{\log\min_i \pi_i}{n}.$$

Finally, this shows that

$$\lim_{n\to\infty}\frac{1}{n}\log\|\mathbf{Z}(\mathbf{Y}_1^n)\|_\infty = \lim_{n\to\infty}\frac{1}{n}\log Pr(\mathbf{Y}_1^n),$$

which completes the proof. □

The following is a restatement of Theorem 7 from [9] and provides a CLT for the largest Lyapunov exponent.

Theorem 4.4.13 (Cohn-Nermann-Peligrad). *Suppose that $\{\mathbf{M}(Y_t)\}_{t\geq 1}$ is a strictly stationary sequence of $r \times r$ non-negative matrices satisfying: (i) there exists an integer n_0 such that $\mathbf{M}(Y_t)\mathbf{M}(Y_{t+1})\ldots\mathbf{M}(Y_{t+n_0})$ is positive with probability 1, (ii) the process $\{Y_t\}_{t\geq 1}$ is geometrically ergodic, and (iii) $E\left[\min^+\left(\log\left[\mathbf{M}(Y_t)\right]\right)^2\right] < \infty$ and $E\left[\max^+\left(\log\left[\mathbf{M}(Y_t)\right]\right)^2\right] < \infty$ where $\min^+\left(\mathbf{M}(Y_t)\right)$ and $\max^+\left(\mathbf{M}(Y_t)\right)$ are the minimum and maximum over the strictly positive elements of $\mathbf{M}(Y_t)$. Then there exists a $\sigma \geq 0$ such that*

$$n^{-1/2}\left[\log\left|\left[\mathbf{M}(Y_1)\mathbf{M}(Y_2)\ldots\mathbf{M}(Y_n)\right]_{ij}\right| - n\gamma_1\right] \xrightarrow{d} N(0,\sigma),$$

converges in distribution and γ_1 is the largest Lyapunov exponent.

The following Corollary provides a CLT for the entropy rate under conditions which are implied by Condition 4.4.8 and a finite variance condition.

Corollary 4.4.14. *Suppose that every observation, $y \in \mathbb{Y}$, is possible during every transition. This implies that $g_{ij}(y) > 0$ for all i,j such that $p_{ij} > 0$. Furthermore, suppose that condition (iii) of Theorem 4.4.13 holds. Then the sample entropy, $\hat{H}_n(\mathcal{Y})$, is asymptotically Gaussian with asymptotic mean $H(\mathcal{Y})$.*

Remark 4.4.15. Using (4.4.7) and the second Lyapunov exponent of the matrix $\mathbf{Z}(\mathbf{Y}_1^n)$, we can also consider the exponential rate at which $\boldsymbol{\alpha}^{(t)}$ forgets its initial condition $\boldsymbol{\alpha}^{(1)}$. Consider the scaled matrix product, $\mathbf{Z}(\mathbf{Y}_1^n)/\left\|\mathbf{Z}(\mathbf{Y}_1^n)\right\|_\infty$, whose maximal row sum will always equal one. The normalized second largest eigenvalue, $|\lambda_2|^{1/n}$, of this scaled matrix product will almost certainly be equal to $e^{\gamma_2-\gamma_1}$. This is because the normalized eigenvalues of $\mathbf{Z}(\mathbf{Y}_1^n)$ will almost surely be given by the Lyapunov spectrum. Therefore, if the largest Lyapunov exponent is simple (i.e., its multiplicity is one), then $\boldsymbol{\alpha}^{(t)}$ forgets its initial condition almost surely at the positive exponential rate given by $\gamma_2 - \gamma_1$. It is important to note that this is the expected rate at which $\boldsymbol{\alpha}^{(t)}$ forgets its initial condition. This does not necessarily imply that the probability of rare events also decays exponentially.

4.5 Capacity Bounds

The capacity of a FSC is given by

$$C = \lim_{n\to\infty} \frac{1}{n} \max_{Pr(X_1^n)} I(X_1^n; Y_1^n),$$

where the limit always exists and is independent of the initial state [14, Chap. 4]. In terms of mutual information rates, this capacity can also be written as

$$C = \sup_{\mathcal{X}} \left[H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X},\mathcal{Y}) \right],$$

where the supremum is taken over all stationary ergodic input processes. Unfortunately, the maximization implied by either formula is over an infinite dimensional distribution and impossible to carry out in practice. The capacity can be sandwiched between two computable quantities, however. Using upper and lower bounds on the entropy rates of the FSPs, we illustrate this in Sections 4.5.1, 4.5.2, and 4.5.3.

4.5.1 Lower Bounds

Lower bounds on the capacity are actually quite straightforward to compute because any achievable rate is a lower bound on the capacity. For example, we consider the maximum rate achievable using Markov input distributions with memory η. These distributions have a simple representation because $Pr(X_i|\mathbf{X}_1^{i-1}) = Pr(X_i|\mathbf{X}_{i-\eta}^{i-1})$. Let M_η be the set of all such input distributions Then the sequence $\left\{\underline{C}_\eta\right\}_{\eta\geq 0}$, defined by

$$\underline{C}_\eta = \lim_{n\to\infty} \frac{1}{n} \max_{Pr(\mathbf{X})\in M_\eta} I(\mathbf{X}_1^n; \mathbf{Y}_1^n),$$

is a sequence of lower bounds on the capacity. The sequence of bounds is non-decreasing because any Markov input process in M_η is also in $M_{\eta+1}$. We also note that the information rate, C_η, is referred to as the Markov-η rate of the channel.

Using standard optimization techniques these bounds were computed numerically for linear ISI channels with Gaussian noise in [24] and [1]. The results given in Section 4.6, however, were generated more accurately and efficiently using Kavčić's elegant generalization of the Arimoto-Blahut algorithm [19]. While no rigorous proof exists for the convergence of this algorithm, theoretical and numerical results strongly imply its correctness. In particular, it always returns a valid information rate and, in all cases tested, it gives results numerically equivalent to standard optimization techniques.

4.5.2 The Vontobel-Arnold Upper Bound

Upper bounds on the capacity are somewhat more difficult to compute because the maximization over all input distributions must be treated very carefully. Vontobel and Arnold propose an upper bound on the capacity of finite state channels in [31]. The first step in this upper bound can be seen as a generalization of the standard upper bound [14, Theorem 4.5.1] for DMCs. Let $I_P(X;Y)$ be the mutual information between the inputs and outputs of a DMC for some input distribution $P(x)$. Then, for any fixed channel (i.e., fixed $Pr(Y|X)$), the upper bound states that

$$C = \max_{P_0(x)} I_{P_0}(X;Y) \le \max_x I_{P_1}(X=x;Y), \tag{4.5.1}$$

where

$$I_P(X=x;Y) = E\left[\log \frac{Pr(Y|X)}{\sum_{x'\in\mathbb{X}} Pr(Y|X=x')P(x')} \middle| X=x \right].$$

The natural generalization of this upper bound to channels with memory implies that

$$C = \lim_{n\to\infty} \frac{1}{n} \max_{P_0(\mathbf{x}_1^n)} I_{P_0}(\mathbf{X}_1^n;\mathbf{Y}_1^n) \le \lim_{n\to\infty} \frac{1}{n} \max_{x_1^n} I_{P_1}(\mathbf{X}_1^n = \mathbf{x}_1^n;\mathbf{Y}_1^n) \tag{4.5.2}$$

for a fixed channel (i.e., fixed $Pr(\mathbf{Y}_1^n|\mathbf{X}_1^n)$) and any $P_1(\mathbf{X}_1^n)$. Vontobel and Arnold start by noting that

$$C \le \lim_{n\to\infty} \frac{1}{n} \max_{\mathbf{x}_1^n} E\left[\log \frac{Pr(\mathbf{Y}_1^n|\mathbf{X}_1^n)}{R(\mathbf{Y}_1^n)} \middle| \mathbf{X}_1^n = \mathbf{x}_1^n \right] \tag{4.5.3}$$

holds for any distribution $R(\mathbf{Y}_1^n)$. By choosing an $R(\mathbf{Y}_1^n)$ which can be factored according to

$$R(\mathbf{y}_1^n) = R_1^L(\mathbf{y}_1^L) \prod_{i=L+1}^{n} R(y_i|\mathbf{y}_{i-L}^{i-1}),$$

they are able to make this bound computable as well. This distribution, $R(Y_i|\mathbf{Y}_{i-L}^{i-1})$, is generally chosen to be the marginal distribution, $Pr(Y_i|\mathbf{Y}_{i-L}^{i-1})$, because this choice minimizes, for any given L, the quantity

$$E\left[\log \frac{Pr(\mathbf{Y}_1^n|\mathbf{X}_1^n)}{R(\mathbf{Y}_1^n)}\right].$$

