On the Design of Finite-State Shaping Encoders for Partial-Response Channels

Joseph B. Soriaga
Qualcomm, Inc.
5775 Morehouse Drive, San Diego, CA 92121
Email: jsoriaga@qualcomm.com

Paul H. Siegel
Center for Magnetic Recording Research
University of California, San Diego
9500 Gilman Drive, La Jolla, CA, USA, 92093-0401
Email: psiegel@ucsd.edu

***Abstract*— Shaping encoders, or encoders which transform an i.i.d. equiprobable process into a near-capacity achieving process, can be concatenated with an outer parity-check code to provide reliable communication at rates near capacity. Here, we examine the problem of designing shaping encoders for binary-input partial-response channels by minimizing the Kullback-Leibler divergence rate between the encoder output process and the target process. Fundamental limits in this case are related to the capacity of channels with cost constraints. Previously introduced encoder constructions are then found to achieve arbitrarily low divergence rates, such as those derived from constraint graphs for typical sequences [1], as well as rate-1 encoders used for near-capacity concatenated coding systems in [2].**

I. Introduction

The technique presented in [3], [4], [5] for accurately estimating achievable rates on binary-input partial-response channels has led to interesting insights for constructing capacity-achieving codes. In particular, the mutual information rate for an independent and identically distributed (i.i.d) equiprobable input process may be appreciably lower than capacity at a moderate-to-low signal-to-noise ratio (SNR), suggesting that turbo codes and low-density parity-check (LDPC) codes alone may not be suitable at lower SNRs. However, through a simple example, Gallager showed [6, p. 208] how capacity could be achieved through the concatenation of an outer linear parity-check code (or coset thereof) with an inner transformer, or *shaping encoder*, that induces the capacity achieving distribution from an i.i.d. binary equiprobable process. Pursuing this idea with the knowledge gained in [3], [4], [5], Ma, Varnica, and Kavčić [7], [8] presented the first complete methodology for near-capacity concatenated coding systems on partial-response channels, including a novel design for an inner finite-state shaping encoder. Soriaga and Siegel [1], [2] later followed with similar concatenated coding systems that used less heuristic shaping code design methods, while Doan and Narayanan [9] provided an alternative and simpler inner code design suitable for low SNRs.

In this paper, we examine the problem of designing shaping encoders from the perspective of minimizing the Kullback-Leibler divergence rate between the encoder-induced process and the target process. Specifically, we consider only target processes which are finite-state Markov. By relating this problem to coding for channels with cost constraints, we are able to characterize the minimum achievable divergence rate. We then revisit two effective shaping encoder construction methods presented earlier in [1], [2], [10], and evaluate their resulting divergence rates and compare them to the minimum achievable. Such example encoders include those derived from constrained graphs for typical sequences [1], as well as rate-1 shaping encoders used in the near-capacity concatenated coding systems of [2]. For completeness, we begin with a brief review of the capacity for channels with cost constraints.

II. Capacity of Channels with Cost Constraints

A channel with cost constraints is a noiseless channel where each symbol x, chosen from some finite alphabet $\mathbb{X}$, is assigned a nonnegative cost $w(x)$, and the entire sequence is constrained to have an average (per-symbol) cost of no more than W. For a *μth-order finite-memory cost function* $w(x_t|\mathbf{x}_{t-\mu}^{t-1})$, the cost of using a symbol x_t at time t is dependent upon the last μ symbols. The average cost for a sequence $\mathbf{x}_1^N$ is then $N^{-1}\sum_{t=1}^{N} w(x_t|\mathbf{x}_{t-\mu}^{t-1})$, where the initial cost for the first μ terms is predefined. Such cost functions can be described with a labeled directed graph G_{cost}, in which each state corresponds to a sequence of μ symbols, and the transitions between states $i = (\mathbf{x}_{t-\mu}^{t-1})$ and $j = (\mathbf{x}_{t-\mu+1}^{t})$ are labeled with the appropriate cost $w(i,j) = w(x_t|\mathbf{x}_{t-\mu}^{t-1})$.

