Capacity, Mutual Information, and Coding for Finite-State Markov Channels

Andrea J. Goldsmith, *Member, IEEE* and Pravin P. Varaiya, *Fellow, IEEE*

Abstract— **The Finite-State Markov Channel (FSMC) is a discrete time-varying channel whose variation is determined by a finite-state Markov process. These channels have memory due to the Markov channel variation. We obtain the FSMC capacity as a function of the conditional channel state probability. We also show that for i.i.d. channel inputs, this conditional probability converges weakly, and the channel's mutual information is then a closed-form continuous function of the input distribution. We next consider coding for FSMC's. In general, the complexity of maximum-likelihood decoding grows exponentially with the channel memory length. Therefore, in practice, interleaving and memoryless channel codes are used. This technique results in some performance loss relative to the inherent capacity of channels with memory. We propose a maximum-likelihood decision-feedback decoder with complexity that is independent of the channel memory. We calculate the capacity and cutoff rate of our technique, and show that it preserves the capacity of certain FSMC's. We also compare the performance of the decision-feedback decoder with that of interleaving and memoryless channel coding on a fading channel with 4PSK modulation.**

Index Terms—**Finite-state Markov channels, capacity, mutual information, decision-feedback maximum-likelihood decoding.**

I. Introduction

THIS PAPER extends the capacity and coding results of Mushkin and Bar-David [1] for the Gilbert–Elliot channel to a more general time-varying channel model. The Gilbert–Elliot channel is a stationary two-state Markov chain, where each state is a binary-symmetric channel (BSC), as in Fig. 1. The transition probabilities between states are g and b, respectively, and the crossover probabilities for the "good" and "bad" BSC's are p_G and p_B, respectively, where $p_G < p_B$. Let $x_n \in \{0,1\}$, $y_n \in \{0,1\}$, and $z_n = x_n \oplus y_n$ denote, respectively, the channel input, channel output, and channel error on the nth transmission. In [1], the capacity of the Gilbert–Elliot channel is derived as

$$C = \lim_{n\to\infty} 1 - E[h(q_n)] = 1 - E[h(q_\infty)] \tag{1}$$

where h is the entropy function, $q_n = p(z_n = 1 \mid z^{n-1})$, q_n converges to q_∞ in distribution, and q_∞ is independent of the initial channel state.

In this paper we derive the capacity of a more general finite-state Markov channel, where the channel states are not necessarily BSC's. We model the channel as a Markov chain S_n which takes values in a finite state space $\mathcal{C}$ of memoryless channels with finite input and output alphabets. The conditional input/output probability is thus $p(y_n \mid x_n, S_n)$, where x_n and y_n denote the channel input and output, respectively. The channel transition probabilities are independent of the input, so our model does not include ISI channels. We refer to the channel model as a finite-state Markov channel (FSMC). If the transmitter and receiver have perfect state information, then the capacity of the FSMC is just the statistical average over all states of the corresponding channel capacity [2]. On the other hand, with no information about the channel state or its transition structure, capacity is reduced to that of the Arbitrarily Varying Channel [3]. We consider the intermediate case, where the channel transition structure of the FSMC is known.

The memory of the FSMC comes from the dependence of the current channel state on past inputs and outputs. As a result, the entropy in the channel output is a function of the channel state conditioned on all past outputs. Similarly, the conditional output entropy given the input is determined by the channel state probability conditioned on all past inputs and outputs. We use this fact to obtain a formula for channel capacity in terms of these conditional probabilities. Our formula can be computed recursively, which significantly reduces its computation complexity. We also show that when the channel inputs are i.i.d., these conditional state probabilities converge in distribution, and their limit distributions are continuous functions of the input distribution. Thus for any i.i.d. input distribution θ, the mutual information of the FSMC is a closed-form continuous function of θ. This continuity allows us to find $I_{\text{i.i.d}}$, the maximum mutual information relative to all i.i.d. input distributions, using straightforward maximization techniques. Since $I_{\text{i.i.d}} < C$, our result provides a simple lower bound for the capacity of general FSMC's.

The Gilbert–Elliot channel has two features which facilitate its capacity analysis: its conditional entropy $H(Y^n \mid X^n)$ is independent of the input distribution, and it is a symmetric channel, so a uniform input distribution induces a uniform output distribution. We extend these properties to a general class of FSMC's and show that for this class, $I_{\text{i.i.d}}$ equals the channel capacity. This class includes channels varying between

Manuscript received February 18, 1994; revised September 15, 1995. This work was supported in part by an IBM graduate fellowship, and in part by the PATH program, Institute of Transportation Studies, University of California, Berkeley. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Trondheim, Norway, June 1994.

A. J. Goldsmith is with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125 USA.

P. P. Varaiya is with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720 USA.

Publisher Item Identifier S 0018-9448(96)02935-5.

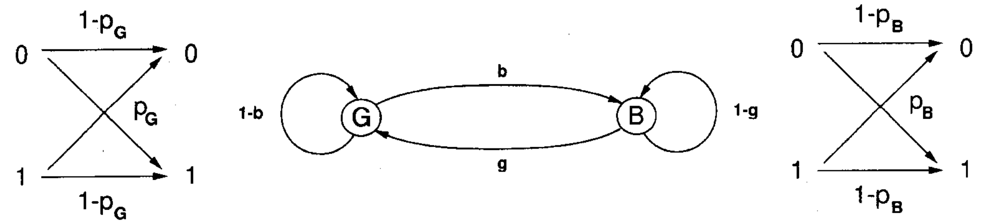

Fig. 1. Gilbert–Elliot channel.

any finite number of BSC's, as well as quantized additive white noise (AWN) channels with symmetric PSK inputs and time-varying noise statistics or amplitude fading.

In principle, communication over a finite-state channel is possible at any rate below the channel capacity. However, good maximum-likelihood (ML) coding strategies for channels with memory are difficult to determine, and the decoder complexity grows exponentially with memory length. Thus a common strategy for channels with memory is to disperse the memory using an interleaver: if the span of the interleaver is long, then the cascade of the interleaver, channel, and deinterleaver can be considered memoryless, and coding techniques for memoryless channels may be used [4]. However, this cascaded channel has a lower inherent Shannon capacity than the original channel, since coding is restricted to memoryless channel codes.

The complexity of ML decoding can be reduced significantly without this capacity degradation by implementing a *decision-feedback decoder*, which consists of a recursive estimator for the channel state distribution conditioned on past inputs and outputs, followed by an ML decoder. We will see that the estimate $\pi_n = p(S_n \mid x_{n-1}, \cdots, x_1, y_{n-1}, \cdots, y_1)$ is a sufficient statistic for the ML decoder input, given all past inputs and outputs. Thus the ML decoder operates on a memoryless system. The only additional complexity of this approach over the conventional method of interleaving and memoryless channel encoding is the recursive calculation of π_n. We will calculate the capacity penalty of the decision-feedback decoder for general FSMC's (ignoring error propagation), and show that this penalty vanishes for a certain class of FSMC's.

The most common example of an FSMC is a correlated fading channel. In [5], an FSMC model for Rayleigh fading is proposed, where the channel state varies over binary-symmetric channels with different crossover probabilities. Our recursive capacity formula is a generalization of the capacity found in [5], and we also prove the convergence of their recursive algorithm. Since capacity is generally unachievable for any practical coding scheme, the channel cutoff rate indicates the practical achievable information rate of a channel with coding. The cutoff rate for correlated fading channels with MPSK inputs, assuming channel state information at the receiver, was obtained in [6]: we obtain the same cutoff rate on this channel using decision-feedback decoding.

Most coding techniques for fading channels rely on built-in time diversity in the code to mitigate the fading effect. Code designs of this type can be found in [7]–[9] and the references therein. These codes use the same time-diversity idea as interleaving and memoryless channel encoding, except that the diversity is implemented with the code metric instead of the interleaver. Thus as with interleaving and memoryless channel encoding, channel correlation information is ignored with these coding schemes. Maximum-likelihood sequence estimation for fading channels without coding has been examined in [10], [11]. However, it is difficult to implement coding with these schemes due to the code delays. In our scheme, coding delays do not result in state decision delays, since the decisions are based on estimates of the coded bits. We can introduce coding in our decision-feedback scheme with a consequent increase in delay and complexity, as we will discuss in Section VI.

The remainder of the paper is organized as follows. In Section II we define the FSMC, and obtain some properties of the channel based on this definition. In Section III we derive a recursive relationship for the distribution of the channel state conditioned on past inputs and outputs, or on past outputs alone. We also show these conditional state distributions converge to limit distributions for i.i.d. channel inputs. In Section IV we obtain the capacity of the FSMC in terms of the condition state distributions, and obtain a simple formula for $I_{\text{i.i.d.}}$. Uniformly symmetric variable-noise FSMC's are defined in Section V. For this channel class (which includes the Gilbert–Elliot channel), capacity is achieved with uniform i.i.d. channel inputs. In Section VI we present the decision-feedback decoder, and obtain the capacity and cutoff rate penalties of the decision-feedback decoding scheme. These penalties vanish for uniformly symmetric variable-noise channels. Numerical results for the capacity and cutoff rate of a two-state variable-noise channel with 4PSK modulation and decision-feedback decoding are presented in Section VII.

II. Channel Model

Let S_n be the state at time n of an irreducible, aperiodic, stationary Markov chain with state space $\mathcal{C} = \{c_1, \cdots, c_K\}$. S_n is positive recurrent and ergodic. The state space $\mathcal{C}$ corresponds to K different discrete memoryless channels (DMC's), with common finite input and output alphabets denoted by $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let P be the matrix of transition probabilities for S, so

$$P_{km} = p(S_{n+1} = c_m \mid S_n = c_k) \tag{2}$$

independent of n by stationarity. We denote the input and output of the FSMC at time n by x_n and y_n, respectively, and we assume that the channel inputs are independent of its states. We will use the notation

$$r^n \triangleq (r_1, \dots, r_n)$$

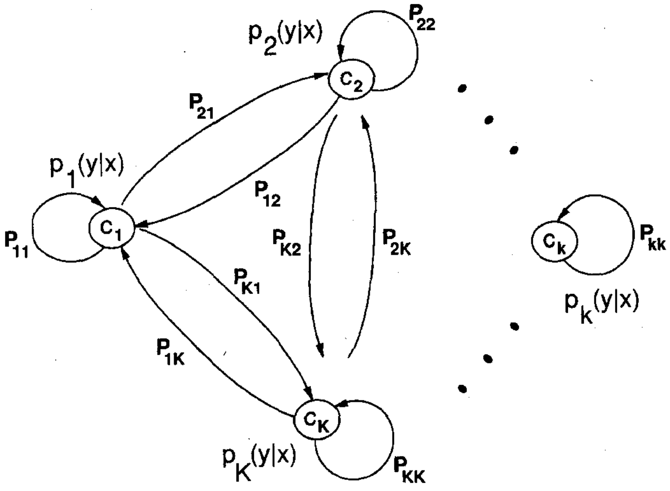

Fig. 2. Finite-state Markov channel.

and

$$r_m^{n+m} \triangleq (r_m, \cdots, r_{n+m})$$

for $r = x, y$, or S.

The FSMC is defined by its conditional input/output probability at time n, which is determined by the channel state at time n

$$p(y_n \mid x_n, S_n) = \sum_{k \in K} p_k(y_n \mid x_n) I[S_n = c_k] \quad . \qquad (3)$$

where $p_k(y \mid x) = p(y \mid x, S = c_k)$, and $I[\cdot]$ denotes the indicator function ($I[S_n = c_k] = 1$ if $S_n = c_k$ and 0 otherwise). The memory of the FSMC is due to the Markov structure of the state transitions, which leads to a dependence of S_n on previous values. The FSMC is memoryless if and only if $P_{km} = P_{jm}$ for all k, j, and m. The finite-state Markov channel is illustrated in Fig. 2.

By assumption, the state at time $n+1$ is independent of previous input/output pairs when conditioned on S_n

$$p(S_{n+1} \mid S_n, x^n, y^n) = p(S_{n+1} \mid S_n). \qquad (4)$$

Since the channels in $\mathcal{C}$ are memoryless

$$p(y_{n+1} \mid S_{n+1}, x_{n+1}, S^n, x^n, y^n) = p(y_{n+1} \mid S_{n+1}, x_{n+1}). \qquad (5)$$

If we also assume that the x_n's are independent, then

$$p(y_{n+1}, x_{n+1} \mid S_{n+1}, S^n, x^n, y^n) = p(y_{n+1}, x_{n+1} \mid S_{n+1}). \qquad (6)$$

From (6)

$$p(y^N, x^N \mid S^N) = \prod_{n=1}^{N} p(y_n, x_n \mid S_n) \qquad (7)$$

and

$$p(y_{n+1} \mid S_{n+1}, S^n, y^n) = p(y_{n+1} \mid S_{n+1}). \qquad (8)$$

III. Conditional State Distribution

The conditional channel state distribution is the key to determining the capacity of the FSMC through a recursive algorithm. It is also a sufficient statistic for the input given all past inputs and outputs, thus allowing for the reduced complexity of the decision-feedback decoder. In this section we show that the state distribution conditioned on past input/output pairs can be calculated using a recursive formula. A similar formula is derived for the state distribution conditioned on past outputs alone, under the assumption of independent channel inputs. We also show that these state distributions converge weakly under i.i.d. inputs, and the resulting limit distributions are continuous functions of the input distribution.

We denote these conditional state distributions by the K-dimensional random vectors $\pi_n = (\pi_n(1), \cdots, \pi_n(K))$ and $\rho_n = (\rho_n(1), \cdots, \rho_n(K))$, respectively, where

$$\rho_n(k) = p(S_n = c_k \mid y^{n-1}) \qquad (9)$$

and

$$\pi_n(k) = p(S_n = c_k \mid x^{n-1}, y^{n-1}). \qquad (10)$$

The following recursive formula for π_n is derived in Appendix I:

$$\pi_{n+1} = \frac{\pi_n D(x_n, y_n) P}{\pi_n D(x_n, y_n) \underline{1}} \triangleq f(x_n, y_n, \pi_n) \qquad (11)$$

where $D(x_n, y_n)$ is a diagonal $K \times K$ matrix with kth diagonal term $p_k(y_n \mid x_n)$, and $\underline{1} = (1, \cdots, 1)^T$ is a K-dimensional vector. Equation (11) defines a recursive relation for π_n which takes values on the state space

$$\Delta = \left\{ \alpha \in R^K \mid \alpha_i \geq 0, \sum \alpha_i = 1 \right\}.$$

The initial value for π_n is

$$\pi_0 = (p(S_0 = c_1), \cdots, p(S_0 = c_K))$$

and its transition probabilities are

$$p(\pi_{n+1} = \alpha \mid \pi_n = \beta) = \sum_{\substack{x_n \in \mathcal{X} \\ y_n \in \mathcal{Y}}} 1[(x_n, y_n)\colon f(x_n, y_n, \beta) = \alpha] \cdot p(y_n \mid \pi_n = \beta, x_n) p(x_n). \qquad (12)$$

Note that (12) is independent of n for stationary inputs.

