

# New Bounds on the Entropy Rate of Hidden Markov Processes

Erik Ordentlich  
 HP Laboratories  
 1501 Page Mill Road  
 Palo Alto, California 94304, USA  
 e-mail: eord@hpl.hp.com

Tsachy Weissman<sup>1</sup>  
 Department of Electrical Engineering  
 Stanford University  
 Stanford, CA 94305, USA  
 e-mail: tsachy@stanford.edu

**Abstract** — Let  $\{X_t\}$  be a stationary finite-alphabet Markov chain and  $\{Z_t\}$  denote its noisy version when corrupted by a discrete memoryless channel. Let  $P(X_t \in \cdot | Z_{-\infty}^t)$  denote the conditional distribution of  $X_t$  given all past and present noisy observations, a simplex-valued random variable. We present a new approach to bounding the entropy rate of  $\{Z_t\}$  by approximating the distribution of this random variable. This approximation is facilitated by the construction and study of a Markov process whose stationary distribution determines the distribution of  $P(X_t \in \cdot | Z_{-\infty}^t)$ . To illustrate the efficacy of this approach, we specialize it and derive concrete bounds for the case of a binary Markov chain corrupted by a binary symmetric channel (BSC). These bounds are seen to capture the behavior of the entropy rate in various asymptotic regimes.

## I. INTRODUCTION

Let  $\{X_t\}$  be a stationary Markov chain and  $\{Z_t\}$  denote its noisy version when corrupted by a discrete memoryless channel. The components of these processes take values, respectively, in the finite alphabets  $\mathcal{X}$  and  $\mathcal{Z}$ . We let  $\mathcal{K}$  denote the transition kernel of the Markov chain (which can be thought of as a  $|\mathcal{X}| \times |\mathcal{X}|$  matrix) and  $\mathcal{C}$  denote the  $|\mathcal{X}| \times |\mathcal{Z}|$  channel transition matrix.  $\{Z_t\}$  is known as a hidden Markov process. Its distribution and, a fortiori, its entropy rate which we denote by  $\overline{H}(Z)$ , are completely determined by the pair  $(\mathcal{K}, \mathcal{C})$ . However, the explicit form of  $\overline{H}(Z)$  as a function of this pair is unknown.

Understanding the entropy rate of hidden Markov processes (HMPs) is motivated by the fact that these processes occur naturally in the modelling of information sources [EM02]. Also, noise processes in additive noise channels are often hidden Markov processes and the characterization of channel capacity in such cases boils down to finding the entropy rate of the noise [MBD89].

Recent approaches to quantifying the entropy rate include use of the bounds of [CT91, Section 4.4] (see also [Bir62]) in [HJ99], Monte Carlo simulation [HGG03], Lyapunov exponents [HGG03, JSS04], statistical mechanics [ZKD04], and more [EBTBH04].

In this work we present another approach to the problem and illustrate its use (for a few special cases) in deriving concrete bounds that are seen to be tight in various asymptotic regimes. For example, we show that in the case of a binary Markov chain observed through a BSC with crossover  $\delta$ , when the chain jumps with probability one from 1 to 0, and with probability  $0 \leq \pi_0 < 1$  from 0 to 1, as  $\delta$  tends to 0,

$$\overline{H}(Z) = \overline{H}(X) + \frac{\pi_0(2 - \pi_0)}{1 + \pi_0} \delta \log \frac{1}{\delta} + O(\delta),$$

<sup>1</sup>Part of this work was performed while this author was visiting Hewlett-Packard Laboratories.

where  $\overline{H}(Z)$  is the entropy rate (in nats) of the BSC( $\delta$ )-corrupted chain, while  $\overline{H}(X)$  is that of the underlying noise-free chain. This should be contrasted with recent results showing that the leading term is of order  $\delta$  rather than  $\delta \log \frac{1}{\delta}$  when the jump from 1 to 0 has probability less than 1 [JSS04, OW04a, ZKD04].

The summary is organized as follows. We start in Section II with a brief description of the basic idea. In the remaining sections we illustrate how it is carried out in the case of a BSC-corrupted Markov chain. Section III details the construction of a Markov process, which is a key component of our approach. In Section IV we then use it to derive bounds for the case where the noiseless binary Markov chain is symmetric, leaving the non-symmetric chain to Section V where we focus on the case where the transition probability from state 1 to 0 is 1. We conclude in Section VI.