Their method of making the bound computable is actually quite clever. It is based upon writing the conditional expectation of (4.5.3) in a form which makes the maximization easy. For FSCs whose state is defined by the previous ν inputs (e.g., any linear ISI channel), we can write

$$E\left[\log \frac{Pr(\mathbf{Y}_1^n|\mathbf{X}_1^n)}{R(\mathbf{Y}_1^n)}\middle|\mathbf{X}_1^n = \mathbf{x}_1^n\right] = K_0(\mathbf{x}_1^{L+\nu}) + E\left[\sum_{i=L+\nu+1}^{n} \log \frac{Pr(Y_i|\mathbf{X}_{i-\nu}^i)}{R(Y_i|\mathbf{Y}_{i-L}^{i-1})}\middle|\mathbf{X}_1^n = \mathbf{x}_1^n\right],$$

where $K_0(\mathbf{x}_1^{L+\nu})$ absorbs the contribution of the neglected $L+\nu$ initial terms of the sum. The conditional expectation of the ith term in the sum only requires knowledge of $\mathbf{x}_{i-L-\nu}^i$. So, using the definition

$$K(\mathbf{x}) = E\left[\log \frac{Pr(Y_i|\mathbf{X}_{i-\nu}^i)}{R(Y_i|\mathbf{Y}_{i-L}^{i-1})}\middle|\mathbf{X}_{i-L-\nu}^i = \mathbf{x}\right], \tag{4.5.4}$$

we have

$$E\left[\log \frac{Pr(\mathbf{Y}_1^n|\mathbf{X}_1^n)}{R(\mathbf{Y}_1^n)}\middle|\mathbf{X}_1^n = \mathbf{x}_1^n\right] = K_0(\mathbf{x}_1^{L+\nu}) + \sum_{i=L+\nu+1}^{n} K(\mathbf{x}_{i-L-\nu}^i).$$

Computing the function $F(\mathbf{x}_1^n) = K_0(\mathbf{x}_1^{k_0}) + \sum_{i=k_0+1}^n K(\mathbf{x}_{i-i_0}^i)$ is equivalent to computing the weight of a path (labeled by $\mathbf{x}_1^n$) through an edge-weighted directed graph. Therefore, the Viterbi algorithm can be used to find the maximum of the function over $\mathbf{x}_1^n$. The quantity we actually want, however, is the limiting value

$$\lim_{n\to\infty} \frac{1}{n} \max_{\mathbf{x}_1^n} F(\mathbf{x}_1^n).$$

In the literature, finding this quantity is known as the *minimum mean cycle* problem [11]. The connection is based on fact that the maximum (or minimum) average weight path spends most of its time walking the same maximum (or minimum) average weight cycle repeatedly. Therefore, the answer is given by the maximum (or minimum) average cycle weight of the graph.

The practical problem of computing the function value, $K(\mathbf{x}_{i-L-\nu}^{i})$, for each binary $(L+\nu)$-tuple can be solved by using a simulation to estimate the expectation in (4.5.4). The complexity of this method linear in the simulation length and exponential in $L+\nu$ because we run one simulation for each $\mathbf{x}$. In some cases, the trade off between complexity and estimation error may also be reduced by using one long simulation and using Bayes' rule to write the conditional expectation as

$$K(\mathbf{x}) = E\left[\frac{Pr(\mathbf{X}_{i-L-\nu}^{i} = \mathbf{x}|\mathbf{Y}_{i-L}^{i-1})}{Pr(\mathbf{X}_{i-L-\nu}^{i} = \mathbf{x})} \log \frac{Pr(Y_i|\mathbf{X}_{i-L-\nu}^{i} = \mathbf{x})}{R(Y_i|\mathbf{Y}_{i-L}^{i-1})}\right].$$

The major drawback of the Vontobel-Arnold Bound is that it can be quite loose for channels whose state sequence is not identifiable from a small number of samples. Consider, for example, a dicode channel with very little noise. The lprobability of observing either a positive or negative transition, after observing L samples near zero, remains large enough to weaken the bound significantly. One might think that the small probability of observing long runs of zeroes at the output would counteract this problem. This is not the case, however, because the maximization over $\mathbf{x}_1^n$ picks the worst-case sequence, regardless of its probability.

4.5.3 A Conjectured Upper Bound

Now, we derive a slightly different expression that we conjecture is also an upper bound on the capacity of IFSCs. This bound also starts with (4.5.2), but uses different simplifications to make the bound computable. We start by considering an IFSC channel, with state sequence $\mathbf{S}_1^n$, driven by a Markov input process with memory $\eta \leq L$. The channel and input state can be combined into a single state variable, and that corresponding state sequence is $\mathbf{Q}_1^n$. The basic idea is that we can upper bound the mutual information by considering a genie-aided decoder which has perfect knowledge of the states Q_{i-L} and Q_{i+L+1} when it is decoding the input X_i. This expression remains a conjectured upper bound because of a subtle gap that remains in our proof.

We begin by upper bounding $I(\mathbf{X}_1^n;\mathbf{Y}_1^n)$ using the chain rule for mutual information and a genie-aided decoder. The chain rule gives

$$I(\mathbf{X}_1^n;\mathbf{Y}_1^n)=\sum_{i=1}^{n} I(X_i;\mathbf{Y}_1^n|\mathbf{X}_1^{i-1})$$

and, neglecting edge effects, the genie-aided upper bound for each term in the sum is given by

$$I(X_i;\mathbf{Y}_1^n|\mathbf{X}_1^{i-1})\leq I(X_i;Q_{i-L},\mathbf{Y}_1^n,Q_{i+L+1}|\mathbf{X}_1^{i-1}).$$

For any input distribution, $P(\mathbf{x}_1^n)$, we define the genie-aided mutual information to be

$$J_P(\mathbf{X}_1^n;\mathbf{Y}_1^n)=\sum_{\mathbf{x}_1^n} P(\mathbf{x}_1^n)J_P(\mathbf{X}_1^n=\mathbf{x}_1^n;\mathbf{Y}_1^n), \tag{4.5.5}$$

where $J_P(\mathbf{X}_1^n=\mathbf{x}_1^n;\mathbf{Y}_1^n)$ is defined with

$$J_P(\mathbf{X}_1^n=\mathbf{x}_1^n;\mathbf{Y}_1^n)=\sum_{i=1}^{n} E\left[\log\frac{Pr(X_i|\mathbf{X}_1^{i-1},Q_{i-L},\mathbf{Y}_1^n,Q_{i+L+1})}{Pr(X_i|\mathbf{X}_1^{i-1})}\middle|\mathbf{X}_1^n=\mathbf{x}_1^n\right].$$

The Markov nature of the input distribution and the channel allow us to write

$$J_P(\mathbf{X}_1^n=\mathbf{x}_1^n;\mathbf{Y}_1^n)=\sum_{i=1}^{n} E\left[\log\frac{Pr(X_i|\mathbf{X}_{i-L}^{i-1},Q_{i-L},\mathbf{Y}_{i-L}^{i+L},Q_{i+L+1})}{Pr(X_i|\mathbf{X}_{i-\eta}^{i-1})}\middle|\mathbf{X}_{i-L}^{i+L}=\mathbf{x}_{i-L}^{i+L}\right],$$

and this provides an upper bound on $I(\mathbf{X}_1^n;\mathbf{Y}_1^n)$ which is computed as the sum of local functions.

The obvious generalization of the Vontobel-Arnold Bound would imply that $C\leq \lim_{n\to\infty}\max_{\mathbf{x}_1^n} J_P(\mathbf{X}_1^n=\mathbf{x}_1^n;\mathbf{Y}_1^n)$. Unfortunately, this does not follow directly from the results that $C\leq \lim_{n\to\infty}\max_{\mathbf{x}_1^n} I(\mathbf{X}_1^n=\mathbf{x}_1^n;\mathbf{Y}_1^n)$ and $I_P(\mathbf{X}_1^n;\mathbf{Y}_1^n)\leq J_P(\mathbf{X}_1^n;\mathbf{Y}_1^n)$. This is because it is possible that the the genie-aided mutual information could be larger when averaged over all sequences, even though its largest per-sequence value is smaller. Our proof of this bound requires that the chain of inequalities,

$$\max_{P_0(\mathbf{x}_1^n)} I_{P_0}(\mathbf{X}_1^n;\mathbf{Y}_1^n)\leq \max_{P_1(\mathbf{x}_1^n)} J_{P_1}(\mathbf{X}_1^n;\mathbf{Y}_1^n)\leq \max_{\mathbf{x}_1^n} J_{P_2}(\mathbf{X}_1^n=\mathbf{x}_1^n;\mathbf{Y}_1^n),$$

holds for all $P_2(\mathbf{x}_1^n)$. The LHS inequality indeed holds because $J_P(\mathbf{X}_1^n;\mathbf{Y}_1^n)$ is an upper bound on $I_P(\mathbf{X}_1^n;\mathbf{Y}_1^n)$ for all $P(\mathbf{x}_1^n)$. We conjecture that the RHS inequality holds as well. Numerical results for the dicode channel are encouraging, however, because our lower bounds on the capacity are quite close to the conjectured upper bound but do not surpass it.