For independent inputs, there is a similar recursive formula for ρ_n

$$\rho_{n+1} = \frac{\rho_n B(y_n) P}{\rho_n B(y_n) \underline{1}} \triangleq \hat{f}(y_n, \rho_n) \qquad (13)$$

where $B(y_n)$ is a diagonal $K \times K$ matrix with kth diagonal term $p(y_n \mid S_n = c_k)$.[1] The derivation of (13) is similar to that of (11) in Appendix I, using (8) instead of (5) and removing all x terms. The variable ρ_n also takes values on the state space Δ, with initial value $\rho_0 = \pi_0$ and transition probabilities

$$p(\rho_{n+1} = \alpha \mid \rho_n = \beta) = \sum_{y_n \in \mathcal{Y}} 1[y_n\colon \hat{f}(y_n, \beta) = \alpha] \cdot p(y_n \mid \rho_n = \beta). \qquad (14)$$

[1] Note that $B(y_n)$ has an implicit dependence on the distribution of x_n.

As for π_n, the transition probabilities in (14) are independent of n when the inputs are stationary.

We show in Appendix II that for i.i.d. inputs, π_n and ρ_n are Markov chains that converge in distribution to limits which are independent of the initial channel state, under some mild constraints on $\mathcal{C}$. These convergence results imply that for any bounded continuous function f, the following limits exist and are equal for all i:

$$\lim_{n\to\infty} E[f(\pi_n)] = \lim_{n\to\infty} E\left[f(\pi_n^i)\right] \tag{15}$$

and

$$\lim_{n\to\infty} E[f(\rho_n)] = \lim_{n\to\infty} E\left[f(\rho_n^i)\right] \tag{16}$$

where

$$\pi_n^i = p(S_n \mid x^{n-1}, y^{n-1}, S_0 = c_i)$$

and

$$\rho_n^i = p(S_n \mid y^{n-1}, S_0 = c_i).$$

This convergence allows us to obtain a closed-form solution for the mutual information under i.i.d. inputs. We also show in Lemmas A2.3 and A2.5 of Appendix II that the limit distributions for π and ρ are continuous functions of the input distribution.

Lemmas A2.6 and A2.7 of Appendix II show the surprising result that π_n and ρ_n are not necessarily Markov chains when the input distribution is Markov. Since the weak convergence of π_n and ρ_n requires this Markov property, (15) and (16) are not valid for general Markov inputs.

IV. Entropy, Mutual Information, and Capacity

We now derive the capacity of the FSMC based on the distributions of π_n and ρ_n. We also obtain some additional properties of the entropy and mutual information when the channel inputs are i.i.d.

By definition, the Markov chain S_n is aperiodic and irreducible over a finite state space, so the effect of its initial state dies away exponentially with time [12]. Thus the FSMC is an indecomposable channel. The capacity of an indecomposable channel is independent of its initial state, and is given by [13, Theorem 4.6.4]

$$C = \lim_{n\to\infty} \max_{\mathcal{P}(X^n)} \frac{1}{n} I(X^n; Y^n) \tag{17}$$

where $I(\cdot;\cdot)$ denotes mutual information and $\mathcal{P}(X^n)$ denotes the set of all input distributions on X^n. The mutual information can be written as

$$I(X^n; Y^n) = H(Y^n) - H(Y^n \mid X^n) \tag{18}$$

where $H(Y) = \boldsymbol{E}[-\log p(y)]$ and $H(Y \mid X) = \boldsymbol{E}[-\log p(y \mid x)]$. It is easily shown [14] that

$$H(Y^n) = \sum_{i=1}^{n} H(Y_i \mid Y^{i-1}) \tag{19}$$

and

$$H(Y^n \mid X^n) = \sum_{i=1}^{n} H(Y_i \mid X_i, Y^{i-1}, X^{i-1}). \tag{20}$$

The following lemma, proved in Appendix III, allows the mutual information to be written in terms of π_n and ρ_n.

Lemma 4.1:

$$\begin{aligned} H(Y_n \mid X_n, X^{n-1}, Y^{n-1}) &= \boldsymbol{E}\left[-\log \sum_{k=1}^{K} p(y_n \mid x_n, S_n = c_k)\pi_n(k)\right] \\ &= H(Y_n \mid X_n, \pi_n) \end{aligned} \tag{21}$$

and

$$\begin{aligned} H(Y_n \mid Y^{n-1}) &= \boldsymbol{E}\left[-\log \sum_{k=1}^{K} p(y_n \mid S_n = c_k)\rho_n(k)\right] \\ &= H(Y_n \mid \rho_n). \end{aligned} \tag{22}$$

Using this lemma in (19) and (20) and substituting into (18) yields the following theorem.

Theorem 4.1: The capacity of the FSMC is given by

$$\begin{aligned} C = \lim_{n\to\infty} \max_{\mathcal{P}(X^n)} \frac{1}{n} &\cdot \sum_{i=1}^{n} \left[\boldsymbol{E}\left[-\log \sum_{k=1}^{K} p(y_i \mid S_i = c_k)\rho_i(k)\right] \right. \\ &\left. -\boldsymbol{E}\left[-\log \sum_{k=1}^{K} p(y_i \mid x_i, S_i = c_k)\pi_i(k)\right]\right] \end{aligned} \tag{23}$$

where the dependence on $\theta \in \mathcal{P}(X^n)$ of the distributions for π_i, ρ_i, and y_i is implicit. This capacity expression is easier to calculate than Gallager's formula (17), since the π_i terms can be computed recursively. The recursive calculation for ρ_i requires independent inputs. However, for many channels of interest $H(Y_i \mid \rho_i)$ will be a constant independent of the input distribution (such channels are discussed in Section V). For these channels, the capacity calculation reduces to minimizing the second term in (23) relative to the input distribution, and the complexity of this minimization is greatly reduced when π_i can be calculated easily.

Using Lemma 4.1, we can also express the capacity as

$$C = \lim_{n\to\infty} \max_{\mathcal{P}(X^n)} \frac{1}{n} \sum_{i=1}^{n} [H(Y_i \mid \rho_i) - H(Y_i \mid X_i, \pi_i)]. \tag{24}$$

Although [13, Theorem 4.6.4] guarantees the convergence of (24), the random vectors π_n and ρ_n do not necessarily converge in distribution for general input distributions. We proved this convergence in Section III for i.i.d. inputs. We now derive some additional properties of the entropy and mutual information under this input restriction. These properties are summarized in Lemmas 4.2–4.7 below, which are proved in Appendix IV.

Lemma 4.2: When the channel inputs are stationary

$$\begin{aligned} H(Y_n \mid X_n, X^{n-1}, Y^{n-1}) &\geq H(Y_{n+1} \mid X_{n+1}, X^n, Y^n) \\ &\geq H(Y_{n+1} \mid X_{n+1}, X^n, Y^n, S_0) \\ &\geq H(Y_n \mid X_n, X^{n-1}, Y^{n-1}, S_0). \end{aligned} \tag{25}$$

Lemma 4.3: For i.i.d. input distributions, the following limits exist and are equal:

$$\begin{aligned} &\lim_{n\to\infty} H(Y_n \mid X_n, X^{n-1}, Y^{n-1}) \\ &\quad = \lim_{n\to\infty} H(Y_n \mid X_n, X^{n-1}, Y^{n-1}, S_0). \end{aligned} \tag{26}$$

We now consider the entropy in the output alone.

Lemma 4.4 For stationary inputs,

$$\begin{aligned} H(Y_n \mid Y^{n-1}) \geq H(Y_{n+1} \mid Y^n) &\geq H(Y_{n+1} \mid Y^n, S_0) \\ &\geq H(Y_n \mid Y^{n-1}, S_0). \end{aligned} \tag{27}$$

Lemma 4.5: For i.i.d. input distributions, the following limits exist and are equal:

$$\lim_{n\to\infty} H(Y_n \mid Y^{n-1}) = \lim_{n\to\infty} H(Y_n \mid Y^{n-1}, S_0). \tag{28}$$

The next lemma is proved using the convergence results for π_n and ρ_n and a change of variables in the entropy expressions (26) and (28).

Lemma 4.6: For any i.i.d. input distribution $\theta \in \mathcal{P}(X)$

$$\begin{aligned} &\lim_{n\to\infty} H(Y_n \mid \rho_n^\theta) - H(Y_n \mid X_n, \pi_n^\theta) \\ &= \int_{\rho\in\Delta} \sum_{y\in\mathcal{Y}} (-\log p^\theta(y \mid \rho)) p^\theta(y \mid \rho) \nu^\theta(d\rho) \\ &\quad - \int_{\pi\in\Delta} \sum_{\substack{y\in\mathcal{Y} \\ x\in\mathcal{X}}} (-\log p(y \mid x, \pi)) p(y \mid x, \pi) \theta(x) \mu^\theta(d\pi) \end{aligned} \tag{29}$$

where the θ superscript on ρ_n, π_n, and $p(y \mid \rho)$ shows their dependence on the input distribution, ν^θ denotes the limiting distribution of ρ_n^θ, and μ^θ denotes the limiting distribution of π_n^θ.

We now combine the above lemmas to get a closed form expression for the mutual information under i.i.d. inputs.

Theorem 4.2: For any i.i.d. input distribution $\theta \in \mathcal{P}(X)$, the average mutual information per channel use is given by

$$\begin{aligned} I_\theta &\triangleq \lim_{n\to\infty} \frac{1}{n} I_\theta(Y^n; X^n) \\ &= \int_\Delta \sum_{y\in\mathcal{Y}} (-\log p^\theta(y \mid \rho)) p^\theta(y \mid \rho) \nu^\theta(d\rho) \\ &\quad - \int_\Delta \sum_{\substack{y\in\mathcal{Y} \\ x\in\mathcal{X}}} (-\log p(y \mid x, \pi)) \\ &\quad \cdot p(y \mid x, \pi) \theta(x) \mu^\theta(d\pi). \end{aligned} \tag{30}$$

Proof: From (18)

$$I(Y^n; X^n) = H(Y^n) - H(Y^n \mid X^n).$$

If we fix $\theta \in \mathcal{P}(X)$

$$H(Y^n \mid X^n) = \sum_{i=1}^n H(Y_i \mid X_i, Y^{i-1}, X^{i-1}) \tag{31}$$

by (20), and the terms of the summation are nonnegative and monotonically decreasing in i by Lemma 4.2. Thus

$$\begin{aligned} &\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n H(Y_i \mid X_i, Y^{i-1}, X^{i-1}) \\ &\quad = \lim_{n\to\infty} H(Y_n \mid X_n, X^{n-1}, Y^{n-1}). \end{aligned} \tag{32}$$

Similarly, from (19)

$$H(Y^n) = \sum_{i=1}^n H(Y_i \mid Y^{i-1}) \tag{33}$$

and by Lemma 4.4, the terms of this summation are nonnegative and monotonically decreasing in i. Hence

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n H(Y_i \mid Y^{i-1}) = \lim_{n\to\infty} H(Y_n \mid Y^{n-1}). \tag{34}$$

Applying Lemmas 4.1 and 4.6 completes the proof. □

It is easily shown that since ν^θ and μ^θ are continuous functions of θ, I_θ is also. Moreover, the calculation of I_θ is relatively simple, since asymptotic values of μ and ν are obtained using the recursive formulas (12) and (14), respectively. For the channel described in Section VII, these recursive formulas closely approach their final values after only 40 iterations. Unfortunately, this simplified formula for mutual information under i.i.d. inputs cannot be extended to Markov inputs, since π_n and ρ_n are no longer Markov chains under these conditions.

We now consider the average mutual information maximized over all i.i.d. input distributions. Define

$$I_{\text{i.i.d.}} \triangleq \sup_{\theta\in\mathcal{P}(X)} I_\theta. \tag{35}$$

Since $\mathcal{P}(X)$ is compact and I_θ continuous in θ, $I_{\text{i.i.d.}}$ achieves its supremum on $\mathcal{P}(X)$, and the maximization can be done using standard techniques for continuous functions. Moreover, it is easily shown that $I_{\text{i.i.d}} \leq C$. Thus (35) provides a relatively simple formula to lower-bound the capacity of general FSMC's.

The next section will describe a class of channels for which uniform i.i.d. channel inputs achieve channel capacity. Thus $I_{\text{i.i.d.}} = C$, and the capacity can be found using the formula of Theorem 4.2. This channel class includes fading or variable-noise channels with symmetric PSK inputs, as well as channels which vary over a finite set of BSC's.

V. Uniformly Symmetric Variable-Noise Channels

In this section we define two classes of FSMC's: uniformly symmetric channels and variable-noise channels. The mutual information and capacity of these channel classes have additional properties which we outline in the lemmas below. Moreover, we will show in the next section that the decision-feedback decoder achieves capacity for uniformly symmetric variable-noise FSMC's.

Definition: For a DMC, let M denote the matrix of input/output probabilities

$$M_{ij} \overset{\Delta}{=} p(y=j \mid x=i), \quad j \in \mathcal{Y},\, i \in \mathcal{X}.$$

A discrete memoryless channel is *output-symmetric* if the rows of M are permutations of each other, and the columns of M are permutations of each other.[2]

Definition: A FSMC is *uniformly symmetric* if every channel $c_k \in \mathcal{C}$ is output-symmetric.

The next lemma, proved in Appendix V, shows that for uniformly symmetric FSMC's, the conditional output entropy is maximized with uniform i.i.d. inputs.

Lemma 5.1: For uniformly symmetric FSMC's and any initial state $S_0 = c_i$, $H(Y_n \mid \rho_n)$, $H(Y_n \mid \rho_n^i)$, $H(Y_n \mid \pi_n)$, and $H(Y_n \mid \pi_n^i)$ are all maximized for a uniform and i.i.d. input distribution, and these maximum values equal $\log |\mathcal{Y}|$.

Definition: Let X_n and Y_n denote the input and output, respectively, of an FSMC. We say that an FSMC is a *variable-noise channel* if there exists a function ϕ such that for $Z_n = \phi(X_n, Y_n)$, $p(Z^n \mid X^n) = p(Z^n)$, and Z^n is a sufficient statistic for S^n (so S^n is independent of X^n and Y^n given Z^n). Typically, ϕ is associated with an additive noise channel, as we discuss in more detail below.