## II. THE GENERIC APPROACH

Let  $H(Q)$  denote the entropy of a distribution  $Q$  on  $\mathcal{Z}$

$$H(Q) = \sum_{z \in \mathcal{Z}} Q(z) \log \frac{1}{Q(z)}.$$

Let  $\mathcal{M}(\mathcal{X})$  denote the simplex of distributions on  $\mathcal{X}$  and  $\beta_t$  be the  $\mathcal{M}(\mathcal{X})$ -valued random variable defined by

$$\beta_t(x) = P(X_t = x | Z_{-\infty}^t),$$

where  $\beta_t(x)$  denotes the  $x$ -th component of  $\beta_t$ . We denote this by

$$\beta_t = P(X_t \in \cdot | Z_{-\infty}^t).$$

The relationship

$$P(Z_t \in \cdot | Z_{-\infty}^t) = \beta_t * \mathcal{K} * \mathcal{C}$$

(where  $*$  denotes composition of kernels<sup>1</sup>) implies

$$\begin{aligned} \overline{H}(Z) &= EH(P(Z_t \in \cdot | Z_{-\infty}^t)) = EH(\beta_t * \mathcal{K} * \mathcal{C}) \\ &= \int_{\mathcal{M}(\mathcal{X})} H(\beta * \mathcal{K} * \mathcal{C}) d\mu(\beta), \end{aligned} \quad (1)$$

where  $\mu$  denotes the distribution of  $\beta_t$  (which is, of course<sup>2</sup>, not explicitly known [Bla57]). An immediate consequence of (1) is

### Observation 1

$$\min_{\beta \in \mathcal{S}} H(\beta * \mathcal{K} * \mathcal{C}) \leq \overline{H}(Z) \leq \max_{\beta \in \mathcal{S}} H(\beta * \mathcal{K} * \mathcal{C}),$$

where  $\mathcal{S}$  denotes the support of  $\mu$ .

<sup>1</sup>Alternatively, viewing  $P(Z_t \in \cdot | Z_{-\infty}^t)$  and  $\beta_t$  as  $|\mathcal{X}|$ -dimensional column vectors and  $\mathcal{K}, \mathcal{C}$  as matrices of appropriate dimensions,  $*$  can be thought of as matrix multiplication.

<sup>2</sup>Otherwise the entropy rate would also be known.

Trivial as this observation may seem, it was shown in [OW04a] to lead to useful bounds in cases where bounds on the support set  $\mathcal{S}$  are obtainable, and this set is significantly smaller than  $\mathcal{M}(\mathcal{X})$ .

The bounds of Observation (1), which depend on  $\mu$  through its support only, can be refined by partitioning  $\mathcal{S}$  into subsets, using a similar bound on each of the subsets, and weighting these according to their probabilities. More precisely:

**Observation 2** *For any countable collection  $\{I_i\}$  of pairwise disjoint sets  $I_i \subseteq \mathcal{M}(\mathcal{A})$  covering  $\mathcal{S}$  (i.e., for which  $\mathcal{S} \subseteq \bigcup_i I_i$ ),*

$$\sum_i \mu(I_i) \inf_{\beta \in I_i} H(\beta * \mathcal{K} * \mathcal{C}) \leq \overline{H}(Z) \leq \sum_i \mu(I_i) \sup_{\beta \in I_i} H(\beta * \mathcal{K} * \mathcal{C}). \quad (2)$$

Since  $\mu$  is unknown,  $\mu(I_i)$  will also be unknown in general. However, for certain choices of  $\{I_i\}$ , and in certain regions of the space of parameters governing the HMP, the bounds in (2) can be either explicitly evaluated or closely bounded. This is done by constructing a Markov process which is more tractable than the  $\{\beta_t\}$  process. The stationary distribution of this Markov process is directly and simply related to the distribution of  $\beta_t$ .

For brevity and concreteness in illustrating this approach and the construction of the Markov process we concentrate below on the case where  $\{Z_t\}$  is a BSC-corrupted binary Markov chain. Description of the more general case will be given in [OW04b].