If we define $S_N(W)$ as the number of sequences of length N with average cost less than or equal to W, then the capacity for a given cost constraint is

$$C(W) \stackrel{\text{def}}{=} \limsup_{N\to\infty} \frac{1}{N} \log_2 |S_N(W)| . \tag{1}$$

In [11], Justesen and Høholdt give the maxentropic Markov chain for a given average cost W. Incidentally, as we shown in Appendix I, this Markov chain also achieves the capacity. Both of these results, summarized below, rely on the *one-step cost-enumerator matrix* $\mathbf{A}(s)$, with $A_{i,j}(s) = s^{w(i,j)}$, and $A_{i,j}(s) = 0$ whenever $w(i,j)$ is infinite (i.e., whenever there is no transition allowed from i to j).

Theorem 1: (Justesen and Høholdt [11, Theorem 1]) Consider a channel with cost constraints given by a finite-memory cost function $w(i,j)$ and labeled graph G_{cost}. For $0 \le s < 1$, let $\mathbf{A}(s)$ be the corresponding one-step cost-enumerator matrix, with maximal eigenvalue $\lambda(s)$ and left and right eigenvectors $\mathbf{u}(s)$ and $\mathbf{v}(s)$, respectively, where $\mathbf{u}(s)\mathbf{v}(s)^T =$

1. For an average cost constraint of

$$W(s) = \frac{1}{\lambda(s)} \sum_{i,j} u_i(s) w(i,j) s^{w(i,j)} v_j(s), \tag{2}$$

the maxentropic Markov chain has state-transition probabilities

$$p_{j|i} = s^{w(i,j)} v_j(s) / \lambda(s) v_i(s). \tag{3}$$

and an entropy rate equal to

$$C(W(s)) = \log_2 \lambda(s) - W(s) \log_2 s. \tag{4}$$

Theorem 2: For a channel with cost constraints given by a finite-memory cost function $w(i,j)$ and labeled graph G_{cost}, the capacity of the channel with an average cost constraint $W(s)$ is $C(W(s))$, as given above in equations (2) and (4), respectively.

Proof: See Appendix I. ∎

III. Minimizing the Kullback-Leibler Divergence Rate for Finite-State Markov Processes

Recall that the *Kullback-Leibler (K-L) divergence rate* between two random processes with distributions Q and P is

$$D(Q||P) \stackrel{\text{def}}{=} \limsup_{N\to\infty} \frac{1}{N} \sum_{\mathbf{x}_1^N \in \mathbb{X}^N} q(\mathbf{x}_1^N) \log_2 \frac{q(\mathbf{x}_1^N)}{p(\mathbf{x}_1^N)}.$$

Note that we do not necessarily have $D(Q||P) = D(P||Q)$, and in some cases it has been shown that $D(Q||P) = 0$ implies $Q = P$ [12]. These properties are analogous those for the divergence between two random variables [15, p. 18]. For a given shaping encoder, we might measure $D(Q||P)$ between the induced encoder-output process Q and the target process P. This is a convenient criterion in that it has a closed-form expression when both processes are Markov, or in some cases Markov-driven, and share the same state space. Additionally, the near-capacity input processes we seek to induce are also finite-state Markov [5], [13].

When designing a shaping encoder of a given rate, it is interesting to characterize the minimum K-L divergence rate that can be achieved. This can be derived once we consider the problem from the perspective of coding for channels with cost constraints. That is, consider a μth-order binary Markov process with distribution P, and define a corresponding finite-memory cost function,

$$w(x_t|x_{t-\mu}^{t-1}) = -\log_2 P(X_t = x_t | \mathbf{X}_{t-\mu}^{t-1} = \mathbf{x}_{t-\mu}^{t-1}).$$

If we let the initial cost of the first μ symbols be $w_0(\mathbf{x}_1^\mu) = -\log_2 P(\mathbf{X}_1^\mu = \mathbf{x}_1^\mu)$, we find that the total cost of a sequence $\mathbf{x}_1^N$ is $w(\mathbf{x}_1^N) = -\log_2 P(\mathbf{X}_1^N = \mathbf{x}_1^N)$.