If Z^n is a sufficient statistic for S^n, then

$$\begin{aligned}\pi_n &\overset{\Delta}{=} p(S_n \mid X^{n-1}, Y^{n-1})\\ &= p(S_n \mid X^{n-1}, Y^{n-1}, Z^{n-1}) = p(S_n \mid Z^{n-1}). \quad (36)\end{aligned}$$

Using (36) and replacing the pairs (X_n, Y_n) with Z_n in the derivation of Appendix I, we can simplify the recursive calculation of π_n

$$\pi_{n+1} = \frac{\pi_n D(z_n) P}{\pi_n D(z_n) \underline{1}} \overset{\Delta}{=} f(z_n, \pi_n) \qquad (37)$$

where $D(z_n)$ is a diagonal $K \times K$ matrix with kth diagonal term $p(z_n \mid S_n = c_k)$. The transition probabilities are also simplified

$$\begin{aligned}p(\pi_{n+1} &= \alpha \mid \pi_n = \beta)\\ &= \sum_{z_n \in \mathcal{Z}} 1[(z_n)\colon f(z_n, \beta) = \alpha] p(z_n \mid \pi_n = \beta). \quad (38)\end{aligned}$$

The next lemma, proved in Appendix V, shows that for a uniformly symmetric variable-noise channel, the output entropy conditioned on the input is independent of the input distribution.

Lemma 5.2: For uniformly symmetric variable-noise FSMC's and all i, $H(Y_n \mid X_n, \pi_n)$ and $H(Y_n \mid X_n, \pi_n^i)$ do not depend on the input distribution.

Consider an FSMC where each $c_k \in \mathcal{C}$ is an AWN channel with noise density n_k. If we let $Z = Y - X$, then it is easily shown that this is a variable-noise channel. However, such channels have an infinite output alphabet. In general, the output of an AWN channel is quantized to the nearest symbol in a finite output alphabet: we call this the quantized AWN (Q-AWN) channel.

If the Q-AWN channel has a symmetric multiphase input alphabet of constant amplitude and output phase quantization [4, p. 80], then it is easily checked that $p_k(y \mid x)$ depends only on $p_k(|y-x|)$, which in turn depends only on the noise density n_k. Thus it is a variable-noise channel.[3] We show in Appendix VI that variable-noise Q-AWN channels with the same input and output alphabets are also uniformly symmetric. Uniformly symmetric variable-noise channels have the property that $I_{\text{i.i.d.}}$ equals the channel capacity, as we show in the following theorem.

Theorem 5.1: Capacity of uniformly symmetric variable-noise channels is achieved with an input distribution that is uniform and i.i.d. The capacity is given by

$$C = I_{\text{i.i.d.}} = \log|\mathcal{Y}| - p\left[\int_{\Delta} \sum_{y \in \mathcal{Y}} -\log p(y \mid x, \pi) \cdot p(y \mid x, \pi)\mu(d\pi)\right] \quad \forall x \in \mathcal{X} \quad (39)$$

where μ is the limiting distribution for π_n under uniform i.i.d. inputs. Moreover, $C = \lim_{n\to\infty} C_n = \lim_{n\to\infty} C_n^i$ for all i, where

$$C_n \overset{\Delta}{=} \max_{\mathcal{P}(X^n)} H(Y_n \mid \rho_n) - H(Y_n \mid X_n, \pi_n) \qquad (40)$$

increases with n, and

$$C_n^i \overset{\Delta}{=} \max_{\mathcal{P}(X^n)} H(Y_n \mid \rho_n^i) - H(Y_n \mid X_n, \pi_n^i) \qquad (41)$$

decreases with n.

Proof: From Lemmas 5.1 and 5.2, C_n, C_n^i, and C are all maximized with uniform i.i.d. inputs. With this input distribution

$$C_n = \log|\mathcal{Y}| - H(Y_n \mid X_n, \pi_n)$$

and

$$C_n^i = \log|\mathcal{Y}| - H(Y_n \mid X_n, \pi_n^i).$$

Applying Lemmas 4.2 and 4.3, we get that $H(Y_n \mid X_n, \pi_n)$ decreases with n, $H(Y_n \mid X_n, \pi_n^i)$ increases with n, and both

[2] Symmetric channels, defined in [13, p. 94], are a more general class of memoryless channels; an output-symmetric channel is a symmetric channel with a single output partition.

[3] If the input alphabet of a Q-AWN channel is not symmetric or the input symbols have different amplitudes, then the distribution of $Z = |Y - X|$ will depend on the input. To see this, consider a Q-AWN channel with a 16-QAM input/output alphabet (so the output is quantized to the nearest input symbol). There are four different sets of $Z = |Y - X|$ values, depending on the amplitude of the input symbol. Thus the distribution of Z over all its possible values (the union of all four sets) will change, depending on the amplitude of the input symbol.

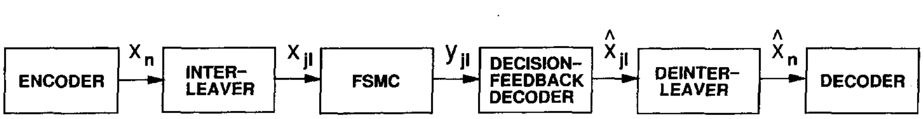

Fig. 3. System model.

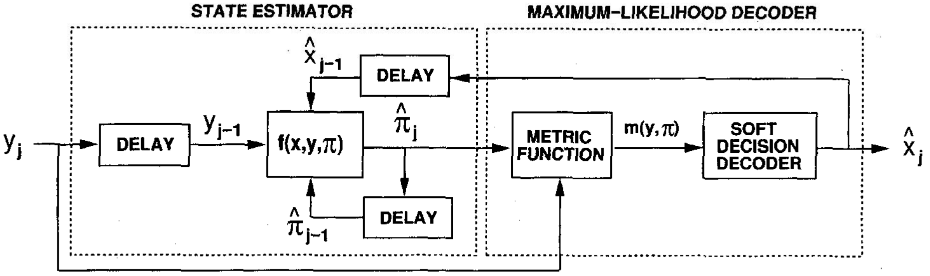

Fig. 4. Decision-feedback decoder.

converge to the same limit. Finally, under uniform i.i.d. inputs

$$\begin{aligned} C &= \log|\mathcal{Y}| - \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} H(Y_i \mid X_i, \pi_i) \\ &= \log|\mathcal{Y}| - \lim_{n\to\infty} H(Y_n \mid X_n, \pi_n) \end{aligned} \tag{42}$$

by Lemma 4.1 and (32). Applying Lemma 4.6 to

$$\lim_{n\to\infty} H(Y_n \mid X_n, \pi_n)$$

completes the proof. □

The BSC is equivalent to a binary-input Q-AWN channel with binary quantization [4]. Thus an FSMC where c_k indexes a set of BSC's with different crossover probabilities is a uniformly symmetric variable-noise channel. Therefore, both [1, Proposition 4] and the capacity formula obtained in [5] are corollaries of Theorem 5.1.

VI. Decision-Feedback Decoder

A block diagram for a system with decision-feedback decoding is depicted in Fig. 3. The system is composed of a conventional (block or convolutional) encoder for memoryless channels, block interleaver, FSMC, decision-feedback decoder, and deinterleaver. Fig. 4 outlines the decision-feedback decoder design, which consists of a channel state estimator followed by an ML decoder. We will show in this section that if we ignore error propagation, a system employing this decision-feedback decoding scheme on uniformly symmetric variable-noise channels is *information-lossless*: it has the same capacity as the original FSMC, given by (30) for i.i.d. uniform inputs. Moreover, we will see that the output of the state estimator is a sufficient statistic for the current output given all past inputs and outputs, which reduces the system of Fig. 3 to a discrete memoryless channel. Thus the ML input sequence is determined on a symbol-by-symbol basis, eliminating the complexity and delay of sequence decoders.

The interleaver works as follows. The output of the encoder is stored row by row in a $J \times L$ interleaver, and transmitted over the channel column by column. The deinterleaver performs the reverse operation. Because the effect of the initial channel state dies away, the received symbols within any row of the deinterleaver become independent as J becomes infinite. However, the symbols within any column of the deinterleaver are received from consecutive channel uses, and are thus dependent. This dependence is called the *latent* channel memory, and the state estimator enables the ML decoder to make use of this memory.

Specifically, the state estimator uses the recursive relationship of (11) to estimate π_n. It will be shown below that the ML decoder operates on a memoryless system, and can therefore determine the ML input sequence on a per-symbol basis. The input to the ML decoder is the channel output y_n and the state estimate $\hat{\pi}_n$, and its output is the x_n which maximizes $\log p(y_n, \hat{\pi}_n \mid x_n)$, assuming equally likely input symbols.[4] The soft-decision decoder uses conventional techniques (e.g., Viterbi decoding) with branch metrics

$$m(y,\pi) \triangleq \log p(y,\pi \mid x). \tag{43}$$

We now evaluate the information, capacity, and cutoff rates of a system using the decision-feedback decoder, assuming $\hat{\pi}_n = \pi_n$ (i.e., ignoring error propagation). We will use the notation $y_{jl} \triangleq y_n$ to explicitly denote that y_n is in the jth row and lth column of the deinterleaver. Similarly, $\pi_{jl} \triangleq \pi_n$ and $x_{jl} \triangleq x_n$ denote, respectively, the state estimate and interleaver input corresponding to y_{jl}. Assume now that the state estimator is reset every J iterations so, for each l, the state estimator goes through j recursions of (11) to calculate π_{jl}. By (12), this recursion induces a distribution $p(\pi_{jl})$ on π_{jl} that depends only on $p(X^{j-1})$. Thus the system up to the output of the state estimator is equivalent to a set of parallel π-output channels, where the π-output channel is defined, for a given j, by the input x_{jl}, the output pair (y_{jl}, π_{jl}), and the input/output probability

$$p(y_{jl}, \pi_{jl} \mid x_{jl}) = \sum_k p_k(y_{jl} \mid x_{jl})\pi_{jl}(k)p(\pi_{jl}). \tag{44}$$

[4] If the x_n are not equally likely, then $\log p(x_n)$ must be added to the decoder metric.

For each j, the π-output channel is the same for $l = 1, 2, \cdots, L$, and therefore there are J different π-output channels, each used L times. We thus drop the l subscript of x_{jl}, y_{jl}, and π_{jl} in the decoder block diagram of Fig. 4. The first π-output channel ($j = 1$) is equivalent to the FSMC with interleaving and memoryless channel encoding, since the estimator is reset and therefore $\pi_{1l} = \pi_0, 1 \leq l \leq L$.

The jth π-output channel is discrete, since x_{jl} and y_{jl} are taken from finite alphabets, and since π_{jl} can have at most $|\mathcal{X}|^j|\mathcal{Y}|^j$ different values. It is also asymptotically memoryless with deep interleaving (large J), which we prove in Appendix VII. Finally, we show in Appendix VIII that for a fixed input distribution, the J π-output channels are independent, and the average mutual information of the parallel channels is

$$\begin{aligned} I_J &= \frac{1}{J} I(Y^J, \pi^J; X^J) \\ &= \frac{1}{J} \sum_{j=1}^{J} H(Y_j \mid \pi_j) - H(Y_j \mid X_j, \pi_j). \end{aligned} \tag{45}$$

Let

$$\begin{aligned} C_J &\triangleq \max_{\mathcal{P}(X^J)} \frac{1}{J} \sum_{j=1} H(Y_j \mid \pi_j) - H(Y_j \mid X_j, \pi_j) \\ &= \max_{\mathcal{P}(X^J)} \frac{1}{J} \sum_{j=1}^{J} C_j \end{aligned} \tag{46}$$

where

$$C_j \triangleq H(Y_j \mid \pi_j) - H(Y_j \mid X_j, \pi_j) \tag{47}$$

for the maximizing distribution $p(X^J)$. The capacity of the decision-feedback decoding system is then

$$C_{\text{df}} = \lim_{J \to \infty} C_J. \tag{48}$$

Comparing (48) to (24), we see that the capacity penalty of the decision-feedback decoder is given by

$$\begin{aligned} C - C_{\text{df}} = \lim_{n \to \infty} \Bigg[& \max_{\mathcal{P}(X^n)} \left(\frac{1}{n} \sum_{j=1}^{n} H(Y_j \mid \rho_j) - H(Y_j \mid X_j, \pi_j) \right) \\ & - \max_{\mathcal{P}(X^n)} \left(\frac{1}{n} \sum_{j=1}^{n} H(Y_j \mid \pi_j) - H(Y_j \mid X_j, \pi_j) \right) \Bigg]. \end{aligned} \tag{49}$$

For uniformly symmetric variable-noise channels, uniform i.i.d. inputs achieve both C and C_{df}, and with this input $C - C_{\text{df}} = 0$. Thus the decision-feedback decoder preserves the inherent capacity of such channels.

Although capacity gives the maximum data rate for any ML encoding scheme, established coding techniques generally operate at or below the channel cutoff rate [4]. Since the π-output channels are independent for a fixed input distribution $p(X^J)$, the random coding exponent for the parallel set is

$$\boldsymbol{E}_o(1, p(X^J)) = \sum_{j=1}^{J} R_j \tag{50}$$

where

$$R_j = -\log \left(\int_{\pi_j \in \Delta} \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} p(x) \sqrt{\sum_{k=1}^{K} p_k(y \mid x) \pi_j(k) p(d\pi_j)} \right]^2 \right). \tag{51}$$

The cutoff rate of the decision-feedback decoding system is

$$R_{\text{df}} \triangleq \lim_{J \to \infty} \max_{\mathcal{P}(X^J)} \frac{1}{J} \sum_{j=1}^{J} R_j. \tag{52}$$

We show in Appendix IX that for uniformly symmetric variable-noise channels, the maximizing input distribution in (52) is uniform and i.i.d., the resulting value of R_j is increasing in j, and the cutoff rate R_{df} becomes

$$\begin{aligned} R_{\text{df}} &= \lim_{j \to \infty} R_j \\ &= -\log \left(\int_{\pi \in \Delta} \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \frac{1}{|\mathcal{X}|} \cdot \sqrt{\sum_{k=1}^{K} p_k(y \mid x) \pi(k) \mu(d\pi)} \right]^2 \right) \end{aligned} \tag{53}$$

where μ is the invariant distribution for π under i.i.d. uniform inputs.