### III. THE BSC-CORRUPTED BINARY MARKOV CHAIN

For this case  $\mathcal{X} = \mathcal{Z} = \{0, 1\}$ . The Markov transition matrix and the channel matrix are, respectively,

$$\mathcal{K} = \begin{pmatrix} 1 - \pi_{01} & \pi_{01} \\ \pi_{10} & 1 - \pi_{10} \end{pmatrix}, \mathcal{C} = \begin{pmatrix} 1 - \delta & \delta \\ \delta & 1 - \delta \end{pmatrix}, \quad (3)$$

where we assume without loss of generality  $\delta \leq 1/2$ ,  $\pi_{01} \leq \pi_{10}$ . For this case the standard forward recursions [EM02] are easily shown [OW04a] to assume the form

$$\frac{\beta_i(0)}{1 - \beta_i(0)} = \left[ \frac{1 - \delta}{\delta} \right]^{1 - 2Z_i} g \left( \frac{\beta_{i-1}(0)}{1 - \beta_{i-1}(0)} \right), \quad (4)$$

where

$$g(x) \triangleq \frac{x(1 - \pi_{0,1}) + \pi_{1,0}}{x\pi_{0,1} + (1 - \pi_{1,0})}. \quad (5)$$

Equivalently, this can be expressed as

$$l_i = (2Z_i - 1) \log \left[ \frac{1 - \delta}{\delta} \right] + h(l_{i-1}), \quad (6)$$

where  $l_i = \log \frac{\beta_i(1)}{1 - \beta_i(1)}$  and

$$h(x) \triangleq \log \frac{\pi_{0,1} + e^x(1 - \pi_{1,0})}{(1 - \pi_{0,1}) + e^x\pi_{1,0}}. \quad (7)$$

Note that for this case (1) gives

$$\begin{aligned} \overline{H}(Z) &= Eh_b([\beta_i(1)(1 - \pi_{10}) + (1 - \beta_i(1))\pi_{01}] * \delta) \\ &= Eh_b \left( \left[ \frac{e^{l_i}}{1 + e^{l_i}}(1 - \pi_{10}) + \frac{1}{1 + e^{l_i}}\pi_{01} \right] * \delta \right), \end{aligned} \quad (8)$$

where henceforth  $*$  boils down to binary convolution defined by  $p * q = (1 - p)q + (1 - q)p$  and

$$h_b(x) = -x \ln x - (1 - x) \ln(1 - x)$$

is the binary entropy function (in nats).

When specialized to the case  $\pi_{10} = \pi_{01} = \pi$ , we obtain the evolution

$$l_i = (2Z_i - 1) \log \left[ \frac{1 - \delta}{\delta} \right] + h(l_{i-1}), \quad (9)$$

where  $h(x) = \log \frac{e^x(1 - \pi) + \pi}{e^x\pi + (1 - \pi)}$ . Specializing (8) for this case gives

$$\overline{H}(Z) = Eh_b \left( \frac{e^{l_i}}{1 + e^{l_i}} * \pi * \delta \right). \quad (10)$$

The distribution of  $\beta_i$  (or, equivalently, of  $l_i$ ) is, evidently, key to the evaluation of the entropy rate. Although  $\{\beta_i\}$  was shown to be a Markov process by Blackwell in [Bla57], its analysis turns out to be quite elusive. In what follows we construct another, more tractable, Markov process, whose stationary distribution is closely related to (and determines) the distribution of  $\beta_i$ .

## A. AN ALTERNATIVE MARKOV PROCESS

### A.1. THE SYMMETRIC CASE

To illustrate the idea behind the construction of the alternative Markov process in its simplest form, we start with the symmetric case. Note that *conditioned on the event  $X_i = 1$* , the two summands on the right side of (4) are *independent* with

$$(2Z_i - 1) \log \left[ \frac{1 - \delta}{\delta} \right] = \begin{cases} \log \frac{1 - \delta}{\delta} & \text{w.p. } 1 - \delta \\ -\log \frac{1 - \delta}{\delta} & \text{w.p. } \delta. \end{cases} \quad (11)$$