For a random process with distribution Q and an entropy rate $H(Q)$, the *average cost* is $\bar{w}(Q) = \limsup_{N\to\infty} N^{-1} E\left[w(\mathbf{X}_1^N)\right]$, where the expectation is with respect to Q. This can be expressed in closed form whenever Q corresponds to a stationary Markov-driven process defined on some graph G with output labels $\{x(e)\}$ and transition probabilities $\{q_e\}$, and the labels of all μ-step paths into each state v equal $\mathbf{x}_1^\mu(v)$; i.e.,

$$\bar{w}(Q) = \sum_{v\in V} \pi_v(Q) \sum_{e:\sigma(e)=v} q_e w\left(x(e) \mid \mathbf{x}_1^\mu(v)\right). \tag{5}$$

More importantly, the average cost is related to the K-L divergence rate by

$$\bar{w}(Q) = H(Q) + D(Q||P). \tag{6}$$

This leads to the following theorem.

Theorem 3: For any finite-order Markov process with distribution P, let

$$D^*(R) = \min_{Q:H(Q)=R} D(Q||P)$$

be the minimum divergence rate over all random process distributions Q that have an entropy rate $H(Q) = R$. $D^*(R)$ can be represented parametrically (using Theorem 2) as

$$D^*(C(W(s))) = W(s) - C(W(s)),$$

where the finite-memory cost function $w(x_t|x_{t-\mu}^{t-1}) = -\log_2 P(X_t = x_t | \mathbf{X}_{t-\mu}^t = \mathbf{x}_{t-\mu}^t)$.

Proof: One can express the capacity as $C(W) = \max H(Q)$ over all Q such that $\bar{w}(Q) \le W$. By duality, we also have that $W = \min \bar{w}(Q)$ over all Q such that $H(Q) \ge C(W)$. From the monotonicity and concavity properties of $C(W)$, this is equal to the minimum over all Q such that $H(Q) = C(W)$. Finally, using equation (6), we have that

$$\begin{aligned} W &= \min_{Q:H(Q)=C(W)} \bar{w}(Q) \\ &= \min_{Q:H(Q)=C(W)} D(Q||P) + C(W). \end{aligned}$$

Therefore, we can solve for $D^*(C(W))$, and using the result in Theorem 2 we can represent this relationship parametrically as $D^*(C(W(s))) = W(s) - C(W(s))$. ∎

Notice that for invertible encoders with rate strictly less than the entropy rate of P, the minimum achievable divergence rate is nonzero. Interestingly, a non-invertible encoder might achieve an arbitrarily small divergence rate, but in this case the entropy rate of the encoder output may be actually less than the encoder rate. Examples for non-invertible and invertible encoders are given in the next section.

IV. Inner Shaping Encoder Design

We now revisit two encoder construction methods developed earlier in [1], [2], and prove that there are cases where each method can attain arbitrarily low K-L divergence rates with respect to the target input process. (Full details for both of these methods can be found in [10].)

A. Shaping Encoders Based on Quantization

One simple approach to shaping encoder design which generalized Gallager's construction [6, p. 208] was presented in [2], and led to near-capacity coding systems. Briefly, given a target finite-state Markov process with M states, the method constructs an M-state, rate $k:n$ encoder by taking the n-th power of the graph that describes the target process, and