Our calculations throughout this section have ignored the impact of error propagation. Referring to Fig. 4, error propagation occurs when the decision-feedback decoder output for the maximum-likelihood input symbol $\hat{x}_j$ is in error, which will then cause the estimate of $\hat{\pi}_j$ to be in error. Since x_j is the value of the coded symbol, the error probability for $\hat{x}_j$ does not benefit from any coding gain. Unfortunately, since block or convolutional decoding introduces delay, the post-decoding decisions cannot be fed back to the decision-feedback decoder to update the $\hat{\pi}_j$ value. This is exactly the difficulty faced by an adaptive decision-feedback equalizer (DFE), where decoding decisions are used to update the DFE tap coefficients [16]. New methods to combine DFE's and coding have recently been proposed, and several of these methods can be used to obtain some coding gain in the estimate of x_j fed back through our decision-feedback decoder. In particular, the structure of our decision-feedback decoder already includes the interleaver/deinterleaver pair proposed by Eyuboğlu for DFE's with coding [17]. In his method, this pair introduced a periodic delay in the received bits such that delayed reliable decisions can be used for feedback. Applying this idea to our system effectively combines the decision-feedback decoder, deinterleaver, and decoder. Specifically, the symbols transmitted over each π-output channel are decoded together, and the symbol decisions output from the decoder are then used by the decision-feedback decoder to update the π values of the subsequent π-output channel. The complexity and delay of this design increases linearly with the block length of the π-output channel code, but it is independent of the

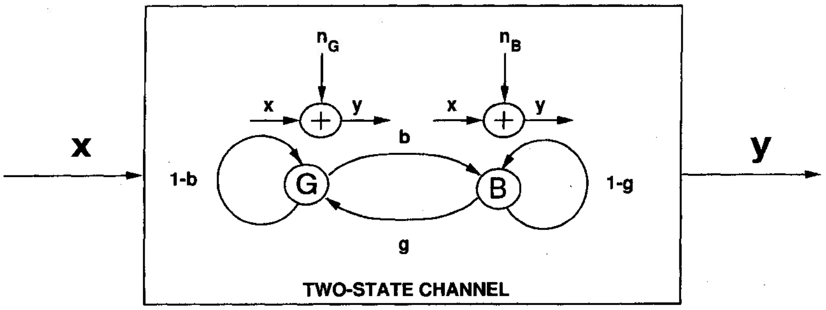

Fig. 5. Two-state fading channel.

channel memory since this memory is captured in the sufficient statistic π_n. Another approach to implement coding gain uses soft decisions on the received symbols to update π_n, then later corrects this initial π_n estimate if the decoded symbols differ from their initial estimates [18]. This method truncates the number of symbols affected by an incorrect decision, at a cost of increased complexity to recalculate and update the π_n values. Finally, decision-feedback decoding can be done in parallel, where each parallel path corresponds to a different estimate of the received symbol. The number of parallel paths will grow exponentially in this case, however we may be able to apply some of the methods outlined in [19] and [20] to reduce the number of paths sustained through the trellis.

VII. Two-State Variable-Noise Channel

We now compute the capacity and cutoff rates of a two-state Q-AWN channel with variable SNR, Gaussian noise, and 4PSK modulation. The variable SNR can represent different fading levels in a multipath channel, or different noise and/or interference levels. The model is shown in Fig. 5. The input to the channel is a 4PSK symbol, to which noise of variance n_G or n_B is added, depending on whether the channel is in state G (good) or B (bad). We assume that the SNR is 10 dB for channel G, and -5 dB for channel B. The channel output is quantized to the nearest input symbol and, since this is a uniformly symmetric variable-noise channel, the capacity and cutoff rates are achieved with uniform i.i.d. inputs. The state transition probabilities are depicted in Fig. 5. We assume a stationary initial distribution of the state process, so $p(S_0 = G) = g/(g+b)$ and $p(S_0 = B) = b/(g+b)$.

Fig. 6 shows the iterative calculation of (12) for $p(\pi_n(G) = \alpha)$, where

$$\pi_n(G) = p(S_n = G \mid x^{n-1}, y^{n-1}).$$

In this example, the difference of subsequent distributions after 40 recursions is below the quantization level ($d\alpha = 0.01$) of the graph. Fig. 7 shows the capacity (C_j) and cutoff rate (R_j) of the jth π-output channel, given by (47) and (52), respectively. Note that $C_{j=1}$ and $R_{j=1}$ in this figure are the capacity and cutoff rate of the FSMC with interleaving and memoryless channel encoding. Thus the difference between the initial and final values of C_j and R_j indicate the performance improvement of the decision-feedback decoder over conventional techniques.

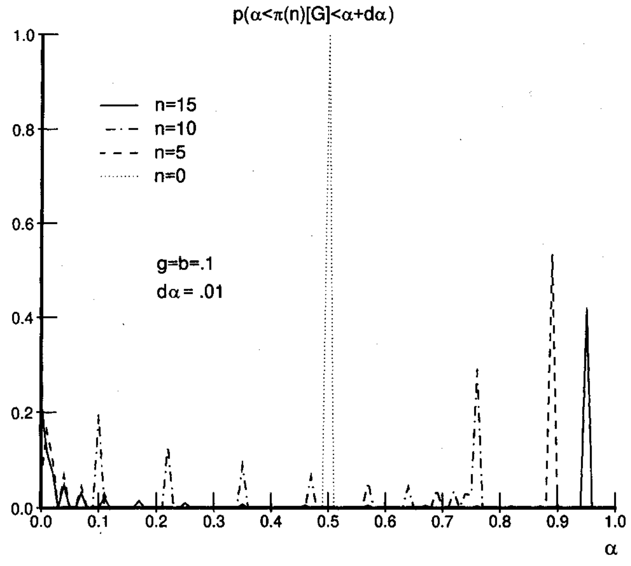

Fig. 6. Recursive distribution of π_n.

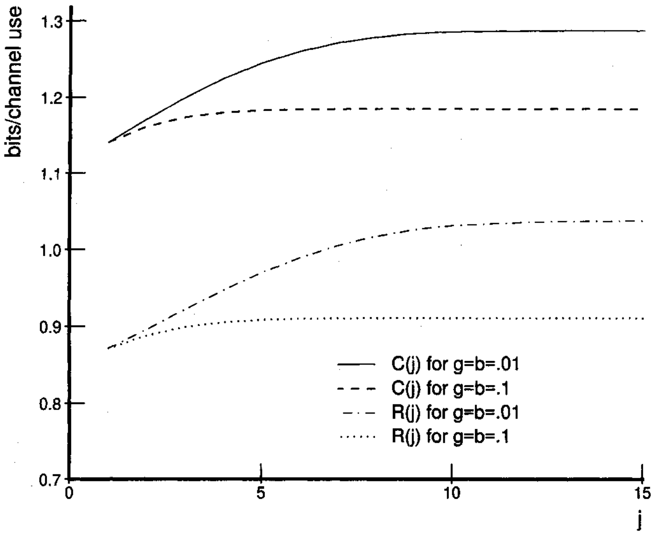

Fig. 7. Capacity and cutoff rate for jth π-output channel.

For this two-state model, the channel memory can be quantified by the parameter $\mu \triangleq 1-g-b$, since for $\sigma \in \{G, B\}$ [1]

$$p(S_n = \sigma \mid S_0 = \sigma) - p(S_n = \sigma \mid S_0 \neq \sigma) = \mu^n. \quad (54)$$

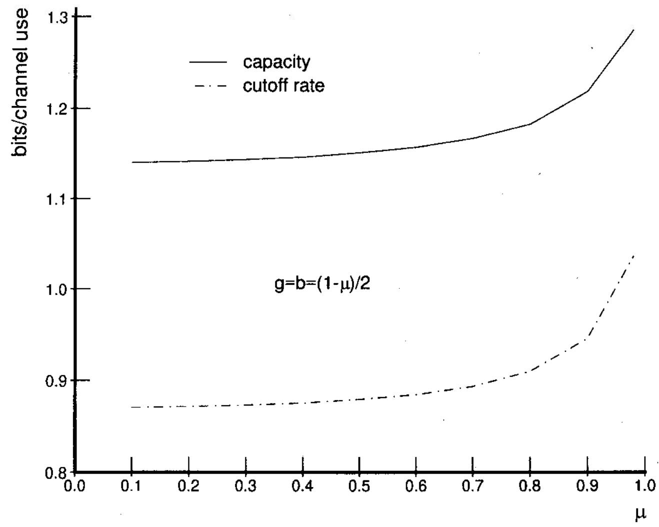

Fig. 8. Decoder performance versus channel memory.

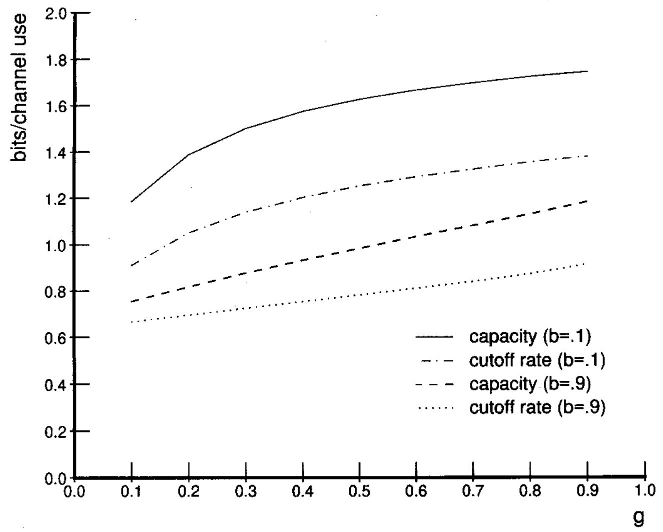

Fig. 9. Decoder performance versus g.

In Fig. 8 we show the decision-feedback decoder's capacity and cutoff rates (C_{df} and R_{df}, respectively) as functions of μ. We expect these performance measures to increase as μ increases, since more latency in the channel should improve the accuracy of the state estimator; Fig. 8 confirms this hypothesis. Finally, in Fig. 9 we show the decision-feedback decoder's capacity and cutoff rates as functions of g. The parameter g is inversely proportional to the average number of consecutive B channel states (which corresponds to a 15 dB fade), thus Fig. 9 can be interpreted as the relationship between the maximum transmission rate and the average fade duration.

VIII. Summary

We have derived the Shannon capacity of an FSMC as a function of the conditional probabilities

$$\rho_n(k) = p(S_n = c_k \mid y^{n-1})$$

and

$$\pi_n(k) = p(S_n = c_k \mid x^{n-1}, y^{n-1}).$$

We also showed that with i.i.d. inputs, these conditional probabilities converge weakly, and the channel's mutual information under this input constraint is then a closed-form continuous function of the input distribution. This continuity allows $I_{\text{i.i.d.}}$, the maximum mutual information of the FSMC over all i.i.d. inputs, to be found using standard maximization techniques. Additional properties of the entropy and capacity for uniformly symmetric variable-noise channels were also derived.

We then proposed an ML decision-feedback decoder, which calculates recursive estimates of π_n from the channel output and the decision-feedback decoder output. We showed that for asymptotically deep interleaving, a system employing the decision-feedback decoder is equivalent to a discrete memoryless channel with input x_n and output (y_n, π_n). Thus the ML sequence decoding can be done on a symbol-by-symbol basis. Moreover, the decision-feedback decoder preserves the inherent capacity of uniformly symmetric variable-noise channels, assuming the effect of error propagation is negligible. This class of FSMC's includes fading or variable-noise channels with symmetric PSK inputs as well as channels which vary over a finite set of BSC's. For general FSMC's, we obtained the capacity and cutoff rate penalties of the decision-feedback decoding scheme.

We also presented numerical results for the performance of the decision-feedback decoder on a two-state variable-noise channel with 4PSK modulation. These results demonstrate significant improvement over conventional schemes which use interleaving and memoryless channel encoding, and the improvement is most pronounced on quasistatic channels. This result is intuitive, since the longer the FSMC stays in a given state, the more accurately the state estimator will predict that state. Finally, we present results for the decoder performance relative to the average fade duration; as expected, the performance improves as the average fade duration decreases.

Appendix I

In this Appendix, we derive the recursive formula (11) for π_n. First, we have (55) at the top of the following page, where a, b, and d follow from Bayes rule, and c follows from (5). Moreover

$$\begin{aligned} p(x^n, y^n) &= \sum_{k\in K} p(x^n, y^n, S_n = c_k) \\ &= \sum_{k\in K} p(x_n, y_n \mid S_n = c_k, x^{n-1}, y^{n-1}) \\ &\quad \cdot p(S_n = c_k, x^{n-1}, y^{n-1}) \\ &= \sum_{k\in K} p(y_n \mid S_n = c_k, x_n, x^{n-1}, y^{n-1}) \\ &\quad \cdot p(x_n \mid S_n, x^{n-1}, y^{n-1}) p(S_n = c_k, x^{n-1}, y^{n-1}) \\ &= \sum_{k\in K} p(y_n \mid S_n = c_k, x_n) p(x_n \mid x^{n-1}) \\ &\quad \cdot p(S_n = c_k \mid x^{n-1}, y^{n-1}) p(x^{n-1}, y^{n-1}). \end{aligned} \tag{56}$$

$$\begin{aligned}p(S_n \mid x^n, y^n) &\overset{a}{=} \frac{p(x_n, y_n \mid S_n, x^{n-1}, y^{n-1})p(S_n, x^{n-1}, y^{n-1})}{p(x^n, y^n)}\\ &\overset{b}{=} \frac{p(y_n \mid S_n, x_n, x^{n-1}, y^{n-1})p(x_n \mid S_n, x^{n-1}, y^{n-1})p(S_n, x^{n-1}, y^{n-1})}{p(x^n, y^n)}\\ &\overset{c}{=} \frac{p(y_n \mid S_n, x_n)p(x_n \mid x^{n-1})p(S_n, x^{n-1}, y^{n-1})}{p(x^n, y^n)}\\ &\overset{d}{=} \frac{p(y_n \mid S_n, x_n)p(x_n \mid x^{n-1})p(S_n \mid x^{n-1}, y^{n-1})p(x^{n-1}, y^{n-1})}{p(x^n, y^n)}\end{aligned} \tag{55}$$

where we again use Bayes rule and the last equality follows from (5). Substituting (56) in the denominator of (55), and canceling the common terms $p(x_n \mid x^{n-1})$ and $p(x^{n-1}, y^{n-1})$ yields

$$\begin{aligned}&p(S_n \mid x^n, y^n)\\ &\quad = \frac{p(y_n \mid S_n, x_n)p(S_n \mid x^{n-1}, y^{n-1})}{\sum\limits_{k \in K} p(y_n \mid S_n = c_k, x_n)p(S_n = c_k \mid x^{n-1}, y^{n-1})}\end{aligned} \tag{57}$$

which, for a particular value of S_n, becomes

$$\begin{aligned}&p(S_n = c_l \mid x^n, y^n)\\ &\quad = \frac{p(y_n \mid S_n = c_l, x_n)p(S_n = c_l \mid x^{n-1}, y^{n-1})}{\sum\limits_{k \in K} p(y_n \mid S_n = c_k, x_n)p(S_n = c_k \mid x^{n-1}, y^{n-1})}.\end{aligned} \tag{58}$$

Finally, from (4)

$$p(S_{n+1} = c_l \mid x^n, y^n) = \sum_{j \in K} p(S_n = c_j \mid x^n, y^n)P_{jl}. \tag{59}$$

Substituting this into (58) yields the desired result.