Following the Vontobel-Arnold approach, we make the conjectured upper bound computable by writing it as the sum of local functions. For ISI channels whose state sequence is a deterministic function of the input sequence, we can simplify the function $J_P(\mathbf{X}_1^n = \mathbf{x}_1^n; \mathbf{Y}_1^n)$ to

$$F(\mathbf{x}_1^n) = \sum_{i=1}^{n} E\left[\log \frac{Pr(X_i|Q_{i-1}, \mathbf{Y}_i^{i+L}, Q_{i+L+1})}{Pr(X_i|\mathbf{X}_{i-\eta}^{i-1})}\middle|\mathbf{X}_{i-\eta}^{i+L} = \mathbf{x}_{i-\eta}^{i+L}\right].$$

The simplification uses the facts that X_i is conditionally independent of the past given Q_{i-1} and Q_{i-1} is computable from Q_{i-L} and $\mathbf{X}_{i-L}^{i-1}$. Let ν be the memory of the channel, η be the memory of the input, and assume that $\eta \geq \nu$ (without loss of generality). This allows us to write

$$F(\mathbf{x}_1^n) = K_0(\mathbf{x}_1^{L+\eta}) + \sum_{i=\eta+L+1}^{n-L} K(\mathbf{x}_{i-\eta}^{i+L}) + K_1(\mathbf{x}_{n-\eta-L}^{n}),$$

where

$$K(\mathbf{x}) = E\left[\log \frac{Pr(X_i|Q_{i-1}, \mathbf{Y}_i^{i+L}, Q_{i+L+1})}{Pr(X_i|\mathbf{X}_{i-\eta}^{i-1})}\middle|\mathbf{X}_{i-\eta}^{i+L} = \mathbf{x}\right].$$

The terms $K_0(\mathbf{x}_1^{L+\eta})$ and $K_1(\mathbf{x}_{n-\eta-L}^{n})$ are asymptotically irrelevant and will be ignored. Since $F(\mathbf{x}_1^n) = J_P(\mathbf{X}_1^n = \mathbf{x}_1^n; \mathbf{Y}_1^n)$, the conjectured upper bound on capacity is given by

$$C \leq \lim_{n\to\infty} \frac{1}{n} \max_{\mathbf{x}_1^n} F(\mathbf{x}_1^n).$$

Once again, the function $F(\mathbf{x}_1^n)$ can be computed as the weight of a path (labeled by $\mathbf{x}_1^n$) through an edge-weighted directed graph. In this case, the states are labeled by $\mathbf{x}_{i-\eta}^{i+L-1}$ and the edge weights are estimated by stochastic averaging of the expectation in $K(\mathbf{x})$. As with the Vontobel-Arnold bound, the conjectured upper bound is given by the maximum average cycle weight of the graph.

4.6 Monte Carlo Results

4.6.1 Partial Response Channels

We start by giving the results for the power normalized binary-input channels listed in Table 4.1. These channels are known as partial response (PR) channels and are sometimes used to model magnetic recording channels. First, we show the SIR of these channels along with binary-input AWGN capacity in Figure 4.6.1. For each channel, the achievable rate is plotted

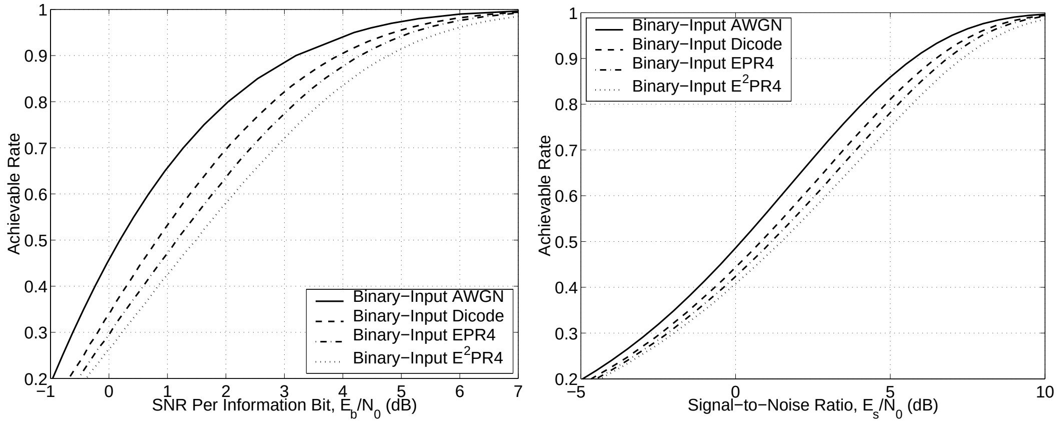

Figure 4.6.1: The SIR for various partial response channels, estimated with $n = 10^7$.

versus both the SNR (E_s/N_0) and the SNR per information bit (E_b/N_0). Notice that increasing the severity of the ISI monotonically decreases the SIR in both cases.

Next, we consider the results attained by optimizing the input distributions for these channels. The elegant generalization of the Arimoto-Blahut algorithm due to Kavčić is used for all of these results [19]. Figure 4.6.2 shows the results for the dicode channel and plots the achievable rate versus E_s/N_0 and E_b/N_0. Figure 4.6.3 shows the results of optimizing input distributions for the EPR4 channel. All of these results show that optimizing the input distribution provides significant gains at low SNR.

Figure 4.6.4 compares the dicode channel lower bound with the Vontobel-Arnold Bound and our own conjectured upper bound. We would like to acknowledge P. Vontobel for providing the data points for the Vontobel-Arnold Bound. Both upper bounds are somewhat loose at low rates, but the conjectured upper bound is actually quite tight at high rates. The conjectured upper bound has an intrinsic advantage at high rates because it is always upper bounded entropy of the input process. We also note that this comparison is not entirely fair because the conjectured upper bound was numerically optimized over the input distribution while the Vontobel-Arnold Bound was not. In fact, without optimization the performance of the conjectured upper bound at low rates does not surpass the Vontobel-Arnold Bound. In the future, we plan to make a fair comparison by optimizing the Vontobel-Arnold Bound as well.

It is worth noting that at low enough SNR, all of the optimized rates actually exceed the capacity of the binary-input AWGN channel. Depending on your perspective, this is either

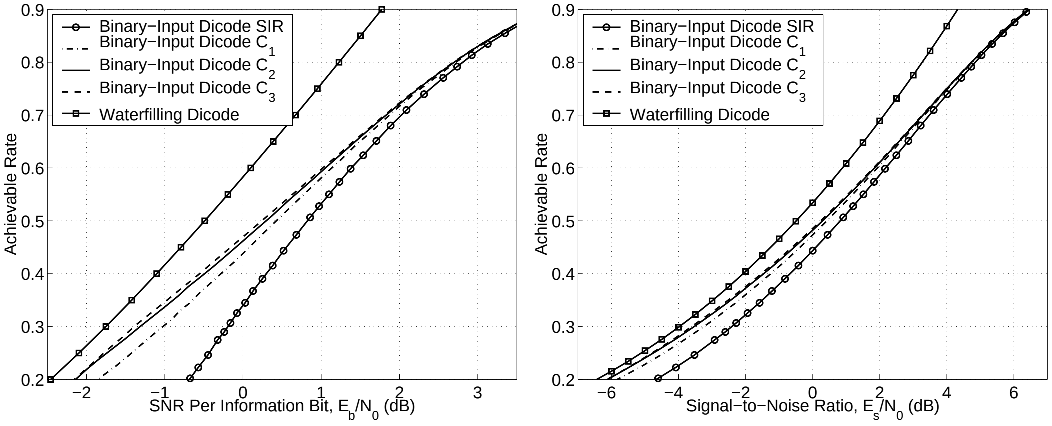

Figure 4.6.2: Monte Carlo lower bounds on the achievable information rate of the dicode channel using optimized Markov input distributions.

an interesting phenomenon or simply a poorly chosen channel normalization. The basic problem is that there is no single normalization which is fair. By convention, we normalize each channel so that the power of a white input signal is unchanged by the channel. This approach seems fair for white input signals such as equiprobable binary inputs. When the channel response is not flat, however, optimizing the input distribution allows the source to concentrate its power around the peaks of the channel response. Now, the signal appears to be receiving a *power gain* from the channel. One solution is to normalize all channels so that the peak of the response is unity. This is merely a different convention, however, and regardless of the chosen normalization, optimizing input distribution will always increase the power output of the channel.