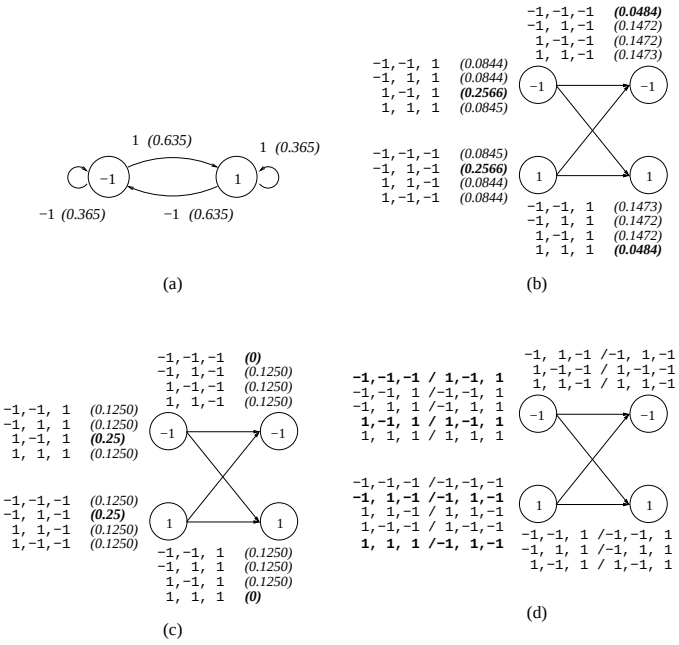

Fig. 1. Construction Method 1 example for the target binary Markov process in (a), with labels for symbols (and branch probabilities). *Step 1:* Set $k, n = 3$. (b) *Step 2:* Derive 3-step branch probabilities for target process. (c) *Step 3:* Approximate to multiples of 1/8. (d) *Step 4:* Tag 3:3 encoder with labels for inputs/outputs.

then approximating the n-step transition probabilities with a distribution in multiples of 2^{-k} (while ensuring that transitions which do not occur have zero probability). An encoder which induces such a distribution can then be produced by assigning each branch with input labels accordingly. We refer to this technique as *Construction Method 1*, and a step-by-step example given in Figure 1.

Theorem 4: Consider any target finite-order binary aperiodic and irreducible Markov process with distribution P and entropy rate H. For any $\epsilon > 0$ and rate $R \geq H$, there exists a rate $\lfloor nR \rfloor : n$ encoder generated by Construction Method 1 which induces a distribution Q such that $D(Q||P) < \epsilon$, for sufficiently large n.

Proof: Let $\{p_e\}$ be the set of branch probabilities and $G = (V, E, L)$ be the graph that together describe the target μ-th order binary Markov process with distribution P. From this, we can determine the n-step branch probabilities $\{p_e^{(n)}\}$, and the power graph $G^n = (V, E', L')$. In this power graph, each edge $e \in E'$ has the output label $\mathbf{x}_1^n(e)$, and for each each state $v \in V$, all of the μ-step paths into that state have the same label $\mathbf{x}_{1-\mu}^0(v)$. This allows us to write

$$p_e^{(n)} = P\left(\mathbf{X}_1^n = \mathbf{x}_1^n(e) \,\middle|\, \mathbf{X}_{1-\mu}^0 = \mathbf{x}_{1-\mu}^0\left(\sigma(e)\right)\right). \quad (7)$$

From Construction Method 1, we know that the induced Markov distribution Q from the encoder output can be represented by the set of probabilities $\{q_e\}$ also defined on this graph G^n. Therefore, noting the relationship between the average cost (5) and the K-L divergence rate (6), we have

$$D(Q||P) = \frac{1}{n} \sum_{v \in V} \pi_v(Q) \sum_{e:\sigma(e)=v} q_e \log_2 \frac{q_e}{p_e^{(n)}}, \quad (8)$$

where $\pi_v(Q)$ is the stationary state distribution of Q. (A similar expression can be found in [14], though we introduce a slight generalization for n symbols per edge.) Note that (8) is always non-negative and well-defined for our situation, because $q_e = 0$ whenever $p_e^{(n)} = 0$.

Now for each state $v \in V$, let $A_\delta^{(n)}(v)$ be the δ-typical set with respect to P [15, p. 51] and conditioned on an initial state v. Following the Shannon-McMillan-Breiman theorem [15, p. 474], for any $0 < \delta < 1$, there exists an N_0 such that $P(\mathbf{X}_1^n \in A_\delta^{(n)}(v)) > 1 - \delta$ for all $n > N_0$. Consequently, for any state v, we can bound the size of the typical set with $\left|A_\delta^{(n)}(v)\right| \geq (1-\delta)2^{n(H-\delta)}$ [15, Theorem 3.1.2], since we assumed the target Markov process was irreducible and aperiodic.