Appendix II

In this Appendix we show that for i.i.d. inputs, π_n and ρ_n are Markov chains that converge in distribution to a limit which is independent of the initial channel state, and that the resulting limit distributions are continuous functions of the input distribution $p(x)$. We also show that the Markov property does not hold for Markov inputs.

We begin by showing the Markov property for independent inputs.

Lemma A2.1: For independent inputs, π_n is a Markov chain.

Proof:

$$\begin{aligned}p(\pi_{n+1} \mid \pi_n, \cdots, \pi_0) &= \sum_{x_n, y_n} p(\pi_{n+1} \mid \pi_n, \cdots, \pi_0, x_n, y_n)\\ &\qquad \cdot p(x_n, y_n \mid \pi_n, \cdots, \pi_0)\\ &= \sum_{x_n, y_n} p(\pi_{n+1} \mid \pi_n, x_n, y_n)p(x_n, y_n \mid \pi_n)\\ &= p(\pi_{n+1} \mid \pi_n)\end{aligned} \tag{60}$$

where the second equality follows from (11) and (6). Thus π_n is Markov. A similar argument using (13) and (8) shows that ρ_n is also Markov for independent inputs.

To obtain the weak convergence of π_n and ρ_n, we also assume that the channel inputs are i.i.d., since we can then apply convergence results for partially observed Markov chains [21]. Consider the new stochastic process $U_n \triangleq (S_n, y_n, x_n)$ defined on the state space $\mathcal{U} = \mathcal{C} \times \mathcal{Y} \times \mathcal{X}$. Since S_n is stationary and ergodic and x_n is i.i.d., U_n is stationary and ergodic. It is easily checked that U_n is Markov.

Let $(S, y, x)_j$ denote the jth element of $\mathcal{U}$, and $J \triangleq |\mathcal{U}|$. To specify its individual components, we use the notation

$$(S_{(j)}, y_{(j)}, x_{(j)}) \triangleq (S, y, x)_j.$$

The $J \times J$ probability transition matrix for U, P^U, is

$$\begin{aligned}P_{kj}^U = p[(S_{n+1}, y_{n+1}, x_{n+1}) = (S, y, x)_j \mid (S_n, y_n, x_n)\\ = (S, y, x)_k)]\end{aligned} \tag{61}$$

independent of n. The initial distribution of U, π_0^U, is given by

$$p(S_0 = c_k, y_0 = y, x_0 = x) = \pi_0(k)p_k(y_0 \mid x_0)p(x_0). \tag{62}$$

Let $g_{y,x}\colon \mathcal{U} \to \mathcal{Y} \times \mathcal{X}$ and $g_y : \mathcal{U} \to \mathcal{Y}$ be the projections

$$g_{y,x}(S_n, y_n, x_n) = (y_n, x_n)$$

and

$$g_y(S_n, y_n, x_n) = (y_n).$$

These projections form the new processes $W_n = g_{y,x}[U_n]$ and $V_n = g_y[U_n]$. We regard W_n and V_n as partial observations of the Markov chain U_n; the pairs (U_n, W_n) and (U_n, V_n) are referred to as partially observed Markov chains. The distribution of U_n conditioned on W_n and V_n, respectively, is

$$\pi_n^U = (\pi_n^U(1), \cdots, \pi_n^U(J))$$

and

$$\rho_n^U = (\rho_n^U(1), \cdots, \rho_n^U(J))$$

where

$$\pi_n^U(j) = p(U_n = (S, y, x)_j \mid W^n) \tag{63}$$

and

$$\rho_n^U(j) = p(U_n = (S, y, x)_j \mid V^n). \tag{64}$$

Note that

$$\begin{aligned}\pi_n^U(j) &= p(U_n = (S, y, x)_j \mid W^n)\\ &= p(S_n = S_{(j)} \mid x^n, y^n)1[x_n = x_{(j)}, y_n = y_{(j)}]\\ &= \pi_n(k)1[x_n = x_{(j)}, y_n = y_{(j)}]\end{aligned} \tag{65}$$

where $S_{(j)} = c_k$. Thus if π_n^U converges in distribution, π_n must also converge in distribution. Similarly, ρ_n converges in distribution if ρ_n^U does.

We will use the following definition for subrectangular matrices in the subsequent theorem.

Definition: Let $D = (D_{ij})$ denote a $d \times d$ matrix. If $D_{i_1,j_1} \neq 0$ and $D_{i_2,j_2} \neq 0$ implies that also $D_{i_1,j_2} \neq 0$ and $D_{i_2,j_1} \neq 0$, then D is called a subrectangular matrix.

We can now state the convergence theorem, due to Kaijser [21], for the distribution of a Markov chain conditioned on partial observations.

Theorem A2.1: Let U_n be a stationary and ergodic Markov chain with transition matrix P^U and state space $\mathcal{U}$. Let g be a function with domain $\mathcal{U}$ and range $\mathcal{Z}$. Define a new process $Z_n = g(U_n)$. For $z \in \mathcal{Z}$ and $U_{(j)}$ the jth element of $\mathcal{U}$, define matrix $M(z)$ by

$$M_{i,j}(z) = \begin{cases} P_{ij}^U, & \text{if } g[U_{(j)}] = z \\ 0, & \text{otherwise.} \end{cases} \tag{66}$$

Suppose that P^U and g are such that there exists a finite sequence $z_1, \cdots, z_m$ of elements in $\mathcal{Z}$ that yield a nonzero subrectangular matrix for the matrix product $M(z_1) \cdots M(z_m)$. Then $p(U_n \mid Z^n)$ converges in distribution and moreover the limit distribution is independent of the initial distribution of U.

We first apply this theorem to π_n^U.

Assumption 1: Assume that there exists a finite sequence $(y_n, x_n), n = 1, \cdots, m$, such that the matrix product $M(y_1, x_1) \cdots M(y_m, x_m)$ is nonzero and subrectangular, where

$$M_{i,j}(y,x) = \begin{cases} P_{ij}^U, & \text{if } g_{y,x}[(S,y,x)_j] = (y,x) \\ 0, & \text{otherwise.} \end{cases} \tag{67}$$

Then by Theorem A2.1, π_n^U converges in distribution to a limit which is independent of the initial distribution. By (65), this implies that π_n also converges in distribution, and its limit distribution is independent of π_0. We thus get the following lemma, which was stated in (15).

Lemma A2.2: For any bounded continuous function f, the following limits exist and are equal for all i

$$\lim_{n\to\infty} E[f(\pi_n)] = \lim_{n\to\infty} E[f(\pi_n^i)]. \tag{68}$$

The subrectangularity condition on M is satisfied if for some input $x \in \mathcal{X}$ there exists a $y \in \mathcal{Y}$ such that $p_k(y \mid x) > 0$ for all k. It is also satisfied if all the elements of the matrix P are nonzero.

From (11) and (12), the limit distribution of π_n is a function of the i.i.d. input distribution. Let $\mathcal{P}(X)$ denote the set of all possible distributions on X. The following lemma, proved below, shows that the limit distribution of π_n is continuous on $\mathcal{P}(X)$.

Lemma A2.3: Let μ^θ denote the limit distribution of π_n as a function of the i.i.d. distribution $\theta \in \mathcal{P}(X)$. Then μ^θ is a continuous function of θ, i.e., $\theta_m \to \theta$ implies that $\mu^{\theta_m} \to \mu^\theta$.

We now consider the convergence and continuity of the distribution for ρ_n. Define the matrix N by

$$N_{i,j}(y) = \begin{cases} P_{ij}^U, & \text{if } g_y[(S,y,x)_j] = y \\ 0, & \text{otherwise.} \end{cases} \tag{69}$$

and note that for any $y \in \mathcal{Y}$ and $x \in \mathcal{X}$

$$M_{i,j}(y,x) = N_{i,j}(y) I(x_{(j)} = x). \tag{70}$$

To apply Theorem A2.1 to ρ_n^U, we must find a sequence $y_1, \cdots, y_l$ which yields a nonzero and subrectangular matrix for the product $N(y_1) \cdots N(y_l)$. Consider the projection onto $\mathcal{Y}$ of the sequence $(y_n, x_n), n = 1, \cdots, m$, from Assumption 1. Let $y_n, n = 1, \cdots, m$ denote this projection. Using (70) and the fact that all the elements of M are nonnegative, it is easily shown that for $M \triangleq M(y_1, x_1) \cdots M(y_m, x_m)$ and $N \triangleq N(y_1) \cdots N(y_m)$, if for any i and j, $M_{i,j}$ is nonnegative, then $N_{i,j}$ is nonnegative also. From this we deduce that if M is nonzero and subrectangular, then N must also be nonzero and subrectangular.

We can now apply Theorem A2.1 to ρ_n^U, which yields the convergence in distribution of ρ_n^U and thus ρ_n. Moreover, the limit distributions of these random vectors are independent of their initial states. Thus we get the following result, which was stated in (16).

Lemma A2.4: For any bounded continuous function f, the following limits exist and are equal for all i:

$$\lim_{n\to\infty} E[f(\rho_n)] = \lim_{n\to\infty} E[f(\rho_n^i)]. \tag{71}$$

From (13) and (14), the limit distribution of ρ_n is also a function of the input distribution. The following lemma shows that the limit distribution of ρ_n is continuous on $\mathcal{P}(X)$.

Lemma A2.5: Let ν^θ denote the limit distribution of ρ_n as a function of the i.i.d. distribution $\theta \in \mathcal{P}(\mathcal{X})$. Then ν^θ is a continuous function of θ, so $\theta_m \to \theta$ implies that $\nu^{\theta_m} \to \nu^\theta$.

Proof of Lemmas A2.3 and A2.5: We must show that for all $\theta_m, \theta \in \mathcal{P}(X)$, if $\theta_m \to \theta$, then $\mu^{\theta_m} \to \mu^\theta$ and $\nu^{\theta_m} \to \nu^\theta$. We first show the convergence of ν^{θ_m}. From [12, p. 346], in order to show that $\nu^{\theta_m} \to \nu^\theta$, it suffices to show that $\{\nu^{\theta_m}\}$ is a tight sequence of probability measures,[5] and that any subsequence of ν^{θ_m} which converges weakly converges to ν^θ.

Tightness of the sequence $\{\nu^{\theta_m}\}$ follows from the fact that Δ is a compact set. Now suppose there is a subsequence $\nu^{\theta_{m_k}} \triangleq \nu^{\theta_k}$ which converges weakly to ψ. We must show that $\psi = \nu^\theta$, where ν^θ is the unique invariant distribution for ρ under the transformation (14) with input distribution $p(x) = \theta$. Thus it suffices to show that for every bounded, continuous, real-valued function ϕ on Δ,

$$\int_\Delta \phi(\alpha)\psi(d\alpha) = \int_\Delta \int_\Delta \phi(\alpha)\psi(d\beta)p^\theta(d\alpha \mid \beta) \tag{72}$$

where $p^\theta(\alpha \mid \beta) \triangleq p(\rho_{n+1} = \alpha \mid \rho_n = \beta)$ is given by (14) under the i.i.d. input distribution θ, and is thus independent of n. Applying the triangle inequality we get that for any k

$$\begin{aligned} &\left| \int_\Delta \phi(\alpha)\psi(d\alpha) - \int_\Delta \int_\Delta \phi(\alpha)\psi(d\beta)p^\theta(d\alpha \mid \beta) \right| \\ &\quad \leq \left| \int_\Delta \phi(\alpha)\psi(d\alpha) - \int_\Delta \phi(\alpha)\nu^{\theta_k}(d\alpha) \right| \end{aligned} \tag{73}$$

[5] A sequence of probability measures $\{\nu_m\}$ is tight if for all $\epsilon > 0$ there exists a compact set K such that $\nu(K) > 1 - \epsilon$ for all $\nu \in \{\nu_m\}$.

$$+\left|\int_{\Delta}\phi(\alpha)\nu^{\theta_k}(d\alpha) - \int_{\Delta}\int_{\Delta}\phi(\alpha)\nu^{\theta_k}(d\beta)p^{\theta_k}(d\alpha \mid \beta)\right| \qquad (74)$$

$$+\left|\int_{\Delta}\int_{\Delta}\phi(\alpha)\nu^{\theta_k}(d\beta)p^{\theta_k}(d\alpha \mid \beta) - \int_{\Delta}\int_{\Delta}\phi(\alpha)\psi(d\beta)p(d\alpha \mid \beta)\right|. \qquad (75)$$

Since this inequality holds for all k, in order to show (72), we need only show that the three terms (73)–(75) all converge to zero as $k \to \infty$. But (73) converges to zero since ν^{θ_k} converges weakly to ψ. Moreover, (74) equals zero for all k, since ν^{θ_k} is the invariant ρ distribution under the transformation (14) with input distribution θ_k. Substituting (14) for $p^{\theta}(\alpha \mid \beta)$ in (75) yields

$$\left|\int_{\Delta}\int_{\Delta}\phi(\alpha)\nu^{\theta_k}(d\beta)p^{\theta_k}(d\alpha \mid \beta) - \int_{\Delta}\int_{\Delta}\phi(\alpha)\psi(d\beta)p^{\theta}(d\alpha \mid \beta)\right|$$
$$= \left|\sum_{y\in\mathcal{Y}}\int_{\Delta}\phi(f^{\theta_k}(y,\beta))p^{\theta_k}(y \mid \beta)\nu^{\theta_k}(d\beta) - \sum_{y\in\mathcal{Y}}\int_{\Delta}\phi(f^{\theta}(y,\beta))p^{\theta}(y \mid \beta)\psi(d\beta)\right| \qquad (76)$$

where f^{θ} is given by (13) with $p(x) = \theta$, and

$$p^{\theta}(y \mid \beta) = \sum_{x\in\mathcal{X}}\sum_{k=1}^{K} p(y \mid x, S = c_k)\beta(k)\theta(x). \qquad (77)$$