Finally, we remark that at low rates these binary-input ISI channels exhibit a threshold behavior similar to the $-1.59\,dB$ limit of the AWGN channel. Essentially, this means that there is an E_b/N_0 threshold below which reliable communication is impossible. Conversely, it can be shown that, at sufficiently low rates, reliable communication is also possible at any E_b/N_0 larger than this threshold. The threshold is known as the low-rate Shannon limit and is discussed in [28].

4.6.2 The Finite State Z-Channel

The SIR and the Markov-1 rate of the finite state Z-channel are shown in Figure 4.6.5. It is interesting to note that, as with the original Z-channel, one can avoid transmission errors by

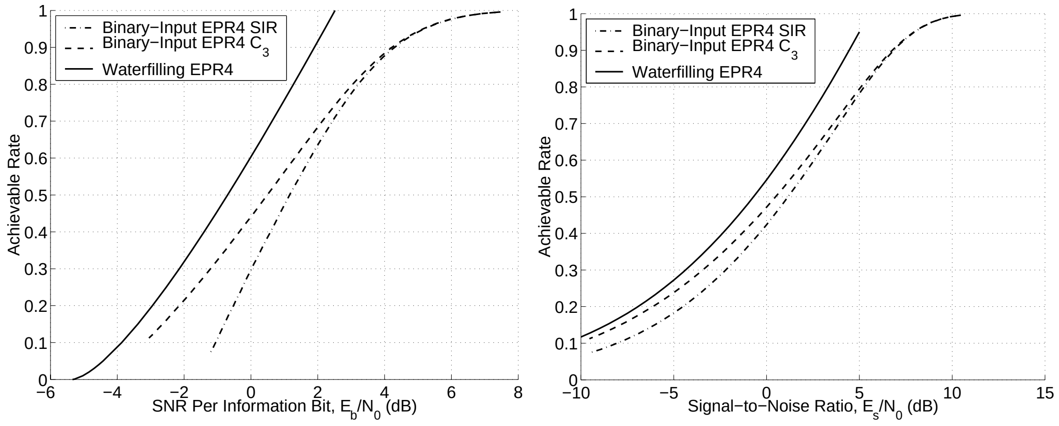

Figure 4.6.3: Monte Carlo lower bounds on the achievable information rate of the EPR4 channel using optimized Markov input distributions.

sending a particular pattern. (e.g., 0101...01). So, while optimizing the input distribution to the dicode channel increases the output power, optimizing the input distribution to the finite state Z-channel decreases the noise. Therefore, the optimized input distribution chooses transitions (i.e., edges from state 0 to state 1 and vice-versa) more frequently than non-transitions (i.e., edges corresponding to self-loops) and thereby incurs fewer channel errors.

4.7 Analytical Results

In this section, we consider analytical methods for computing achievable information rates. We start by using the results of Section 4.4.4 to compute exact information rates for the DEC. This analysis of the DEC is made possible by the fact that the stationary distribution of the joint Markov chain is supported on a countably infinite set. This follows from Theorem 4.4.7 because the reception of a $+$ or $-$ symbol gives the observer perfect state knowledge. Next, we describe a pseudo-analytical method based on density evolution that can be used to estimate rates more efficiently for arbitrary two state channels.

4.7.1 The Symmetric Information Rate of the DEC

Consider a DEC with erasure probability ϵ and equiprobable inputs. Let $\mathbf{X}_1^n$ be the channel input sequence, $\mathbf{S}_1^{n+1}$ be the channel state sequence, and $\mathbf{Y}_1^n$ be the channel output sequence. The input and output alphabets of the DEC are defined so that $X_i \in \{0, 1\}$ and

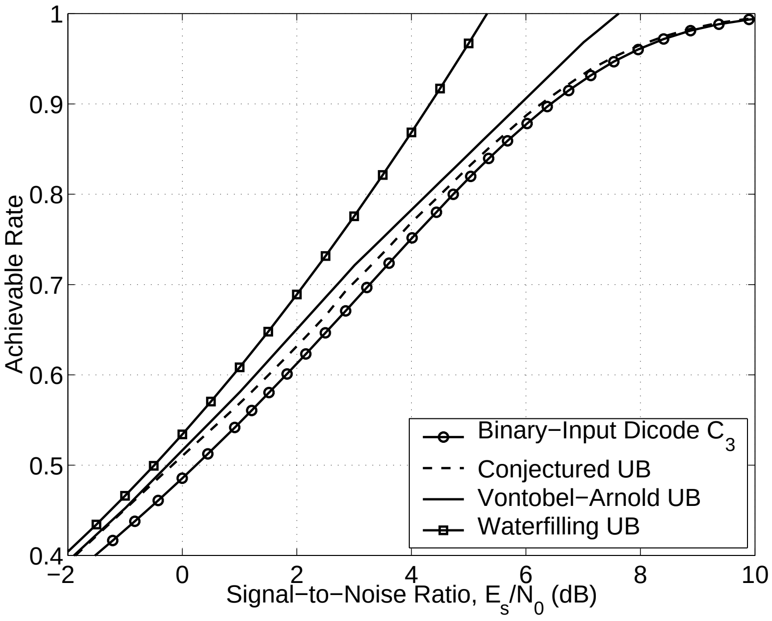

Figure 4.6.4: Monte Carlo upper and lower bounds on the achievable information rate of the dicode channel using optimized Markov input distributions.

$Y_i \in \{-,0,+,e\}$. In this section, we compute the exact SIR analytically by characterizing the forward recursion of the APP algorithm, which computes $Pr(Y_t|Y_1^{t-1})$, in terms of the random variable $\alpha_i^{(t)} = Pr(S_t = i|Y_1^t)$. Parts of this analysis were motivated by a method used to compute iterative decoding thresholds for turbo codes on the binary erasure channel [30].

Since the channel has only two states, it suffices to consider the quantity $\alpha^{(t)} \triangleq \alpha_0^{(t)} = 1 - \alpha_1^{(t)}$. The real simplification, however, comes from the fact that the distribution of $\alpha^{(t)}$ has finite support when $X \sim B(1/2)$. We can observe this fact by writing the APP recursion as (4.4.7) where $\boldsymbol{\alpha}^{(t)} = \begin{bmatrix}\alpha^{(t)} & 1-\alpha^{(t)}\end{bmatrix}$ and

$$\mathbf{M}(e)=\begin{bmatrix} \frac{\epsilon}{2} & \frac{\epsilon}{2} \\ \frac{\epsilon}{2} & \frac{\epsilon}{2} \end{bmatrix}, \mathbf{M}(0)=\begin{bmatrix} \frac{1-\epsilon}{2} & 0 \\ 0 & \frac{1-\epsilon}{2} \end{bmatrix}, \mathbf{M}(+)=\begin{bmatrix} 0 & \frac{1-\epsilon}{2} \\ 0 & 0 \end{bmatrix}, \mathbf{M}(-)=\begin{bmatrix} 0 & 0 \\ \frac{1-\epsilon}{2} & 0 \end{bmatrix}.$$

Using this, it is easy to verify that we can use instead the simpler recursion,

$$\alpha^{(t+1)} = \begin{cases} 1/2 & \text{if } Y_t = e \\ \alpha^{(t)} & \text{if } Y_t = 0 \\ 0 & \text{if } Y_t = + \\ 1 & \text{if } Y_t = - \end{cases}.$$

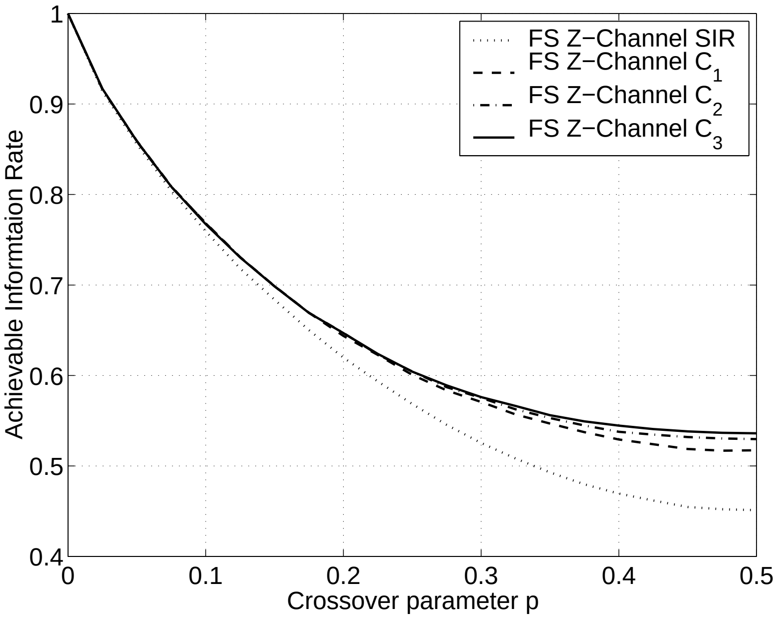

Figure 4.6.5: The SIR and the Markov-1 rate of the finite state Z-channel.