When applying Construction Method 1, this asymptotic equipartition property also allows us to closely approximate $p_e^{(n)}$ with a uniform distribution when n is large. That is, assuming $n > N_0$, we find some subset $B_\delta^{(n)}(v) \subset A_\delta^{(n)}(v)$ for each state $v \in V$ such that $\log_2 \left|B_\delta^{(n)}(v)\right| = M \stackrel{\text{def}}{=} \lfloor n(H-\delta) + \log_2(1-\delta) \rfloor$. Then we assign

$$q_e = \begin{cases} 2^{-M}, & \text{if } \mathbf{x}_1^n(e) \in B_\delta^{(n)}(\sigma(e)) \\ 0 & \text{otherwise.} \end{cases}$$

Finally, for us to derive an encoder that induces this uniform distribution, we need to make sure that all $2^{\lfloor nR \rfloor}$ encoder input sequences can be evenly distributed among the 2^M sequences in $B_\delta^{(n)}(v)$; i.e., we need $\lfloor nR \rfloor \geq M$. But because R is chosen so that $R \geq H$, we already have that $\lfloor nR \rfloor \geq M$ for any $1 > \delta > 0$. Thus, the approach above yields an encoder which approximates $\{p_e^{(n)}\}$ and has $q_e = 0$ whenever $p_e^{(n)} = 0$.

Given the encoder above, we can then calculate the divergence rate between the two process distributions Q and P. By substitution of q_e into (8), we get

$$\begin{aligned} D(Q||P) &= \frac{1}{n} \sum_v \pi_\nu(Q) \sum_{e \in B_\delta^{(n)}(v)} 2^{-M} \log_2 \frac{2^{-M}}{p_e^{(n)}} \\ &\leq \frac{1}{n} \sum_v \pi_\nu(Q) \sum_{e \in B_\delta^{(n)}(v)} 2^{-M} \log_2 \frac{2^{-M}}{2^{-n(H+\delta)}} \\ &= \frac{1}{n}\left(-M + n(H+\delta)\right). \end{aligned}$$

The inequality reflects the fact that, since the edges e satisfy $\mathbf{x}_1^n(e) \in B_\delta^{(n)}(v) \subset A_\delta^{(n)}$, we have $p_e^{(n)} \geq 2^{-n(H+\delta)}$. By further noting that $M \geq n(H-\delta) + \log_2(1-\delta) - 1$, we have $D(Q||P) \leq 2\delta + \frac{1}{n}\left(1 - \log_2(1-\delta)\right)$. Since this last bound holds for any $1 > \delta > 0$ and all sufficiently large n, $D(Q||P)$ can be made arbitrarily small. ∎

B. Graphical Representations for Typical Sequences

In [1] (see also [10]), another method for encoder design was developed by exploiting the connection between typicality

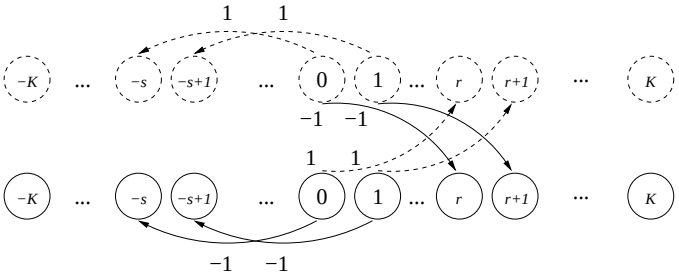

Fig. 2. Constraint graph which generates typical sequences with respect to a binary Markov process distribution with $P(X_t = 1 \,|\, X_{t-1} = -1) = P(X_t = 1 \,|\, X_{t-1} = -1) = s/(r+s)$. Such a graph is used in Construction Method 2.

and graphical representations of sequences. More specifically, for some target process with distribution P and entropy H, it was shown that, for any $\epsilon > 0$, a constraint graph G could always be constructed such that the capacity of the graph was greater than $H - \epsilon$, and such that sequences generated from G were typical with respect to P, i.e., $S_N(G) \subset A_\epsilon^{(N)}$ for large enough N. An example of such a constraint graph is given in Figure 2. Moreover, one could then derive finite-state encoders for typical sequences by applying the state-splitting algorithm or other code design methodologies to these graphs [16], and ultimately produce an encoder for typical sequences that has a rate of at least $H - \delta$, for any $\delta > 0$ [1]. By further examining this overall technique, which we term *Construction Method 2*, we arrive at the next theorem.