Since $\mathcal{Y}$ is a finite set, (76) converges to zero if for every $y \in \mathcal{Y}$

$$\left|\int_{\Delta}\phi(f^{\theta_k}(y,\beta))p^{\theta_k}(y \mid \beta)\nu^{\theta_k}(d\beta) - \int_{\Delta}\phi(f^{\theta}(y,\beta))p^{\theta}(y \mid \beta)\psi(d\beta)\right| \to 0. \qquad (78)$$

Fix an arbitrary $y \in \mathcal{Y}$. Then applying the triangle inequality to (78) yields

$$\left|\int_{\Delta}\phi(f^{\theta_k}(y,\beta))p^{\theta_k}(y \mid \beta)\nu^{\theta_k}(d\beta) - \int_{\Delta}\phi(f^{\theta}(y,\beta))p^{\theta}(y \mid \beta)\psi(d\beta)\right|$$
$$\le \left|\int_{\Delta}\phi(f^{\theta_k}(y,\beta))p^{\theta_k}(y \mid \beta)\nu^{\theta_k}(d\beta) - \int_{\Delta}\phi(f^{\theta}(y,\beta))p^{\theta}(y \mid \beta)\nu^{\theta_k}(d\beta)\right| \qquad (79)$$
$$+ \left|\int_{\Delta}\phi(f^{\theta}(y,\beta))p^{\theta}(y \mid \beta)\nu^{\theta_k}(d\beta) - \int_{\Delta}\phi(f^{\theta}(y,\beta))p^{\theta}(y \mid \beta)\psi(d\beta)\right|. \qquad (80)$$

But for any fixed y and β, $\theta_k \to \theta$ implies that $f^{\theta_k}(y,\beta) \to f^{\theta}(y,\beta)$, since from (13), the numerator and denominator of f are linear functions of θ, and the denominator is nonzero. Similarly, $\theta_k \to \theta$ implies that for fixed y and β, $p^{\theta_k}(y \mid \beta) \to p^{\theta}(y \mid \beta)$, since $p^{\theta}(y \mid \beta)$ is linear in θ. Since ϕ is continuous, this implies that for fixed y and β

$$\phi(f^{\theta_k}(y,\beta))p^{\theta_k}(y \mid \beta) \to \phi(f^{\theta}(y,\beta))p^{\theta}(y \mid \beta).$$

Thus for any ϵ we can find k sufficiently large such that

$$\int_{\Delta}(\phi(f^{\theta_k}(y,\beta))p^{\theta_k}(y \mid \beta) - \phi(f^{\theta}(y,\beta))p^{\theta}(y \mid \beta))\nu^{\theta_k}(d\beta) \le \epsilon\int_{\Delta}\nu^{\theta_k}(d\beta) = \epsilon. \qquad (81)$$

So (79) converges to zero. Finally, for fixed y and θ, $f^{\theta}(y,\beta)$ and $p^{\theta}(y \mid \beta)$ are linear in β, so $\phi(f^{\theta}(y,\beta))p^{\theta}(y \mid \beta)$ is a bounded continuous function of β. Thus (80) converges to zero by the weak convergence of ν^{θ_k} to ψ [12, Theorem 25.8]. □

Since the $\{\mu^{\theta_m}\}$ sequence is also tight, the proof that $\mu^{\theta_m} \to \mu^{\theta}$ follows if the limit of any convergent subsequence of $\{\mu^{\theta_m}\}$ is the invariant distribution for π under (12). This is shown with essentially the same argument as above for $\nu^{\theta_k} \to \nu^{\theta}$, using (12) instead of (14) for $p(\alpha \mid \beta)$, $p^{\theta}(y \mid x, \beta)$ instead of $p^{\theta}(y \mid \beta)$, and summations over $\mathcal{X} \times \mathcal{Y}$ instead of $\mathcal{Y}$. The details are omitted.

Lemma A2.6: In general, the Markov property does not hold for π_n under Markov inputs.

Proof: We show this using a counterexample. Let $\mathcal{C} = \{c_1, c_2, c_3\}$ be the state space for S_n, with transition probabilities

$$P = \begin{pmatrix} 2/3 & 0 & 1/3 \\ 0 & 2/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \qquad (82)$$

and initial distribution $\pi_0 = (1/3, 1/3, 1/3)$. This Markov chain is irreducible, aperiodic, and stationary. Each of the states correspond to a memoryless channel, where the input alphabet is $\{0,1\}$ and the output alphabet is $\{0,1,2\}$. The memoryless channels c_1, c_2, and c_3 are defined as follows:

$$\begin{array}{lll} c1: & p_1(0 \mid 0) = p_1(2 \mid 1) = 1, & \text{otherwise } p_1(y \mid x) = 0. \\ c2: & p_2(1 \mid 0) = p_2(2 \mid 1) = 1, & \text{otherwise } p_2(y \mid x) = 0. \\ c3: & p_3(2 \mid 0) = 1, & \\ & p_3(0 \mid 1) = p_3(1 \mid 1) = 1/2, & \text{otherwise } p_3(y \mid x) = 0. \end{array}$$

The stochastic process $\{\pi_n\}_{n=0}^{\infty}$ then takes values on the three points $\alpha_0 = (1/3, 1/3, 1/3)$, $\alpha_1 = (2/3, 0, 1/3)$, and $\alpha_2 = (0, 2/3, 1/3)$.

Let the Markov input distribution be given by $p(x_0 = 0) = p(x_0 = 1) = 1/2$ and $p(x_n = x_{n-1}) = 1$ for $n > 0$. Then

$$p(\pi_3 = \alpha_0 \mid \pi_2 = \alpha_0, \pi_1 = \alpha_1) = 1/3$$

while

$$p(\pi_3 = \alpha_0 \mid \pi_2 = \alpha_0) = 5/6.$$

So $\{\pi_n\}_{n=0}^{\infty}$ is not a Markov process.

Lemma A2.7: In general, the Markov property does not hold for ρ_n under Markov inputs.

Proof: We prove this using a counterexample similar to that of Lemma A2.6. Let the FSMC be as in Lemma A2.6 with the following change in the definition of the memoryless channels c_1, c_2, and c_3:

$$\begin{array}{lll} c1: & p_1(1 \mid 0) = p_1(1 \mid 1) = 1, & \text{otherwise } p_1(y \mid x) = 0. \\ c2: & p_2(2 \mid 0) = p_2(2 \mid 1) = 1, & \text{otherwise } p_2(y \mid x) = 0. \\ c3: & p_3(0 \mid 0) = p_3(2 \mid 0) = 1/2, & \\ & p_3(0 \mid 1) = 1/4, & \\ & p_3(2 \mid 1) = 3/4, & \text{otherwise } p_3(y \mid x) = 0. \end{array}$$

It is easily shown that the state space for the stochastic process $\{\rho_n\}_{n=0}^{\infty}$ includes the points α_0 and α_1 defined in Lemma A2.6. Using the same Markov input distribution defined there, we have

$$p(\rho_3 = \alpha_0 \mid \rho_2 = \alpha_0, \rho_1 = \alpha_1) = 5/36$$

while

$$p(\rho_3 = \alpha_0 \mid \rho_2 = \alpha_0) = 8/57.$$

So $\{\rho_n\}_{n=1}^{\infty}$ is not a Markov process.

APPENDIX III

In this Appendix, we prove Lemma 4.1. Consider first $H(Y_n \mid X_n, X^{n-1}, Y^{n-1})$. We have

$$\begin{aligned} H(Y_n \mid X_n, X^{n-1}, Y^{n-1}) &= \boldsymbol{E}[-\log p(y_n \mid x_n, x^{n-1}, y^{n-1})] \\ &= \boldsymbol{E}\Bigg[-\log \sum_{k=1}^{K} p(y_n \mid x_n, S_n = c_k) \\ &\quad \cdot p(S_n = c_k \mid x^{n-1}, y^{n-1})\Bigg] \\ &= \boldsymbol{E}\left[-\log \sum_{k=1}^{K} p_k(y_n \mid x_n)\pi_n(k)\right] \\ &= \boldsymbol{E}[-\log p(y_n \mid x_n, \pi_n)] \\ &= H(Y_n \mid X_n, \pi_n). \end{aligned} \tag{83}$$

The argument that $H(Y_n \mid Y^{n-1})) = H(Y_n \mid \rho_n)$ is the same, with all the x terms removed and π_n replaced by ρ_n. □

APPENDIX IV

In this Appendix, we prove Lemmas 4.2–4.6.

Proof of Lemma 4.2: We first note that the conditional entropy $H(W \mid V) = \boldsymbol{E} \log p(w \mid v)$, where the log function is concave on $[0, 1]$. To show the first inequality in (25), let f denote any concave function. Then

$$\begin{aligned} &\boldsymbol{E}f(p[y_n \mid x_n, x^{n-1}, y^{n-1}]) \\ &\quad \overset{a}{=} \boldsymbol{E}f(p[y_{n+1} \mid x_{n+1}, x_2^n, y_2^n]) \\ &\quad \overset{b}{=} \boldsymbol{E}f(\boldsymbol{E}(p[y_{n+1} \mid x_{n+1}, x^n, y^n] \mid x_{n+1}, x_2^n, y_2^n)) \\ &\quad \overset{c}{\geq} \boldsymbol{E}\boldsymbol{E}(f(p[y_{n+1} \mid x_{n+1}, x^n, y^n]) \mid x_{n+1}, x_2^n, y_2^n) \\ &\quad \overset{d}{=} \boldsymbol{E}f(p[y_{n+1} \mid x_{n+1}, x^n, y^n]) \end{aligned} \tag{84}$$

where a follows from the stationarity of the channel and the inputs, b and d follow from properties of conditional expectation [12], and c is a consequence of Jensen's inequality.

The second inequality in (25) results from the fact that conditioning on an additional random variable, in this case the initial state S_0, always reduces the entropy [14]. The proof of the third inequality in (25) is similar to that of the first

$$\begin{aligned} &\boldsymbol{E}f(p[y_{n+1} \mid x_{n+1}, x^n, y^n, S_0]) \\ &\quad \overset{a}{=} \boldsymbol{E}f(\boldsymbol{E}p[y_{n+1} \mid x_{n+1}, x^n, y^n, S_0^1] \mid x_{n+1}, x^n, y^n, S_0)) \\ &\quad \overset{b}{=} \boldsymbol{E}f(\boldsymbol{E}(p[y_{n+1} \mid x_{n+1}, x_2^n, y_2^n, S_1] \mid x_{n+1}, x^n, y^n, S_0)) \\ &\quad \overset{c}{\geq} \boldsymbol{E}\boldsymbol{E}(f(p[y_{n+1} \mid x_{n+1}, x_2^n, y_2^n, S_1]) \mid x_{n+1}, x^n, y^n, S_0) \\ &\quad \overset{d}{=} \boldsymbol{E}f(p[y_{n+1} \mid x_{n+1}, x_2^n, y_2^n, S_1]) \\ &\quad \overset{e}{=} \boldsymbol{E}f(p[y_n \mid x_n, x^{n-1}, y^{n-1}, S_0]) \end{aligned} \tag{85}$$

where a and d follow from properties of conditional expectation, b follows from (4) and (5), c follows from Jensen's inequality, and e follows from the channel and input stationarity. □

Proof of Lemma 4.3: From Lemma 4.1

$$\begin{aligned} &\lim_{n\to\infty} H(Y_n \mid X_n, X^{n-1}, Y^{n-1}) \\ &\quad = \lim_{n\to\infty} \boldsymbol{E}\left[-\log \sum_{k=1}^{K} p(y \mid x, S = c_k)\pi_n(k)\right]. \end{aligned} \tag{86}$$

Similarly

$$\begin{aligned} &\lim_{n\to\infty} H(Y_n \mid X_n, X^{n-1}, Y^{n-1}, S_0) \\ &\quad = \lim_{n\to\infty} \boldsymbol{E}\left[-\log \sum_{k=1}^{K} p(y \mid x, S = c_k)\pi_n^*(k)\right] \end{aligned} \tag{87}$$

where $\pi_n^* \triangleq \pi_n^i$ for some i. Applying (15) to (86) and (87) completes the proof. □

Proof of Lemma 4.4: The proof of this lemma is similar to that of Lemma 4.2 above. For the first inequality in (27), we have

$$\begin{aligned} \boldsymbol{E}f(p[y_n \mid y^{n-1}]) &\overset{a}{=} \boldsymbol{E}f(p[y_{n+1} \mid y_2^n]) \\ &\overset{b}{=} \boldsymbol{E}f(\boldsymbol{E}(p[y_{n+1} \mid y^n] \mid y_2^n)) \\ &\overset{c}{\geq} \boldsymbol{E}\boldsymbol{E}(f(p[y_{n+1} \mid y^n]) \mid y_2^n) \\ &\overset{d}{=} \boldsymbol{E}f(p[y_{n+1} \mid y^n]) \end{aligned} \tag{88}$$

where a follows from the stationarity of the inputs and channel, b and d follow from properties of conditional expectation [12], and c is a consequence of Jensen's inequality.