Furthermore, we have

$$P(h(l_{i-1})|X_i = 1) = \sum_j P(h(l_{i-1}), X_{i-1} = j|X_i = 1)$$

$$= \sum_j P(X_{i-1} = j|X_i = 1)P(h(l_{i-1})|X_{i-1} = j)$$

$$= \pi P(h(l_{i-1})|X_{i-1} = 0) + (1 - \pi)P(h(l_{i-1})|X_{i-1} = 1)$$

$$= \pi P(-h(l_{i-1})|X_{i-1} = 1) + (1 - \pi)P(h(l_{i-1})|X_{i-1} = 1),$$

the last equality following since, by symmetry,  $l_{i-1}|X_{i-1} = 0 \stackrel{d}{=} -l_{i-1}|X_{i-1} = 1$  and due to the fact that  $h(\cdot)$  is anti-symmetric. As a consequence, the stationary marginal distribution (which can be shown to be unique [OW04b]) of the following auto-regressively defined 1st-order Markov process is seen to be given by  $P(l_i|X_i = 1)$  (i.e., the distribution of  $l_i$  conditioned on the event  $X_i = 1$ ):

$$Y_i = r_i \log \frac{1 - \delta}{\delta} + s_i h(Y_{i-1}), \quad (12)$$

where  $\{r_i\}$  and  $\{s_i\}$  are independent i.i.d. sequences with

$$r_i = \begin{cases} -1 & \text{w.p. } \delta \\ 1 & \text{w.p. } 1 - \delta, \end{cases} \quad s_i = \begin{cases} -1 & \text{w.p. } \pi \\ 1 & \text{w.p. } 1 - \pi. \end{cases} \quad (13)$$

Note that in terms of  $Y_i$  (assuming it is started from its stationary distribution), from (10) we get

$$\bar{H}(Z) = \frac{1}{2}E \left[ h_b \left( \frac{e^{l_i}}{1 + e^{l_i}} * \pi * \delta \right) | X_i = 1 \right] \quad (14)$$

$$+ \frac{1}{2}E \left[ h_b \left( \frac{e^{l_i}}{1 + e^{l_i}} * \pi * \delta \right) | X_i = 0 \right] \quad (15)$$

$$= Eh_b \left( \frac{e^{Y_i}}{1 + e^{Y_i}} * \pi * \delta \right), \quad (16)$$

the second equality following by the facts that  $l_i | X_i = 0 \stackrel{d}{=} l_i | X_i = 1$  and that  $h_b \left( \frac{e^y}{1 + e^y} * \pi * \delta \right) = h_b \left( \frac{e^{-y}}{1 + e^{-y}} * \pi * \delta \right)$  for all  $y$ .

### A.2. THE NON-SYMMETRIC CASE

Conditioned on the event  $X_i = 1$ , the two summands on the right side of (6) are independent with

$$(2Z_i - 1) \log \left[ \frac{1 - \delta}{\delta} \right] = \begin{cases} \log \frac{1 - \delta}{\delta} & \text{w.p. } 1 - \delta \\ -\log \frac{1 - \delta}{\delta} & \text{w.p. } \delta \end{cases} \quad (17)$$

Furthermore, we have

$$P(h(l_{i-1}) | X_i = 1) = \sum_j P(h(l_{i-1}), X_{i-1} = j | X_i = 1)$$

$$= \sum_j P(X_{i-1} = j | X_i = 1) P(h(l_{i-1}) | X_{i-1} = j)$$

$$= \pi_{10} P(h(l_{i-1}) | X_{i-1} = 0) + (1 - \pi_{10}) P(h(l_{i-1}) | X_{i-1} = 1).$$

Similarly, conditioned on the event  $X_i = 0$ , the two summands on the right side of (6) are independent with

$$(2Z_i - 1) \log \left[ \frac{1 - \delta}{\delta} \right] = \begin{cases} \log \frac{1 - \delta}{\delta} & \text{w.p. } \delta \\ -\log \frac{1 - \delta}{\delta} & \text{w.p. } 1 - \delta \end{cases} \quad (18)$$

and

$$P(h(l_{i-1}) | X_i = 0) = \sum_j P(h(l_{i-1}), X_{i-1} = j | X_i = 0)$$

$$= \sum_j P(X_{i-1} = j | X_i = 0) P(h(l_{i-1}) | X_{i-1} = j)$$

$$= (1 - \pi_{01}) P(h(l_{i-1}) | X_{i-1} = 0) + \pi_{01} P(h(l_{i-1}) | X_{i-1} = 1).$$