Theorem 5: Let P be the distribution of the target process with entropy rate H, and let Q be the distribution of the output process for a rate $k : n$ encoder generated by Construction Method 2 driven by i.i.d. equiprobable binary inputs. Then $D(Q||P) = H - k/n$, and thus, the divergence rate between the encoder output process and target process can be made arbitrarily small.

Proof: This can actually be proved without referring to the explicit details of the encoder graph. That is, let $S_{nN}(\mathcal{E}, v)$ be the set of all length-nN sequences generated from state v of a rate $k : n$ encoder $\mathcal{E}$ obtained by Construction Method 2. Noting that $\mathcal{E}$ is an invertible encoder, it follows that $|S_{nN}(\mathcal{E}, v)| = 2^{kN}$. All of these sequences are equiprobable when the encoder is driven with i.i.d. equiprobable binary inputs. This allows us to bound the K-L divergence for a sequence of N blocks, i.e.,

$$\begin{aligned} D_{(nN)}(Q||P) &= \frac{1}{nN} \sum_{\mathbf{x}_1^{nN} \in S_{nN}(\mathcal{E})} q(\mathbf{x}_1^{nN}) \log_2 \frac{q(\mathbf{x}_1^{nN})}{p(\mathbf{x}_1^{nN})} \\ &= \frac{1}{nN} \sum_{\mathbf{x}_1^{nN} \in S_{nN}(\mathcal{E})} 2^{-kN} \log_2 \frac{2^{-kN}}{p(\mathbf{x}_1^{nN})} \\ &\le \frac{1}{nN} \sum_{\mathbf{x}_1^{nN} \in S_{nN}(\mathcal{E})} 2^{-kN} \log_2 \frac{2^{-kN}}{2^{-(H+\epsilon)nN}} \\ &= H - \frac{k}{n} + \epsilon. \end{aligned}$$

More importantly, this implies that for any $\epsilon > 0$,

$$\limsup_{N\to\infty} D_{(nN)}(Q||P) \le H - \frac{k}{n} + \epsilon.$$

By similar arguments we can show that for any $\epsilon > 0$ the limit-infimum is bounded from below by $H - k/n - \epsilon$, and so the theorem is proved. ∎

V. Further Remarks

Alternatively, one might design the inner code such that mutual information rate on the channel for the encoder-induced input process, $I(\mathcal{X}; \mathcal{Y})$, is close to that of the target input process, $I(\mathcal{X}^*; \mathcal{Y})$. However, closed-form expressions for these quantities are not yet known, so in this paper we adopted the more convenient criterion of minimizing the divergence rate $D(Q||P)$. Although a vanishing $D(Q||P)$ may be sufficient if one desires an arbitrarily small gap between $I(\mathcal{X}; \mathcal{Y})$ and $I(\mathcal{X}^*; \mathcal{Y})$, this condition is not actually necessary. The analysis of [17] examines the set of necessary conditions, and the construction methods of [8] are based on other criteria and still achieve rates very near capacity.

Finally, since we related shaping encoder design to coding for channels with cost constraints, one might use another method to construct encoders which minimize the divergence rate, such as that of Khayrallah and Neuhoff [18].

Appendix I
Proof of Theorem 2

To our knowledge, the relationship between Theorem 2 and Theorem 1 (from [11]) has not appeared in the literature, but the analysis we employ below arises in other problems such as the computation of asymptotic weight spectra for convolutional codes (e.g., see Pfister [19, Theorem 3.3.6], and references therein). Moreover, it is generally accepted that the result in [11] is a lower bound to capacity even though an explicit code construction is not given. Details for the straightforward code construction from typical sequences can be found in [10].

Proof: We need only show that the capacity of a channel with cost constraints can be upper bounded by $C(W(s)) \le \log_2 \lambda(s) - W(s) \log_2 s$, for $0 \le s < 1$, where

$$W(s) = \frac{s}{\lambda(s)} \frac{\partial}{\partial s} \lambda(s) = \frac{1}{\lambda(s)} \sum_{i,j} u_i(s) w(i,j) s^{w(i,j)} v_j(s). \tag{9}$$

The lower bound follows from Theorem 1. (See [10].)