The second inequality results from the fact that conditioning on an additional random variable reduces entropy. Finally, for the third inequality, we have

$$\begin{aligned} \boldsymbol{E}f(p[y_{n+1} \mid y^n, S_0]) &\overset{a}{=} \boldsymbol{E}f(\boldsymbol{E}(p[y_{n+1} \mid y^n, S_1] \mid y^n, S_0)) \\ &\overset{b}{=} \boldsymbol{E}f(\boldsymbol{E}(p[y_{n+1} \mid y_2^n, S_1] \mid y^n, S_0)) \\ &\overset{c}{\geq} \boldsymbol{E}\boldsymbol{E}(f(p[y_{n+1} \mid y_2^n, S_1]) \mid y^n, S_0) \\ &\overset{d}{=} \boldsymbol{E}f(p[y_{n+1} \mid y_2^n, S_1]) \\ &\overset{e}{=} \boldsymbol{E}f(p[y_n \mid y^{n-1}, S_0]) \end{aligned} \tag{89}$$

where a and d follow from properties of conditional expectation, b follows from (6), c follows from Jensen's inequality, and e follows from the channel and input stationarity. □

Proof of Lemma 4.5: Following a similar argument as in the proof of Lemma 4.3, we have that

$$\begin{aligned}\lim_{n\to\infty} H(Y_n \mid Y^{n-1}) &= \lim_{n\to\infty} \boldsymbol{E}\left[-\log \sum_{k=1}^{K} p(y \mid S = c_k)\rho_n(k)\right] \qquad (90)\end{aligned}$$

and

$$\begin{aligned}\lim_{n\to\infty} H(Y_n \mid Y^{n-1}, S_0) &= \lim_{n\to\infty} \boldsymbol{E}\left[-\log \sum_{k=1}^{K} p(y \mid S = c_k)\rho_n^*(k)\right] \qquad (91)\end{aligned}$$

where $\rho_n^* \triangleq \rho_n^i$ for some i. Applying (16) to (90) and (91) completes the proof. □

Proof of Lemma 4.6: We first consider the limiting conditional entropy $H(Y_n \mid \rho_n^\theta)$ as $n \to \infty$. Let ν_n^θ denote the distribution of ρ_n^θ and ν^θ denote the corresponding limit distribution. Also, let $p_\theta(y \mid \cdot)$ explicitly denote the dependence of the (conditional) output probability on θ. Then

$$\begin{aligned}
&\lim_{n\to\infty} H(Y_n \mid \rho_n^\theta)\\
&= \lim_{n\to\infty} \boldsymbol{E}[-\log p_\theta(y_n \mid \rho_n^\theta)]\\
&= \lim_{n\to\infty} \sum_{y^n \in \mathcal{Y}^n} -\log p_\theta(y_n \mid \rho_n^\theta(y^{n-1}))p_\theta(y^n)\\
&= \lim_{n\to\infty} \sum_{y^n \in \mathcal{Y}^n} -\log p_\theta(y_n \mid \rho_n^\theta(y^{n-1}))\\
&\quad \cdot p_\theta(y_n \mid y^{n-1})p_\theta(y^{n-1})\\
&= \lim_{n\to\infty} \sum_{y^{n-1}\in\mathcal{Y}^{n-1}} \left[\sum_{y_n\in\mathcal{Y}} -\log p_\theta(y_n \mid \rho_n^\theta(y^{n-1}))\right.\\
&\qquad\qquad \left. \cdot p_\theta(y_n \mid \rho_n^\theta(y^{n-1}))\right] p_\theta(y^{n-1})\\
&= \lim_{n\to\infty} \int_\Delta \left[\sum_{y_n\in\mathcal{Y}} -\log p_\theta(y_n \mid \rho_n^\theta)p_\theta(y_n \mid \rho_n^\theta)\right] p(d\rho_n^\theta)\\
&= \lim_{n\to\infty} \int_\Delta \left[\sum_{y\in\mathcal{Y}} -\log p_\theta(y \mid \rho)p_\theta(y \mid \rho)\right] \nu_n^\theta(d\rho)\\
&= \int_\Delta \left[\sum_{y\in\mathcal{Y}} -\log p_\theta(y \mid \rho)p_\theta(y \mid \rho)\right] \nu^\theta(d\rho). \qquad (92)
\end{aligned}$$

The second and fourth equalities in (92) follow from the fact that ρ_n is a function of y^{n-1}. We also use this in the fifth equality to take expectations relative to ρ_n instead of y^{n-1}. The sixth equality follows from the definition of ν_n and the stationarity of the channel inputs. The last equality follows from the weak convergence of ρ_n^θ and the fact that the entropy is continuous in ρ and is bounded by log $\mid \mathcal{Y} \mid$ [12, Theorem 25.8].

The limiting conditional entropy $H(Y_n \mid X_n, \pi_n)$ is obtained with a similar argument. Let μ_n^θ denote the distribution of π_n^θ and μ^θ denote the corresponding limit distribution. Then

$$\begin{aligned}
&\lim_{n\to\infty} H(Y_n \mid X_n, \pi_n^\theta)\\
&= \lim_{n\to\infty} \boldsymbol{E}[-\log p_\theta(y_n \mid x_n, \pi_n^\theta)]\\
&= \lim_{n\to\infty} \sum_{\substack{y^n\in\mathcal{Y}^n\\ x^n\in\mathcal{X}^n}} -\log p_\theta(y_n \mid x_n, \pi_n^\theta)p_\theta(y^n, x^n)\\
&= \lim_{n\to\infty} \sum_{\substack{y^n\in\mathcal{Y}^n\\ x^n\in\mathcal{X}^n}} -\log p_\theta(y_n \mid x_n, \pi_n^\theta)\\
&\quad \cdot p_\theta(y_n, x_n \mid y^{n-1}, x^{n-1})p_\theta(y^{n-1}, x^{n-1})\\
&= \lim_{n\to\infty} \sum_{\substack{y^{n-1}\in\mathcal{Y}^{n-1}\\ x^{n-1}\in\mathcal{X}^{n-1}}} \left[\sum_{\substack{y_n\in\mathcal{Y}\\ x_n\in\mathcal{X}}} -\log p_\theta(y_n \mid x_n, \pi_n^\theta)\right.\\
&\qquad\qquad \left. \cdot p_\theta(y_n \mid x_n, y^{n-1}, x^{n-1})\theta(x_n)\right]\\
&\quad \cdot p_\theta(y^{n-1}, x^{n-1})\\
&= \lim_{n\to\infty} \sum_{\substack{y^{n-1}\in\mathcal{Y}^{n-1}\\ x^{n-1}\in\mathcal{X}^{n-1}}} \left[\sum_{\substack{y_n\in\mathcal{Y}\\ x_n\in\mathcal{X}}} -\log p_\theta(y_n \mid x_n, \pi_n^\theta)\right.\\
&\qquad\qquad \left. \cdot p_\theta(y_n \mid x_n, \pi_n^\theta)\theta(x_n)\right]\\
&\quad \cdot p_\theta(y^{n-1}, x^{n-1})\\
&= \lim_{n\to\infty} \int_\Delta \left[\sum_{\substack{y_n\in\mathcal{Y}\\ x_n\in\mathcal{X}}} -\log p_\theta(y_n \mid x_n, \pi_n)\right.\\
&\qquad\qquad \left. \cdot p_\theta(y_n \mid x_n, \pi_n^\theta)\theta(x_n)\right] p(d\pi_n^\theta)\\
&= \lim_{n\to\infty} \int_\Delta \left[\sum_{\substack{y\in\mathcal{Y}\\ x\in\mathcal{X}}} -\log p(y \mid x, \pi)p(y \mid x, \pi)\theta(x)\right] \mu_n^\theta(d\pi)\\
&= \lim_{n\to\infty} \int_\Delta \left[\sum_{\substack{y\in\mathcal{Y}\\ x\in\mathcal{X}}} -\log p(y \mid x, \pi)p(y \mid x, \pi)\theta(x)\right] \mu^\theta(d\pi)
\end{aligned}$$
$$(93)$$

where we use the fact that π_n is a function of x^{n-1} and y^{n-1}, and the last equality follows from the weak convergence of π_n^θ to π^θ. □

APPENDIX V

In this Appendix, we prove Lemmas 5.1 and 5.2.

Proof of Lemma 5.1: From [14],

$$H(Y_n \mid \rho_n) \leq H(Y_n) \leq \log \mid \mathcal{Y} \mid$$

and similarly

$$H(Y_n \mid \rho_n^i) \leq H(Y_n) \leq \log \mid \mathcal{Y} \mid$$

for any i. But since each $c_k \in \mathcal{C}$ is output symmetric, for each k the columns of $M^k \triangleq \{M_{lj}^k = p_k(j \mid l), l \in \mathcal{X}, j \in \mathcal{Y}\}$ are permutations of each other. Thus, if the marginal $p(x_n)$ is uniform, then $p(y_n \mid S_n = c_k)$ is also uniform, i.e., $p(y_n \mid S_n = c_k) = 1/ \mid \mathcal{Y} \mid$. Hence for any $\rho_n \in \Delta$

$$\begin{aligned} p(y_n \mid \rho_n) &= \sum_{k=1}^{K} p(y_n \mid S_n = c_k)\rho_n(k) \\ &= \frac{1}{\mid \mathcal{Y} \mid} \sum_{k=1}^{K} \rho_n(k) = \frac{1}{\mid \mathcal{Y} \mid} \end{aligned} \tag{94}$$

and similarly $p(y_n \mid \rho_n^i) = 1/ \mid \mathcal{Y} \mid$ for any i. Thus

$$\begin{aligned} H(Y_n \mid \rho_n) &= \int_{\rho_n \in \Delta} \sum_{y_n \in \mathcal{Y}} p(\rho_n) p(y_n \mid \rho_n)[-\log p(y_n \mid \rho_n)] \\ &= \int_{\rho_n \in \Delta} p(\rho_n) \sum_{y_n \in \mathcal{Y}} p(y_n \mid \rho_n)[-\log p(y_n \mid \rho_n)] \\ &= \int_{\rho_n \in \Delta} p(\rho_n) \sum_{y_n \in \mathcal{Y}} \frac{1}{\mid \mathcal{Y} \mid} \log \mid \mathcal{Y} \mid \\ &= \log \mid \mathcal{Y} \mid \end{aligned} \tag{95}$$

and similarly

$$H(Y_n \mid \rho_n^i) = \log \mid \mathcal{Y} \mid$$

for any i. Since (95) only requires that $p(x_n)$ is uniform for each n, an i.i.d. uniform input distribution achieves this maximum. Substituting π for ρ in the above argument yields the result for $H(Y_n \mid \pi_n)$ and $H(Y_n \mid \pi_n^i)$. □

Proof of Lemma 5.2: We consider only $H(Y_n \mid X_n, \pi_n)$, since the same argument applies for $H(Y_n \mid X_n, \pi_n^i)$. By the output symmetry of each $c_k \in \mathcal{C}$, the sets

$$\{p_k(y \mid x) \colon y \in \mathcal{Y}\}_{x \in \mathcal{X}}$$

are permutations of each other. Thus

$$\begin{aligned} H(Y_n \mid X_n, \pi_n) = &\sum_{\pi_n} \sum_{x_n} \left[\sum_{y_n} \left(-\log \sum_k p_k(y_n \mid x_n) \pi_n(k) \right) \right. \\ &\left. \cdot \sum_k p_k(y_n \mid x_n) \pi_n(k) \right] p(x_n) p(\pi_n) \\ = &\sum_{\pi_n} \sum_{y_n} \left(-\log \sum_k p_k(y_n \mid x_n) \pi_n(k) \right) \\ &\cdot \left(\sum_k p_k(y_n \mid x) \pi_n(k) \right) p(\pi_n) \quad \forall x \in \mathcal{X}. \end{aligned} \tag{96}$$

So $H(Y_n \mid X_n, \pi_n)$ depends only on the distribution of π_n. But by (38), this distribution depends only on the distribution of Z^{n-1}. The proof then follows from the fact that $p(Z^n \mid X^n) = p(Z^n)$. □

APPENDIX VI

We consider a Q-AWN channel where the output is quantized to the nearest input symbol and the input alphabet consists of symmetric PSK symbols. We want to show that for any k $P_{ij}^k \triangleq p_k(y = j \mid x = i)$ has rows which are permutations of each other and columns which are permutations of each other. The input/output symbols are given by

$$y_m = x_m = Ae^{j2\pi m/M}, \quad m = 1, \cdots, M. \tag{97}$$

Define the $M \times M$ matrix Z by $Z_{ij} = |y_i - x_j|$ and let $q_k(Z_{ij})$ denote the distribution of the quantized noise, which is determined by the noise density n_k and the values of A and M from (97). By symmetry of the input/output symbols and the noise, the rows of Z are permutations of each other, and the columns are also permutations of each other.

If M is odd, then

$$p_k(y \mid x) = \begin{cases} q_k(|y - x|), & |y - x| = 0 \\ q_k(|y - x|)/2, & \text{else} \end{cases} \tag{98}$$

and if M is even

$$p_k(y \mid x) = \begin{cases} q_k(|y - x|), & |y - x| = 0 \text{ or } |y - x| = 2A \\ q_k(|y - x|)/2, & \text{else.} \end{cases} \tag{99}$$

Thus P_{ij}^k depends only on the value of Z_{ij}; the rows of P_{ij}^k are therefore permutations of each other, and so are the columns.

APPENDIX VII

We will show that the π-output channel is asymptotically memoryless as $J \to \infty$. Indeed, since the FSMC is indecomposable and stationary

$$\lim_{J \to \infty} p(S_{n+J}, S_n) = \lim_{J \to \infty} p(S_{n+J}) p(S_n) \tag{100}$$

for any n, and thus also

$$\lim_{J \to \infty} p(\pi_{n+J}, \pi_n) = \lim_{J \to \infty} p(\pi_{n+J}) p(\pi_n). \tag{101}$$

Therefore, since π_{jl} and $\pi_{j(l-1)}$ are J iterations apart, π_{jl} and $\pi_{j(l-1)}$ are asymptotically independent as $J \to \infty$.

In order to show that the π-output channel is memoryless, we must show that for any j and L

$$p(y^{jL}, \pi^{jL} \mid x^{jL}) = \prod_{l=1}^{L} p(y_{jl}, \pi_{jl} \mid x_{jl}). \tag{102}$$

We can decompose $p(y^{jL}, \pi^{jL} \mid x^{jL})$ as follows:

$$\begin{aligned} &p(y^{jL}, \pi^{jL} \mid x^{jL}) \\ &\quad = \prod_{l=1}^{L} p(y_{jl}, \pi_{jl} \mid x_{jl}, y^{j(l-1)}, \pi^{j(l-1)} x^{j(l-1)}). \end{aligned} \tag{103}$$

Thus we need only show that the lth factor in the right-hand side of (103) equals $p(y_{jl}, \pi_{jl} \mid x_{jl})$ in the limit as $J \to \infty$. This result is proved in the following lemma.