Letting now  $\{q_i\}$ ,  $\{r_i\}$ ,  $\{s_i\}$ ,  $\{t_i\}$  be independent i.i.d. sequences with

$$q_i = \begin{cases} 1 & \text{w.p. } \delta \\ -1 & \text{w.p. } 1 - \delta \end{cases}, \quad r_i = \begin{cases} -1 & \text{w.p. } \delta \\ 1 & \text{w.p. } 1 - \delta \end{cases}, \quad (19)$$

and  $s_i \sim \text{Bernoulli}(\pi_{10})$ ,  $t_i \sim \text{Bernoulli}(\pi_{01})$ , we define the process  $\{(Y_i, U_i)\}$  via

$$Y_i = r_i \log \frac{1 - \delta}{\delta} + s_i h(U_{i-1}) + (1 - s_i) h(Y_{i-1}) \quad (20)$$

and

$$U_i = q_i \log \frac{1 - \delta}{\delta} + (1 - t_i) h(U_{i-1}) + t_i h(Y_{i-1}), \quad (21)$$

a Markov process with state space  $\mathbb{R}^2$ . Letting  $(Y, U)$  denote a generic pair having the stationary distribution of that process, it is clear that  $Y \stackrel{d}{=} l_i | X_i = 1$  and  $U \stackrel{d}{=} l_i | X_i = 0$ . When

combined with (8) this implies that the entropy rate  $\bar{H}(Z)$  is given by

$$\frac{\pi_{01}}{\pi_{01} + \pi_{10}} Eh_b \left( \left[ \frac{e^Y}{1 + e^Y} (1 - \pi_{10}) + \frac{1}{1 + e^Y} \pi_{01} \right] * \delta \right) + \frac{\pi_{10}}{\pi_{01} + \pi_{10}} Eh_b \left( \left[ \frac{e^U}{1 + e^U} (1 - \pi_{10}) + \frac{1}{1 + e^U} \pi_{01} \right] * \delta \right). \quad (22)$$

## IV. BOUNDS ON THE ENTROPY RATE FOR THE SYMMETRIC CHAIN

Assume throughout this section the case of a BSC-corrupted symmetric Markov chain with  $0 < \pi_{10} = \pi_{01} = \pi \leq 1/2$ . There is no loss of generality in assuming  $\pi \leq 1/2$  since the argument in [OW04a, Subsection 4-C] implies that the entropy rate when the Markov chain is symmetric with transition probability  $1 - \pi$  is the same as when it is  $\pi$ . The derivation of Subsection A.1 above implies:

**Theorem 1** *Let  $\{Y_i\}$  be the stationary Markov process whose evolution is given by (12). Let  $\{a_i\}_{i=1}^M, \{b_i\}_{i=1}^M$  be strictly increasing sequences of nonnegative reals such that  $a_k \leq b_k$  and  $a_{k+1} > b_k$  (i.e., the intervals  $[a_k, b_k]$  do not intersect). Assume further that  $\bigcup_{k=1}^M [a_k, b_k] \cup \bigcup_{k=1}^M [-b_k, -a_k]$  contains the support of  $Y_i$ . Then  $\bar{H}(Z)$  is lower bounded by*

$$\sum_{k=1}^M P(Y_i \in [-b_k, -a_k] \cup [a_k, b_k]) h_b \left( \frac{e^{b_k}}{1 + e^{b_k}} * \pi * \delta \right)$$

and upper bounded by

$$\sum_{k=1}^M P(Y_i \in [-b_k, -a_k] \cup [a_k, b_k]) h_b \left( \frac{e^{a_k}}{1 + e^{a_k}} * \pi * \delta \right).$$

*Proof:* Immediate from (16) and the decreasing monotonicity of  $h_b \left( \frac{e^y}{1 + e^y} * \pi * \delta \right)$  in the absolute value of  $y$ .  $\square$

When specialized to the case  $M = 1$ , Theorem 1 yields

**Corollary 1** *Let  $\{Y_i\}$  be the process in (12). Let  $0 \leq b \leq A$  be such that  $[-A, -b] \cup [b, A]$  contains the support of  $Y_i$ . Then*

$$h_b \left( \frac{e^A}{1 + e^A} * \pi * \delta \right) \leq \bar{H}(Z) \leq h_b \left( \frac{e^b}{1 + e^b} * \pi * \delta \right). \quad (23)$$