Let us begin by considering some fixed cost constraint W, and for the set $S_N(W)$ let us define a cost-enumerating function $B(N, W, s) = \sum_{h \le WN} b_h s^h$, where b_h is the number of sequences with total cost h. This enumerator can also be written as

$$B(N, W, s) = |S_N(W)| \sum_{h \le WN} \alpha_h s^h,$$

where α_h is the fraction of sequences in $S_N(W)$ with total cost h. For any $0 \le s < 1$, as a consequence of Jensen's

inequality (e.g., [15, p. 25]), we have

$$\begin{aligned} |S_N(W)| \sum_{h \le WN} \alpha_h s^h &\ge |S_N(W)|\, s^{\sum_{h \le WN} \alpha_h h} \\ &\ge |S_N(W)|\, s^{WN \sum_{h \le WN} \alpha_h} \\ &= |S_N(W)|\, s^{WN}. \end{aligned}$$

Thus, $B(N, W, s) \ge |S_N(W)|\, s^{WN}$.

Now it also follows that

$$\sum_{i,j} \left[\mathbf{A}(s)^N\right]_{i,j} \ge B(N, W, s),$$

because $B(N, W, s)$ only enumerates a subset of all sequences. If we assume that the enumerator matrix is irreducible, i.e., for each (i, j) there exists an ℓ such that $\left[\mathbf{A}(s)^\ell\right]_{i,j} > 0$, then its largest (nonnegative) eigenvalue $\lambda(s)$ must have left and right eigenvectors with strictly positive components; and this can be used to show that

$$\sum_{i,j} \left[\mathbf{A}(s)^N\right]_{i.j} \le \lambda(s)^N \frac{\sum v_i(s)}{\min_i v_i(s)} = K\lambda(s)^N,$$

as in [16, p. 1667]. Therefore, we can conclude that

$$|S_N(W)|\, s^{WN} \le B(N, W, s) \le K\lambda(s)^N,$$

and thus

$$C(W) = \limsup_{N\to\infty} \frac{1}{N} \log_2 |S_N(W)| \le \log_2 \lambda(s) - W \log_2 s.$$

Moreover, this upper bound can be tightened by maximizing the right-hand side with respect to s, i.e., by taking derivatives and determining the extremal points. This results in a dependence of W on s corresponding to the middle expression in equation (9). Therefore, to complete the proof of the upper bound it remains to be shown that

$$\frac{s}{\lambda(s)} \frac{\partial}{\partial s} \lambda(s) = \frac{1}{\lambda(s)} \sum_{i,j} u_i(s) w(i, j) s^{w(i,j)} v_j(s).$$

Recall that the eigenvectors $\mathbf{u}(s)$ and $\mathbf{v}(s)$ are assumed to be normalized so that $\mathbf{u}(s)\mathbf{v}(s)^T = 1$. Then, $\lambda(s) = \mathbf{u}(s)\mathbf{A}(s)\mathbf{v}(s)^T = \sum_{i,j} u_i(s) s^{w(i,j)} v_j(s)$. Taking the derivative, we find

$$\begin{aligned} \frac{\partial}{\partial s}\lambda(s) &= \sum_{i,j} u_i(s) w(i, j) s^{w(i,j)-1} v_j(s) \; + \\ &\quad \sum_{i,j} s^{w(i,j)} \left(v_j(s) \frac{\partial}{\partial s} u_i(s) + u_i(s) \frac{\partial}{\partial s} v_j(s) \right). \end{aligned} \tag{10}$$

Note that

$$\begin{aligned} \sum_{i,j} s^{w(i,j)} v_j(s) \frac{\partial}{\partial s} u_i(s) &= \sum_i \left(\frac{\partial}{\partial s} u_i(s) \right) \sum_j s^{w(i,j)} v_j(s) \\ &= \sum_i \left(\frac{\partial}{\partial s} u_i(s) \right) \lambda(s) v_j(s). \end{aligned}$$