Lemma A7.1: For asymptotically large J

$$p\big(y_{jl}, \pi_{jl} \mid x_{jl}, y^{j(l-1)}, \pi^{j(l-1)}, x^{j(l-1)}\big) = p(y_{jl}, \pi_{jl} \mid x_{jl}). \tag{104}$$

Proof:

$$\begin{aligned} p\big(y_{jl}, \pi_{jl} &\mid x_{jl}, y^{j(l-1)}, \pi^{j(l-1)}, x^{j(l-1)}\big) \\ &= p\big(y_{jl} \mid \pi_{jl}, x_{jl}, y^{j(l-1)}, \pi^{j(l-1)}, x^{j(l-1)}\big) \\ &\quad \cdot p\big(\pi_{jl} \mid x_{jl}, y^{j(l-1)}, \pi^{j(l-1)}, x^{j(l-1)}\big) \\ &= p\big(y_{jl} \mid \pi_{jl}, x_{jl}\big) p\big(\pi_{jl} \mid y^{j(l-1)}, \pi^{j(l-1)}, x^{j(l-1)}\big) \\ &= p\big(y_{jl} \mid \pi_{jl}, x_{jl}\big) p\big(\pi_{jl} \mid \pi_{(j+1)(l-1)}\big) \\ &= p\big(y_{jl} \mid \pi_{jl}, x_{jl}\big) p\big(\pi_{jl}\big) \\ &= p\big(y_{jl}, \pi_{jl} \mid x_{jl}\big) \end{aligned} \tag{105}$$

where the second equality follows from (4) and (5), the third equality follows from (4) and (11), and the fourth equality follows from (101) in the asymptotic limit of deep interleaving. □

Appendix VIII

The π-output channels are independent if

$$p(y^J, \pi^J \mid x^J) = \prod_{j=1}^{J} p(y_j, \pi_j \mid x_j). \tag{106}$$

This is shown in the following string of equalities:

$$\begin{aligned} p(y^J&, \pi^J \mid x^J) \\ &= \prod_{j=1}^{J} p(y_j, \pi_j \mid x_j, y^{j-1}, \pi^{j-1}, x^{j-1}) \\ &= \prod_{j=1}^{J} p(y_j \mid \pi_j, x_j, y^{j-1}, \pi^{j-1}, x^{j-1}) \\ &\quad \cdot p(\pi_j \mid x_j, y^{j-1}, \pi^{j-1}, x^{j-1}) \\ &= \prod_{j=1}^{J} p(y_j \mid \pi_j, x_j) p(\pi_j \mid x_j, y^{j-1}, \pi^{j-1}, x^{j-1}) \\ &= \prod_{j=1}^{J} p(y_j \mid \pi_j, x_j) p(\pi_j) \end{aligned} \tag{107}$$

where the third equality follows from (5) and the last equality follows from the fact that we ignore error propagation, so x^{j-1}, y^{j-1}, and π^{j-1} are all known constants at time j.

We now determine the average mutual information of the parallel π-output channels for a fixed input distribution $p(X^J)$. The average mutual information of the parallel set is

$$I_J = \frac{1}{J} I(Y^J, \pi^J; X^J). \tag{108}$$

From above, the parallel channels are independent, and each channel is memoryless with asymptotically deep interleaving. Thus we obtain (45) as follows:

$$\begin{aligned} \frac{1}{J}&I(Y^J, \pi^J; X^J) \\ &= H(Y^J, \pi^J) - H(Y^J, \pi^J \mid X^J) \\ &= H(Y^J \mid \pi^J) + H(\pi^J) \\ &\quad - (H(Y^J \mid \pi^J, X^J) + H(\pi^J \mid X^J)) \\ &= H(Y^J \mid \pi^J) - H(Y^J \mid \pi^J, X^J) \\ &= \sum_{j=1}^{J} H(Y_j \mid Y^{j-1}, \pi^J) - H(Y_j \mid Y^{j-1}, \pi^J, X^J) \\ &= \sum_{j=1}^{J} H(Y_j \mid \pi_j) - H(Y_j \mid \pi_j, X_j) \end{aligned} \tag{109}$$

where the third equality follows from the fact that

$$p(\pi^J \mid x^J) = p(\pi^J \mid x^{J-1}) = p(\pi^J)$$

by definition of π^J and by the memoryless property of the π_j channels. The last inequality follows from the fact that

$$H(Y_j \mid Y^{j-1}, \pi^J) = H(Y_j \mid \rho_j, \pi^J) = H(Y_j \mid \pi_J) \tag{110}$$

since the π_j channels are memoryless and $\rho_j = \boldsymbol{E}_{x^{j-1}} \pi_j$.

Appendix IX

In this Appendix we examine the cutoff rate for uniformly symmetric variable-noise channels. The first three lemmas show that for these channels, the maximizing distribution of (52) is uniform and i.i.d. We then determine that R_j, as given by (52), is monotonically increasing in j, and use this to get a simplified formula for R_{df} in terms of the limiting value of R_j.

Lemma A9.1: For all j, R_j depends only on $p(x_j)$.

Proof: From the proof of Lemma 5.2, π_j is a function of Z^{j-1}, and is independent of X^{j-1}. So $p(\pi_j)$ does not depend on the input distribution. The result then follows from the definition of R_j. □

Corollary: An independent input distribution achieves the maximum of R_{df}.

Lemma A9.2: For a fixed input distribution $p(X^J)$, the J corresponding π-output channels are all symmetric [13, p. 94].

Proof: We must show that for any $j < J$, the set of outputs for the jth π-output channel can be partitioned into subsets such that the corresponding submatrices of transition probabilities have rows which are permutations of each other and columns which are permutations of each other. We will call such a matrix *row/column-permutable*.

Let $n_j \leq |\mathcal{X}|^j |\mathcal{Y}|^j$ be the number of points $\delta \in \Delta$ with $p(\pi_j = \delta) > 0$, and let $\{\delta_i\}_{i=1}^{n_j}$ explicitly denote this set. Then we can partition the output into n_j sets, where the ith set consists of the pairs $\{(y, \delta_i) \colon y \in \mathcal{Y}\}$. We want to show that the transition probability matrix associated with each of these output partitions is row/column-permutable, i.e., that for all i, $1 \leq i \leq n_j$, the $|\mathcal{X}| \times |\mathcal{Y}|$ matrix

$$P^i \triangleq p(y_j = y, \pi_j = \delta_i \mid x_j = x), \quad x \in \mathcal{X}, y \in \mathcal{Y} \tag{111}$$

has rows which are permutations of each other, and columns which are permutations of each other.

Since the FSMC is a variable-noise channel, there is a function f such that $p_k(y \mid x)$ depends only on $z \triangleq f(x,y)$ for all k, $1 \leq k \leq K$. Therefore, if for some k', $p_{k'}(y \mid x) = p_{k'}(y' \mid x')$, then $f(x,y) = f(x',y')$. But since $z = f(x,y)$ is the same for all k, this implies that

$$p_k(y \mid x) = p_k(y' \mid x') \quad \forall k, 1 \leq k \leq K. \tag{112}$$

Fix k'. Then by definition of uniform symmetry, $p_{k'}(y \mid x)$ is row/column-permutable. Using (112), we get that the $|\mathcal{X}| \times |\mathcal{Y}|$ matrix

$$P_\Sigma = \sum_{k=1}^{K} p_k(y \mid x), \quad x \in \mathcal{X}, y \in \mathcal{Y} \tag{113}$$

is also row/column-permutable. Moreover, multiplying a matrix by any constant will not change the permutability of its rows and columns, hence the matrix

$$P_\Sigma^i = \left[\sum_{k=1}^{K} p_k(y \mid x)\right] \delta_i p(\pi_j = \delta_i), \quad x \in \mathcal{X}, y \in \mathcal{Y} \tag{114}$$

is also row/column-permutable. But this completes the proof, since

$$\begin{aligned} p(y_j = y, \pi_j = \delta_i \mid x_j = x) \\ = \sum_{k=1}^{K} p_k(y_j = y \mid x_j = x) \delta_i p(\pi_j = \delta_i). \end{aligned} \tag{115}$$

□

Lemma A9.3: For i.i.d. uniform inputs, R_j is monotonically increasing in j.

Proof: For uniform i.i.d. inputs

$$\begin{aligned} R_j = -\log \Bigg(\frac{1}{|\mathcal{X}|^2} \int_{\pi_j \in \Delta} p(\pi_j) \\ \cdot \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y \mid x, S = c_k) \pi_j(k)} \right]^2 \Bigg). \end{aligned} \tag{116}$$

Let

$$f(\pi_j) \triangleq \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y \mid x, S = c_k) \pi_j(k)} \right]^2. \tag{117}$$

Then

$$R_j = -\log \left(\frac{1}{|\mathcal{X}|^2} \boldsymbol{E}[f(\pi_j)] \right).$$

We want to show that

$$-\log \left(\frac{1}{|\mathcal{X}|^2} \boldsymbol{E}[f(\pi_j)] \right) \leq -\log \left(\frac{1}{|\mathcal{X}|^2} \boldsymbol{E}[f(\pi_{j+1})] \right)$$

or, equivalently, that

$$\boldsymbol{E}[f(\pi_j)] \geq \boldsymbol{E}[f(\pi_{j+1})].$$

Following an argument similar to that of Lemma 4.2, we have

$$\begin{aligned} &\boldsymbol{E} f(\pi_j) \\ &= \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y \mid x) \pi_j(k)} \right]^2 \\ &= \boldsymbol{E} \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y \mid x) p(S_j = c_k \mid x^{n-1}, y^{n-1})} \right]^2 \\ &\overset{a}{=} \boldsymbol{E} \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y \mid x) p\big(S_{j+1} = c_k \mid x_2^n, y_2^n\big)} \right]^2 \\ &= \boldsymbol{E} \sum_{y \in \mathcal{Y}} \\ &\quad \cdot \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y \mid x) \boldsymbol{E}\big[p\big(S_{j+1} = c_k \mid x^n, y^n \mid x_2^n, y_2^n\big)\big]} \right]^2 \\ &\overset{b}{\geq} \boldsymbol{E} \sum_{y \in \mathcal{Y}} \\ &\quad \cdot \left[\sum_{x \in \mathcal{X}} \boldsymbol{E} \left(\sqrt{\sum_{k=1}^{K} p_k(y \mid x) p(S_{j+1} = c_k \mid x^n, y^n)} \,\middle|\, x_2^n, y_2^n \right) \right]^2 \\ &= \boldsymbol{E} \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} p_k(y \mid x) p(S_{j+1} = c_k \mid x^n, y^n)} \right]^2 \\ &= \boldsymbol{E} f(\pi_{j+1}) \end{aligned} \tag{118}$$

where a follows from stationarity and b follows from Jensen's inequality. □

Lemma A9.4: For uniformly symmetric variable-noise channels, a uniform i.i.d. input distribution maximizes R_{df}. Moreover

$$R_{\text{df}} = \lim_{j \to \infty} R_j. \tag{119}$$

Proof: From Lemma A9.2, the maximizing distribution for R_{df} is independent. Moreover, from Lemma A9.2, each of the π-output channels are symmetric, therefore from [13, p. 144], a uniform distribution for $p(X_j)$ maximizes R_j for all j, and therefore it maximizes R_{df}. By Lemma A9.3, R_j is monotonically increasing in j for i.i.d. uniform inputs. Finally, by Lemma A2.2, for $f(\pi_j)$ as defined in (117), $\boldsymbol{E}f(\pi_j)$ converges to a limit which is independent of the initial channel state, and thus so does $R_j = -\log(\frac{1}{|\mathcal{X}|^2} \boldsymbol{E} f(\pi_j))$. Therefore

$$R_{\text{df}} = \lim_{J \to \infty} \frac{1}{J} \sum_{j=1}^{J} R_j = \lim_{j \to \infty} R_j. \tag{120}$$

Acknowledgment

The authors wish to thank V. Borkar for suggesting the proof of Lemma A2.3. They also wish to thank the reviewers for their helpful comments and suggestions, and for providing the counterexample of Lemma A2.6.

REFERENCES

[1] M. Mushkin and I. Bar-David, "Capacity and coding for the Gilbert-Elliot channels," *IEEE Trans. Inform. Theory*, vol. 35, no. 6, pp. 1277–1290, Nov. 1989.

[2] A. J. Goldsmith, "The capacity of time-varying multipath channels," Masters thesis, Dept. of Elec. Eng. Comput. Sci., Univ. of California at Berkeley, May 1991.

[3] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Channels.* New York: Academic Press, 1981.

[4] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding.* New York: McGraw-Hill, 1979.

[5] H. S. Wang and N. Moayeri, "Modeling, capacity, and joint source/channel coding for Rayleigh fading channels," Tech. Rep. WINLAB-TR-32, Wireless Information Network Lab., Rutgers Univ., New Brunswick, NJ, May 1992. Also "Finite-state Markov channel—A useful model for radio communication channels," *IEEE Trans. Vehic. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[6] K. Leeuwin-Boullé and J. C. Belfiore, "The cutoff rate of time correlated fading channels," *IEEE Trans. Inform. Theory*, vol. 39, no. 2, pp. 612–617, Mar. 1993.

[7] N. Seshadri and C.-E. W. Sundberg, "Coded modulations for fading channels—An overview," *European Trans. Telecommun. Related Technol.*, vol. ET-4, no. 3, pp. 309–324, May–June 1993.

[8] L.-F. Wei, "Coded M-DPSK with built-in time diversity for fading channels," *IEEE Trans. Inform. Theory*, vol. 39, no. 6, pp. 1820–1839, Nov. 1993.

[9] D. Divsalar and M. K. Simon, "The design of trellis coded MPSK for fading channels: Set partitioning for optimum code design," *IEEE Trans. Commun.*, vol. 36, no. 9, pp. 1013–1021, Sept. 1988.

[10] W. C. Dam and D. P. Taylor, "An adaptive maximum likelihood receiver for correlated Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 42, no. 9, pp. 2684–2692, Sept. 1994.

[11] J. H. Lodge and M. L. Moher, "Maximum-likelihood sequence estimation of CPM signals transmitted over Rayleigh flat-fading channels," *IEEE Trans. Commun.*, vol. 38, no. 6, pp. 787–794, June 1990.

[12] P. Billingsley, *Probability and Measure.* New York: Wiley, 1986

[13] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

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

[15] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control.* Englewood Cliffs, NJ: Prentice-Hall, 1986.

[16] J. G. Proakis, *Digital Communications*, 2nd ed. New York: McGraw-Hill, 1989.

[17] M. V. Eyuboğlu, "Detection of coded modulation signals on linear, severely distorted channels using decision-feedback noise prediction with interleaving," *IEEE Trans. Commun.*, vol. 36, no. 4, pp. 401–409, Apr. 1988.

[18] J. C. S. Cheung and R. Steele, "Soft-decision feedback equalizer for continuous phase modulated signals in wideband mobile radio channels," *IEEE Trans. Commun.*, vol. 42, no. 2/3/4, pp. 1628–1638, Feb.–Apr. 1994.

[19] A. Duel-Hallen and C. Heegard, "Delayed decision-feedback sequence estimation," *IEEE Trans. Commun.*, vol. 37, no. 5, pp. 428–436, May 1989.

[20] M. V. Eyuboğlu and S. U. H. Qureshi, "Reduced-state sequence estimation with set partitioning and decision feedback," *IEEE Trans. Commun.*, vol. 36, no. 1, pp. 13–20, Jan. 1988.

[21] T. Kaijser, "A limit theorem for partially observed Markov chains," *Ann. Probab.*, vol. 3, no. 4, pp. 677–696, 1975.