Using this, we see that, for all $t \geq \min\{i \geq 1 | Y_i \neq 0\}$, $\alpha^{(t)}$ will be confined to the finite set $\{0, 1/2, 1\}$.

With the results of Section 4.4.4 in mind, we proceed by finding the stationary distribution of the joint Markov chain, $\left\{Q_t, \boldsymbol{\alpha}^{(t)}\right\}_{t\geq 1}$. Each state of this Markov chain can be indexed by the pair $(q_t, \alpha^{(t)})$, and the stationary distribution is supported on the set $(0,1)$, $(1,0)$, $(0,1/2)$, and $(1,1/2)$. The first two states correspond to the known (K) state condition, while the second two correspond to the unknown (U) state condition. The symmetry of the problem allows us to write $Pr(K)/2 = \pi(0,1) = \pi(1,0)$ and $Pr(U)/2 = \pi(0,1/2) = \pi(1,1/2)$.

Using a two state Markov chain, we can compute the steady state probabilities of the joint Markov chain. First, we note that the joint Markov chain transitions from the known state condition to the unknown state condition only if $Y = e$. Therefore, we have $Pr(K \rightarrow U) = 1 - Pr(K \rightarrow K) = \epsilon$. Secondly, we note that the joint Markov chain transitions from the unknown state condition to the known state condition only if $Y = +$ or $Y = -$. This means that we have $Pr(U \rightarrow K) = 1 - Pr(U \rightarrow U) = (1-\epsilon)/2$. The steady state probabilities $Pr(K)$

and $Pr(U)$ can be found using the eigenvector equation,

$$\left[\begin{array}{cc} Pr(K) & Pr(U) \end{array}\right]\left[\begin{array}{cc} 1-\epsilon & \epsilon \\ \frac{1-\epsilon}{2} & \frac{1+\epsilon}{2} \end{array}\right] = \left[\begin{array}{cc} Pr(K) & Pr(U) \end{array}\right],$$

whose solution is $Pr(U) = 1 - Pr(K) = 2\epsilon/(1+\epsilon)$.

Now, we can compute the exact entropy rate of Y_1^n using the definition $H(\mathcal{Y}) = \lim_{t\to\infty} H(Y_t|Y_1^{t-1})$. When the joint Markov chain is in a known state, the observer knows that one of only two edges can be traversed in the next step. In this case, the conditional entropy of the next output given the past is denoted $H(Y|K)$ and is given by

$$H(Y|K) = H\left(\left[\epsilon, \frac{1-\epsilon}{2}, \frac{1-\epsilon}{2}\right]\right) = h(\epsilon) + (1-\epsilon),$$

where $H\left([p_1,\ldots,p_k]\right) = -\sum_{i=1}^{k} p_i \log_2 p_i$ and $h(\epsilon) = -\epsilon\log_2\epsilon - (1-\epsilon)\log_2(1-\epsilon)$. When the joint Markov chain is in an unknown state, the observer must allow the possibility of that any of four edges may actually be traversed in the next step. In this case, the conditional entropy of the next output given the past is denoted $H(Y|U)$ and is given by

$$H(Y|U) = H\left(\left[\epsilon, \frac{1-\epsilon}{4}, \frac{1-\epsilon}{2}, \frac{1-\epsilon}{4}\right]\right) = h(\epsilon) + \frac{3(1-\epsilon)}{2}.$$

Since the stationary distribution of the joint Markov chain determines how often each of these events occurs, we can write

$$H(\mathcal{Y}) = Pr(K)H(Y|K) + Pr(U)H(Y|U).$$

Substituting exact values into this expression and simplifying gives

$$H(\mathcal{Y}) = 1 - \frac{2\epsilon^2}{1+\epsilon} + h(\epsilon).$$

Since the entropy rate of $\mathbf{Y}_1^n$ given $\mathbf{X}_1^n$, $H(\mathcal{Y}|\mathcal{X})$, is simply the entropy rate of the erasure process (i.e., $h(\epsilon)$), the SIR is given by

$$H(\mathcal{Y}) - H(\mathcal{Y}|\mathcal{X}) = H(\mathcal{Y}) - h(\epsilon) = 1 - \frac{2\epsilon^2}{1+\epsilon}.$$

4.7.2 The Markov-1 Rate of the DEC

In this section, we derive the achievable information rate of the DEC using a Markov-1 input distribution. Based on the symmetry of the channel, we use an input distribution which changes state with probability p and remains in the same state with probability $1-p$. The combined state space of the input process and the channel still only has two states, so again it suffices to consider only $\alpha^{(t)} \triangleq \alpha_0^{(t)} = 1-\alpha_1^{(t)}$.

In this case, we can write the APP recursion as (4.4.7) where $\boldsymbol{\alpha}^{(t)} = \left[\alpha^{(t)} \;\; 1-\alpha^{(t)}\right]$,

$$\mathbf{M}(e) = \begin{bmatrix} (1-p)\epsilon & p\epsilon \\ p\epsilon & (1-p)\epsilon \end{bmatrix}, \mathbf{M}(0) = \begin{bmatrix} (1-p)(1-\epsilon) & 0 \\ 0 & (1-p)(1-\epsilon) \end{bmatrix},$$

$$\mathbf{M}(+) \begin{bmatrix} 0 & p(1-\epsilon) \\ 0 & 0 \end{bmatrix}, \text{ and } \mathbf{M}(-) = \begin{bmatrix} 0 & 0 \\ p(1-\epsilon) & 0 \end{bmatrix}.$$

A simple recursion also exists in this case and is given by

$$\alpha^{(t+1)} = \begin{cases} \alpha^{(t)}(1-p) + (1-\alpha^{(t)})p & \text{if } Y_t = e \\ \alpha^{(t)} & \text{if } Y_t = 0 \\ 0 & \text{if } Y_t = + \\ 1 & \text{if } Y_t = - \end{cases}.$$

The major complication in completing the analysis is the fact that the support set of $\alpha^{(t)}$ is now countably infinite. For example, the $\alpha^{(t)}$ that results from observing a $-$ first and then observing a mixture of k erasures and any number of 0's (but no more $+$'s and $-$'s) is given by $\left(1+(1-2p)^k\right)/2$. Likewise, if the first observation was a $+$, then we would have $\left(1-(1-2p)^k\right)/2$. These two cases, with $k \in \{0,\ldots,\infty\}$, constitute the entire support set of $\alpha^{(t)}$. For simplicity, we refer to these values using the shorthand,

$$\gamma_k^{\pm} = \frac{1 \pm (1-2p)^k}{2}. \tag{4.7.1}$$

Now, we define the countably infinite Markov chain that will be used to help analyze the joint Markov chain. Each state in the new Markov chain is labeled by a letter (A or B) and a non-negative integer. The A_k state corresponds to the event that $(S_t = 0, \alpha^{(t)} = \gamma_k^+)$ or $(S_t = 1, \alpha^{(t)} = \gamma_k^-)$. The symmetry of the system can be used to show these events occur with equal probability. The B_k state corresponds to the event that $(S_t = 0, \alpha^{(t)} = \gamma_k^-)$ or $(S_t = 1, \alpha^{(t)} = \gamma_k^+)$. Again, symmetry forces these events to occur with equal probability. The state probabilities, as a function of time, are defined by

$$\begin{aligned} A_k^{(t)} &= Pr(S_t = 0, \alpha^{(t)} = \gamma_k^+) + Pr(S_t = 1, \alpha^{(t)} = \gamma_k^-) \\ B_k^{(t)} &= Pr(S_t = 0, \alpha^{(t)} = \gamma_k^-) + Pr(S_t = 1, \alpha^{(t)} = \gamma_k^+). \end{aligned}$$

The basic idea of the new Markov chain is to simultaneously track the true state and count the number of erasures since the last instance of perfect knowledge. In doing this, we find that an observed 0 causes the transitions $A_k \rightarrow A_k$ and $B_k \rightarrow B_k$, an erased 0 causes the transitions $A_k \rightarrow A_{k+1}$ and $B_k \rightarrow B_{k+1}$, an erased $+$ or $-$ causes the transitions $A_k \rightarrow B_{k+1}$ and $B_k \rightarrow A_{k+1}$, and an observed $+$ or $-$ causes the transitions $A_k \rightarrow A_0$ and $B_k \rightarrow A_0$. Using the probabilities of these events gives the recursions

$$\begin{aligned} A_0^{(t+1)} &= A_0^{(t)}(1-\epsilon) + (1-A_0^{(t)})p(1-\epsilon) \\ A_k^{(t+1)} &= A_k^{(t)}(1-p)(1-\epsilon) + A_{k-1}^{(t)}(1-p)\epsilon + B_{k-1}^{(t)}p\epsilon \\ B_k^{(t+1)} &= B_k^{(t)}(1-p)(1-\epsilon) + B_{k-1}^{(t)}(1-p)\epsilon + A_{k-1}^{(t)}p\epsilon. \end{aligned}$$