Using this argument, the second expression on the right side of (10) becomes

$$\lambda(s) \sum_i \left(\left(\frac{\partial}{\partial s} u_i(s) \right) v_i(s) + \left(\frac{\partial}{\partial s} v_i(s) \right) u_i(s) \right),$$

which equals $\lambda(s) \frac{\partial}{\partial s} \left(\mathbf{u}(s)\mathbf{v}(s)^T \right)$. But from the assumption that $\mathbf{u}(s)\mathbf{v}(s)^T = 1$, this derivative is zero. Therefore, we conclude that equation (10) simplifies to

$$\frac{\partial}{\partial s}\lambda(s) = \sum_{i,j} u_i(s) w(i, j) s^{w(i,j)-1} v_j(s),$$

and the result in (9) readily follows. Thus the upper bound is proved. ■

REFERENCES

[1] J. B. Soriaga and P. H. Siegel, "On distribution shaping codes for partial-response channels," in *Proceedings 41st Annual Allerton Conference on Communication, Control, and Computing*, (Monticello, IL, USA), pp. 468–477, October 2003.

[2] J. B. Soriaga and P. H. Siegel, "Near-capacity coding systems for partial-response channels," in *Proceedings of IEEE International Symposium on Information Theory*, (Chicago, IL, USA), p. 267, June 2004.

[3] D. Arnold and H. Loeliger, "On the information rate of binary-input channels with memory," in *Proceedings IEEE International Conference on Communications*, (Helsinki, Finland), pp. 2692–2695, June 2001.

[4] V. Sharma and S. K. Singh, "Entropy and channel capacity in the regenerative setup with applications to Markov channels," in *Proceedings of IEEE International Symposium on Information Theory*, (Washington, DC, USA), p. 283, June 2001.

[5] H. D. Pfister, J. B. Soriaga, and P. H. Siegel, "On the achievable information rates of finite state ISI channels," in *Proceedings IEEE Global Telecommunications Conference*, (San Antonio, Texas, USA), pp. 2992–2996, November 2001.

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

[7] X. Ma, N. Varnica, and A. Kavčić, "Matched information rate codes for binary ISI channels," in *Proceedings of IEEE International Symposium on Information Theory*, (Lausanne, Switzerland), p. 269, June 2002.

[8] N. Varnica, X. Ma, and A. Kavčić, "Capacity-approaching codes for partial response channels," in *Handbook of Coding and Signal Processing for Recording Systems* (B. Vasic, ed.), CRC Press, 2004.

[9] D. Doan and K. R. Narayanan, "Design of good low rate codes for ISI channels based on spectral shaping," in *Proceedings International Symposium on Turbo Codes & Related Topics*, (Brest, France), pp. 31–34, September 2003.

[10] J. Soriaga, *On Near-Capacity Code Design for Partial-Response Channels*. PhD thesis, University of California, San Diego, La Jolla, CA, USA, March 2005.

[11] J. Justesen and T. Høholdt, "Maxentropic Markov chains," *IEEE Transactions on Information Theory*, vol. 30, pp. 665–667, July 1984.

[12] K. Marton and P. C. Shields, "Ergodic processes and zero divergence," in *Proceedings of IEEE International Symposium on Information Theory*, (Budapest, Hungary), p. 76, IEEE, June 1991.

[13] J. Chen and P. H. Siegel, "Markov processes asymptotically achieve the capacity of finite state intersymbol interference channels," in *Proceedings of IEEE International Symposium on Information Theory*, (Chicago, IL, USA), p. 349, June 2004.

[14] Z. Rached, F. Alajaji, and L. L. Campbell, "The Kullback-Leibler divergence rate between Markov sources," *IEEE Transactions on Information Theory*, vol. 50, pp. 917–921, May 2004.

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

[16] B. H. Marcus, R. M. Roth, and P. H. Siegel, *Handbook of Coding Theory*, ch. 20, pp. 1635–1764. New York, NY, USA: Elsevier Science, 1998.

[17] S. Shamai and S. Verdú, "The empirical distribution of good codes," *IEEE Transactions on Information Theory*, vol. 43, pp. 836–846, May 1997.

[18] A. S. Khayrallah and D. L. Neuhoff, "Coding for channels with cost constraints," *IEEE Transactions on Information Theory*, vol. 42, pp. 854–867, May 1996.

[19] H. D. Pfister, *On the Capacity of Finite State Channels and the Analysis of Convolutional Accumulate-m Codes*. PhD thesis, University of California, San Diego, La Jolla, CA, USA, March 2003.