Solving for the stationary distribution of the new Markov chain gives

$$\begin{aligned} Pr(A_0) &= \frac{p(1-\epsilon)}{p(1-\epsilon)+\epsilon} \\ Pr(A_k) &= Pr(A_0)\omega^k \gamma_k^+ \\ Pr(B_k) &= Pr(A_0)\omega^k \gamma_k^- \end{aligned}$$

where $\omega = \frac{\epsilon}{1-(1-p)(1-\epsilon)}$. Since the forward state probabilities must give a consistent state estimate, the events $(S_t = 0, \alpha^{(t)} = 0)$ and $(S_t = 1, \alpha^{(t)} = 1)$ must have probability zero. This also implies that $B_0^{(t)} = Pr(S_t = 0, \alpha^{(t)} = 0) + Pr(S_t = 1, \alpha^{(t)} = 1) = 0$. Finally, the the new Markov chain is uniformly ergodic by Theorem 4.4.7 and converges to its unique stationary distribution exponentially fast.

Now, we can compute the entropy rate using the limit $H(\mathcal{Y}) = \lim_{t\rightarrow\infty} H(Y_t|\mathbf{Y}_1^{t-1})$. When the new Markov chain is in state A_k or B_k, the conditional entropy is denoted by $H(Y|A_k)$ or $H(Y|B_k)$, respectively. These two expressions are given by

$$H(Y|A_k) = -\epsilon \log_2 \epsilon - (1-p)(1-\epsilon)\log_2\left((1-p)(1-\epsilon)\right) - p(1-\epsilon)\log_2\left(p(1-\epsilon)\gamma_k^+\right)$$

and

$$H(Y|B_k) = -\epsilon \log_2 \epsilon - (1-p)(1-\epsilon)\log_2\left((1-p)(1-\epsilon)\right) - p(1-\epsilon)\log_2\left(p(1-\epsilon)\gamma_k^-\right).$$

The first term of each expression is associated with the observation probability of an erasure, the second term with the observation probability of a 0, and the third term with the observation probability of either a $+$ or a $-$. Combining these with the stationary distribution of the new Markov chain gives final entropy rate

$$\begin{aligned} H(\mathcal{Y}) &= \sum_{k=0}^{\infty} Pr(A_k)H(Y|A_k) + Pr(B_k)H(Y|B_k) \\ &= D - p(1-\epsilon)\sum_{k=0}^{\infty} Pr(A_0)\omega^k \left(\gamma_k^+ \log_2 \gamma_k^+ + \gamma_k^- \log_2 \gamma_k^-\right), \end{aligned} \tag{4.7.2}$$

where

$$D = -\epsilon \log_2 \epsilon - (1-p)(1-\epsilon)\log_2\left((1-p)(1-\epsilon)\right) - p(1-\epsilon)\log_2\left(p(1-\epsilon)\right).$$

While we could find no closed form solution for this infinite sum, we did find a relatively simple approximation. Using (4.7.1), we can write

$$\log_2 \gamma_k^{\pm} = \log_2\left(1 \pm (1-2p)^k\right) - 1,$$

and then use the two term Taylor expansion, $\log_2(1+x) \approx (x - x^2/2)/\ln 2$, to get

$$\log_2 \gamma_k^{\pm} \approx \pm\frac{(1-2p)^k}{\ln 2} - \frac{(1-2p)^{2k}}{2\ln 2} - 1.$$

This lead us to the approximation

$$\gamma_k^+ \log_2 \gamma_k^+ + \gamma_k^- \log_2 \gamma_k^- \approx \frac{(1-2p)^{2k}}{2\log 2} - 1,$$

which allows the infinite sum (4.7.2) to be approximated in closed form by

$$\sum_{k=1}^{\infty} a_{0,\infty}\omega^k \left(\frac{(1-2p)^{2k}}{2\log 2} - 1\right) = a_{0,\infty}\left(\frac{1}{2\log 2}\frac{\omega(1-2p)^2}{1-\omega(1-2p)^2} - \frac{\omega}{1-\omega}\right).$$

The resulting entropy rate approximation, which we believe is actually an upper bound, is

$$H(\mathcal{Y}) \approx D - \frac{p^2(1-\epsilon)^2}{p(1-\epsilon)+\epsilon}\left(\frac{1}{2\log 2}\frac{\omega(1-2p)^2}{1-\omega(1-2p)^2} - \frac{\omega}{1-\omega}\right).$$

We evaluated the numerical error in this approximation over the rectangle formed by $\epsilon \in [0,1]$ and $p \in [1/2, 2/3]$, and found its maximum value to be roughly 0.0002.

The results of Sections 4.7.1 and 4.7.2 are shown in Figure 4.7.1, along with the capacity of the binary erasure channel (BEC) and the ternary erasure channel (TEC). While one might expect that the SIR and Markov-1 rate should be upper bounded by the capacity of the BEC, we see that this is definitely not the case. This is because the output alphabet of the BEC has only three symbols, while the output alphabet of the DEC has four symbols. Therefore, the rates of the DEC should be upper bounded by the capacity of the TEC. Surprisingly, the achievable rates of the DEC are quite close to the capacity of the TEC when ϵ is close to one.

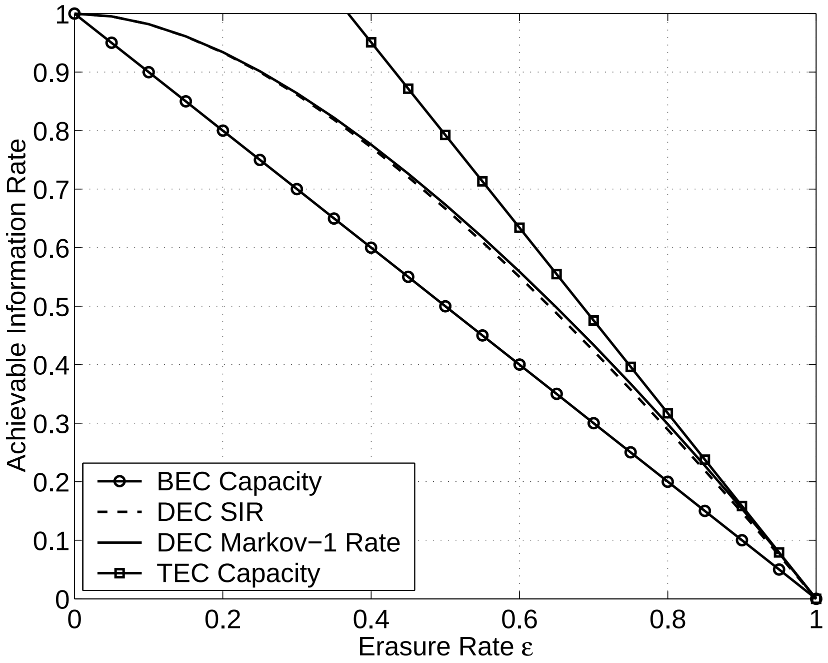

Figure 4.7.1: The SIR and Markov-1 rate of the DEC compared with the capacity of the binary erasure channel (BEC) and the ternary erasure channel (TEC).

4.7.3 Density Evolution for Finite State Channels

Density evolution is a pseudo-analytical method of analyzing LDPC codes that was introduced by Richardson and Urbanke in [25]. It analyzes a decoder by tracking the probabilistic evolution of messages passed around the decoder. In general, it is implemented by quantizing the continuous set of messages to a finite set and then tracking a probability distribution over that set.

Now, we consider a density evolution approach to the forward recursion of the BCJR algorithm. Since the state probability vector acts like a message in the BCJR algorithm, the first step is quantizing these vectors. For a two state channel, the vector is defined by a single parameter, and therefore we can use any scalar quantizer. We note that this idea was applied to two state fading channels by Goldsmith and Varaiya in [15]. For more complicated channels, the natural generalization amounts to using a vector quantizer rather than a scalar quantizer. We note that density evolution on the quantized vectors can be viewed either as an approximate analysis of the true algorithm or an exact analysis of the quantized algorithm.

Let the quantizer, $V(\mathbf{x})$, be a mapping from probability vectors of length N_Q to the index set, $\{1,\dots,N_V\}$. We abuse notation slightly and define the inverse of this mapping, $V^{-1}(i)$, to be some generalized centroid of the ith quantization cell $\{\mathbf{x} \in \mathfrak{D}(\mathcal{Q}) | V(\mathbf{x}) = i\}$. The forward variable of the quantized algorithm, A_t, is therefore characterized by the update equation

$$A_{t+1} = V\left(\frac{V^{-1}(A_t)\mathbf{M}(Y_t)}{\left\|V^{-1}(A_t)\mathbf{M}(Y_t)\right\|_1}\right), \tag{4.7.3}$$

which is simply a quantized version of (4.4.7).

Now, we consider the evolution of $Pr(A_t)$ while the underlying finite state Markov process transitions from state i to state j. In this case, the output, Y_t, is drawn from the distribution $g_{ij}(y)$. This gives rise to the transition matrix, $\mathbf{A}^{(i,j)}$, defined by

$$\left[\mathbf{A}^{(i,j)}\right]_{kl} = Pr(A_{t+1} = l | A_t = k, Q_t = i, Q_{t+1} = j).$$

This matrix can be constructed for channels with a finite output alphabet by evaluating (4.7.3) for all $A_t \in \{1,\dots,N_V\}$ and $Y_t \in \mathbb{Y}$ and assuming the corresponding probabilities. For channels with continuous output alphabets, one can either integrate over the appropriate regions of $\mathbb{Y}$ or approximate these probabilities by quantizing the output alphabet. The number of non-zero entries in each $\mathbf{A}^{(i,j)}$ matrix is also upper bounded by $N_V\,|\mathbb{Y}|$, and will therefore be sparse if $N_V \gg |\mathbb{Y}|$.

Next, we analyze the quantized algorithm completely by combining the $\mathbf{A}^{(i,j)}$ matrices with the state transition probabilities, p_{ij}. This allows us to define the $(N_Q N_V) \times (N_Q N_V)$ matrix,

$$\mathbf{A} = \begin{bmatrix} p_{1,1}\mathbf{A}^{(1,1)} & \cdots & p_{1,N_Q}\mathbf{A}^{(1,N_Q)} \\ \vdots & \ddots & \vdots \\ p_{N_Q,1}\mathbf{A}^{(N_Q,1)} & \cdots & p_{N_Q,N_Q}\mathbf{A}^{(N_Q,N_Q)} \end{bmatrix},$$

and point out that

$$[\mathbf{A}]_{(i-1)N_V+k,(j-1)N_V+l} = Pr(A_{t+1} = l, Q_{t+1} = j | A_t = k, Q_t = i).$$

While we expect that the stochastic matrix, $\mathbf{A}$, will generally have a unique stationary distribution, one can also consider the following pair of stationary distributions. Let the lower stationary distribution, $\underline{\mathbf{v}}$, be defined by $\lim_{n\to\infty} \mathbf{x}\frac{1}{n}\sum_{i=1}^{n}\mathbf{A}^i$ where $\mathbf{x}$ is given by quantizing the distribution (4.4.10). Likewise, let the upper stationary distribution, $\overline{\mathbf{v}}$, be defined by the same limit except that $\mathbf{x}$ is given by quantizing the distribution (4.4.11). These limits are always well defined and can be computed using the eigenvalue decomposition of $\mathbf{A}$. We note that the sparsity of $\mathbf{A}$ may also be exploited to reduce complexity.

Consider the probability vectors, $\mathbf{v}^{(t)}$, defined by $\mathbf{v}^{(t+1)} = \mathbf{v}^{(t)}\mathbf{A}$. Based on the definition of $\mathbf{A}$, these vectors have the implicit definition,

$$\left[\mathbf{v}^{(t)}\right]_{(i-1)N_V+k} = Pr(A_t = k, Q_t = i).$$

Using this, we find that the entropy estimate at time t is given by

$$H(Y_t|\mathbf{Y}_1^{t-1}, W) \approx \sum_{i,j,k}\left[\mathbf{v}^{(t)}\right]_{(i-1)N_V+k} p_{ij} E\left[\log\left\|V^{-1}(k)\mathbf{M}(Y_t)\right\|_1 |Q_t = i, Q_{t+1} = j\right],$$

where W is any random variable which gives rise to the initial distribution, $\mathbf{v}^{(1)}$. This same formula can be used with the stationary distributions $\underline{\mathbf{v}}$ and $\overline{\mathbf{v}}$ to estimate the upper and lower entropy rate bounds.

Since we have a valid probabilistic analysis of the quantized algorithm, we can actually show that any entropy computed in this manner is an upper bound on the same entropy computed via an exact algorithm. For example, suppose we compute the entropy $H(Y_t|\mathbf{Y}_1^{t-1}, W)$ where W is initialized by the vector $\mathbf{v}^{(1)}$. In this case, the entropy computed by the quantized algorithm will always be larger because its state probability estimates are less accurate and therefore increase the entropy.

While the approximation error of this algorithm is quite dependent on the particular quantizer used, we can still make a few general statements. We note that all of these statements are based on the fact that the entropy expression is continuous function on the state probability vector. This means that one would expect the entropy approximation error from using a uniform quantizer to decay like $O\left(N_V^{-1/(N_Q-1)}\right)$. When using an optimized vector quantizer, one would expect the error to decay like $O\left(N_V^{-1/d}\right)$, where d is the (possibly fractal) dimension of the true stationary distribution of the joint Markov chain. This means that this type of analysis may actually be less efficient than Monte Carlo methods when $d > 2$.

This method has been applied successfully to the dicode channel with AWGN. In particular, we used a non-linear scalar quantizer based on the uniform quantization of log-likelihood ratios. This type of quantization is widely used in the density evolution analysis of LDPC codes [25]. For the dicode channel, the Monte Carlo method can easily achieve tolerances of 10^{-3} while the density evolution approach can achieve tolerances around 10^{-6} with some effort. We note that the density evolution results were in complete agreement with the Monte Carlo results from Section 4.6. The practical value of achieving tolerances less than 10^{-3} is questionable, however.

Remark 4.7.1. While this section discusses only the forward recursion of the BCJR algorithm, the same type of analysis may be applied to the backwards recursion and the output stage. This gives a valid probabilistic analysis of a quantized BCJR algorithm that can be used to approximate the log-likelihood density at the output of the BCJR algorithm. In particular, these densities can be used to optimize LDPC codes and compute information rates for the multilevel coding approach proposed in [24].

4.8 Concluding Remarks

This chapter discusses a number of issues related to entropy rates and capacity for finite state channels. All of the results which are not expressly attributed to other authors were developed independently by us. That said, this field is currently the subject of great interest, and many of the same ideas have recently been developed independently by other authors. For example, the simple Monte Carlo method was published in 2001 by three separate groups [1][24][27]. The formulation of the entropy rate as a Lyapunov exponent was also discovered independently and reported in [17]. Finally, the quantized density evolution approach for information rates is quite natural for two state channels was introduced in [15]. The move to vector quantization is a natural generalization and is also used in a slightly different manner in [32] to help estimate the feedback capacity of finite state channels.

4A Formal Channel Definitions

4A.1 Discrete Input Linear Filter Channels with AWGN

The formal definition, $(\mathbb{X}, \mathbb{Y}, \mathbf{F}(\cdot, \cdot))$, of this finite state channel depends solely on ν, $(h_0, h_1, \ldots, h_\nu)$, σ^2, and $\mathbb{X}$. We start by noting that the number of channel states is given by

$N_S = |\mathbb{X}|^\nu$ and defining the output alphabet, in terms of the input alphabet and channel taps, with

$$\mathbb{Y} = \left\{ y \in \mathbb{R} | y = \sum_{i=0}^{\nu} h_i x_i,\ (x_0, \ldots, x_\nu) \in \mathbb{X}^\nu \right\}.$$

Next, we define $N_X = |\mathbb{X}|$ and let ξ be any one to one mapping from $\mathbb{X}$ to the set $\{0, 1, \ldots, N_X - 1\}$. Although, the channel state is clearly defined by the last $\nu - 1$ inputs, we would also like an integer representation of this quantity. Using a base conversion from $\nu - 1$ digits of $\mathbb{X}$ to the integers, we have the integer state, $S_t = \sum_{i=1}^{\nu} (N_X)^{\nu-i} \xi(X_{t-i})$, and the one step update, $S_{t+1} = \lfloor S_t / N_X \rfloor + \xi(X_t)(N_X)^\nu$. Finally, we define $[\mathbf{F}(x,y)]_{ij} = f_{ij}(x,y)$ with

$$f_{ij}(x,y) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-m_{ij})^2}{2\sigma^2}} & \text{if } j = \lfloor i/N_X \rfloor + \xi(x)(N_X)^\nu \\ 0 & \text{otherwise} \end{cases},$$

where $m_{ij} = h_\nu \xi^{-1}(i \mod N_X) + \sum_{l=0}^{\nu-1} h_l \xi^{-1} \left(\lfloor j/(N_X)^{\nu-l-1} \rfloor \mod N_X \right)$.

4A.2 Dicode Erasure Channel

The formal definition, $(\mathbb{X}, \mathbb{Y}, \mathbf{F}(\cdot,\cdot))$, of this finite state channel depends only on ϵ. The input and output alphabets are defined by $\mathbb{X} = \{0, 1\}$ and $\mathbb{Y} = \{+, 0, -, e\}$, and $N_S = 2$. The conditional transition-observation probabilities are given by $[\mathbf{F}(x,y)]_{ij} = f_{ij}(x,y)$ where $f_{ij}(x,y) = 0$ unless defined by $f_{00}(0,0) = f_{11}(1,0) = f_{01}(1,+) = f_{10}(0,-) = 1 - \epsilon$ or $f_{00}(0,e) = f_{11}(1,e) = f_{01}(1,e) = f_{10}(0,e) = \epsilon$.

4A.3 Finite State Z-Channel

The formal definition, $(\mathbb{X}, \mathbb{Y}, \mathbf{F}(\cdot,\cdot))$, of this finite state channel has $\mathbb{X} = \mathbb{Y} = \{0, 1\}$ and $N_S = 2$. The conditional transition-observation probabilities are given by $[\mathbf{F}(x,y)]_{ij} = f_{ij}(x,y)$ where $f_{ij}(x,y) = 0$ unless defined by $f_{0,0}(0,1) = f_{1,1}(1,0) = p$, $f_{0,0}(0,0) = f_{1,1}(1,1) = 1 - p$, or $f_{0,1}(1,1) = f_{1,0}(0,0) = 1$.

4A.4 The Finite State Z-Channel with Markov-1 Inputs

Consider a finite state Z-channel with a stochastic input sequence. Using the notation above, we define the input process, $(\mathbf{\Theta}, \mathbf{\Phi})$, with

$$\mathbf{\Theta} = \begin{bmatrix} 1-q & q \\ q & 1-q \end{bmatrix}, \ \mathbf{\Phi} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

This allows us to define the stochastic output sequence with the triple, $(\mathbb{Y}, \mathbf{P}, \mathbf{G}(\cdot))$, where $\mathbb{Y} = \{0, 1\}$,

$$\mathbf{P} = \begin{bmatrix} 1-q & 0 & 0 & q \\ 1-q & 0 & 0 & q \\ q & 0 & 0 & 1-q \\ q & 0 & 0 & 1-q \end{bmatrix}, \ \mathbf{G}(0) = \begin{bmatrix} 1-p & 0 & 0 & 0 \\ 1 & 0 & 0 & p \\ 1-p & 0 & 0 & 0 \\ 1 & 0 & 0 & p \end{bmatrix}, \ \mathbf{G}(1) = \begin{bmatrix} p & 0 & 0 & 1 \\ 0 & 0 & 0 & 1-p \\ p & 0 & 0 & 1 \\ 0 & 0 & 0 & 1-p \end{bmatrix}.$$

In this case, the transition probability matrix $\mathbf{P}$, and therefore underlying Markov chain, is reducible. Therefore, we simplify our description of the process $\{Y_t\}_{t\geq 1}$by removing any state whose stationary probability is zero. Removing states 1 and 2 results in the simplified description

$$\mathbf{P} = \begin{bmatrix} 1-q & q \\ q & 1-q \end{bmatrix}, \ \mathbf{G}(0) = \begin{bmatrix} 1-p & 0 \\ 1 & p \end{bmatrix}, \ \mathbf{G}(1) = \begin{bmatrix} p & 1 \\ 0 & 1-p \end{bmatrix}.$$

Bibliography

[1] D. Arnold and H. Loeliger. On the information rate of binary-input channels with memory. In *Proc. IEEE Int. Conf. Commun.*, pages 2692–2695, Helsinki, Finland, June 2001.

[2] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Trans. Inform. Theory*, 20(2):284–287, March 1974.

[3] A. R. Barron. The strong ergodic theorem for densities: Generalized Shannon-McMillan-Breiman theorem. *Ann. Probab.*, 13(4):1292–1303, Nov. 1985.

[4] M. Benda. A central limit theorem for contractive stochastic dynamical systems. *J. Appl. Prob.*, 35:200–205, 1998.

[5] J. J. Birch. On information rates of finite-state channels. *Inform. and Control*, 6:372–380, 1963.

[6] D. Blackwell. Entropy of functions of finite-state Markov chains. *Trans. First Prague Conf. on Inform. Theory, Stat. Dec. Fun., Rand. Processes*, pages 13–20, 1957.

[7] D. Blackwell, L. Breiman, and A. J. Thomasian. Proof of shannon's transmission theorem for finite-state indecomposable channels. *Ann. Math. Stats.*, 29:1209–1220, Dec. 1958.

[8] X. Chen. Limit theorems for functionals of ergodic Markov chains with general state space. *Memoirs of the AMS*, 139(664), May 1999.

[9] H. Cohn, O. Nerman, and M. Peligrad. Weak ergodicity and products of random matrices. *J. Theor. Prob.*, 6:389–405, July 1993.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[11] A. Dasdan, S. S. Irani, and R. K. Gupta. Efficient algorithms for optimum cycle mean and optimum cost to time ratio problems. In *Proc. 36th Design Automation Conf.*, pages 37–42, June 1999.

[12] P. Diaconis and D. Freedman. Iterated random functions. *SIAM Review*, 41(1):45–76, Jan. 1999.

[13] H. Furstenberg and H. Kesten. Products of random matrices. *Ann. Math. Stats.*, 31:457–469, June 1960.

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

[15] A. J. Goldsmith and P. P. Varaiya. Capacity, mutual information, and coding for finite-state Markov channels. *IEEE Trans. Inform. Theory*, 42(3):868–886, May 1996.

[16] W. Hirt. *Capacity and Information Rates of Discrete-Time Channels with Memory*. PhD thesis, E.T.H., Zurich, Switzerland, 1988.

[17] T. Holliday, A. Goldsmith, and P. Glynn. Entropy and mutual information for markov channels with general inputs. In *Proc. 40th Annual Allerton Conf. on Commun., Control, and Comp.*, Monticello, IL, USA, Oct. 2002.

[18] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, New Jersey, USA, 1985.

[19] A. Kavčić. On the capacity of Markov sources over noisy channels. In *Proc. IEEE Global Telecom. Conf.*, pages 2997–3001, San Antonio, Texas, USA, Nov. 2001.

[20] F. Le Gland and L. Mevel. Basic properties of the projective product with application to products of column-allowable nonnegative matrices. *Math. Control Signals Systems*, 13(1):41–62, July 2000.

[21] F. Le Gland and L. Mevel. Exponential forgetting and geometric ergodicity in hidden Markov models. *Math. Control Signals Systems*, 13(1):63–93, July 2000.

[22] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

[23] V. I. Oseledec. A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems. *Trans. Moscow Math. Soc.*, pages 197–231, 1968.

[24] H. D. Pfister, J. B. Soriaga, and P. H. Siegel. On the achievable information rates of finite state ISI channels. In *Proc. IEEE Global Telecom. Conf.*, pages 2992–2996, San Antonio, Texas, USA, Nov. 2001.

[25] T. J. Richardson and R. L. Urbanke. The capacity of low-density parity check codes under message-passing decoding. *IEEE Trans. Inform. Theory*, 47(2):599–618, Feb. 2001.

[26] S. Shamai, L. H. Ozarow, and A. D. Wyner. Information rates for a discrete-time Gaussian channel with intersymbol interference and stationary inputs. *IEEE Trans. Inform. Theory*, 37(6):1527–1539, Nov. 1991.

[27] V. Sharma and S. K. Singh. Entropy and channel capacity in the regenerative setup with applications to Markov channels. In *Proc. IEEE Int. Symp. Information Theory*, page 283, Washington, DC, USA, June 2001.

[28] J. B. Soriaga, H. D. Pfister, and P. H. Siegel. On the low rate Shannon limit for binary intersymbol interference channels. submitted to *IEEE Trans. Commun.*, Oct. 2003.

[29] Ö. Stenflo. Ergodic theorems for Markov chains represented by iterated function systems. *Bull. Polish Acad. Sci. Math.*, 49(1):27–43, 2001.

[30] R. Urbanke. Iterative coding systems. http://www.calit2.net/events/2001/courses/ics.pdf, Aug. 2001.

[31] P. O. Vontobel and D. M. Arnold. An upper bound on the capacity of channels with memory and constraint input. In *Proc. IEEE Inform. Theory Workshop*, pages 147–149, Cairns, Australia, Sept. 2001.

[32] S. Yang and A. Kavčić. Markov sources achieve the feedback capacity of finite-state machine channels. In *Proc. IEEE Int. Symp. Information Theory*, page 361, Lausanne, Switzerland, June 2002.