The lower bound of Corollary 1 is clearly optimized when taking  $A$  to be the upper endpoint of the support of  $Y_i$ . This point is readily seen, by observation of the dynamics of the process  $\{Y_i\}$  in (12), to be the solution to the equation

$$A = h(A) + \log \frac{1 - \delta}{\delta}, \quad (24)$$

namely

$$A = \log \frac{\alpha - 1 + (1 - \alpha)\pi + \sqrt{4\alpha\pi^2 + (1 - \alpha - (1 - \alpha)\pi)^2}}{2\pi}, \quad (25)$$

where  $\alpha = \frac{1 - \delta}{\delta}$ . Similarly, to optimize the upper bound,  $b$  should be taken as the lower endpoint of this support in the positive half of the real line. For the case where  $\delta$  is small enough so that the first term on the right side of (12) uniquely determines the sign of  $Y_i$  ("small enough" will be made explicit below), the value of this lower endpoint can be read from the

dynamics of the process in (12) (see proof of Lemma 1 below) to be given by

$$b = -h(A) + \log \frac{1-\delta}{\delta}. \quad (26)$$

Crude as the bounds of Corollary 1 may seem, they were shown in [OW04a] to convey non-trivial information (when substituting the values from (25) and (26)) regarding the behavior of the entropy rate in various asymptotic regimes (some will be mentioned below). In the remainder of this section we take one step of refinement beyond Corollary 1, studying the form of the bounds of Theorem 1 in the case  $M = 2$ , and their implications in some asymptotic regimes.

Define, in addition to  $A$  and  $b$  in (24) and (26),

$$a = -h(b) + \log \frac{1-\delta}{\delta} \quad (27)$$

and

$$B = h(b) + \log \frac{1-\delta}{\delta}. \quad (28)$$

**Lemma 1** *Assume either  $\pi \geq 1/4$  and  $\delta \leq 1/2$ , or  $\pi < 1/4$  and  $\delta < \frac{1}{2}(1 - \sqrt{1-4\pi})$ . More compactly, assume  $\delta \leq \frac{1}{2}(1 - \sqrt{\max\{1-4\pi, 0\}})$ . Then  $A, b, a$  and  $B$  (defined in (25), (26), (27) and (28)) satisfy  $0 \leq b \leq a < B \leq A$ , as well as:  $P(Y_i \in [B, A]) = (1-\delta)[\pi * (1-\delta)]$ ,  $P(Y_i \in [b, a]) = (1-\delta)[\pi * \delta]$ ,  $P(Y_i \in [-a, -b]) = \delta[\pi * (1-\delta)]$ , and  $P(Y_i \in [-A, -B]) = \delta[\pi * \delta]$ . In particular, the support of  $Y_i$  is contained  $[-A, -B] \cup [-a, -b] \cup [b, a] \cup [B, A]$ .*

*Proof:* That the  $A$  solving (24) is the upper end point of the support of  $Y_i$  and, by symmetry,  $-A$  its lower end point, is evident from (12). It was shown in [OW04a, Corollary 3] that in this region of the  $\pi - \delta$  plane  $Y_i \geq 0$  if and only if  $r_i = 1$ , in which case the smallest value  $Y_i$  can take is  $b = \log \frac{1-\delta}{\delta} - h(A)$ . This implies, by symmetry of the support of  $Y_i$ , that this support is contained in  $[-A, -b] \cup [b, A]$ . Furthermore, when  $Y_i > 0$  (i.e.,  $r_i = 1$ ), there are two possibilities. The first is that the second term on the right side of (12) is negative, in which case the most (least negative) it can be is  $-h(b)$ , implying that in this case  $Y_i \leq \log \frac{1-\delta}{\delta} - h(b) = a$ . The second possibility is that this second term is positive, in which case the least it can be is  $h(b)$  implying  $Y_i \geq \log \frac{1-\delta}{\delta} + h(b) = B$ . It follows that when  $Y_i > 0$  either  $Y_i \in [b, a]$  or  $Y_i \in [B, A]$ . Symmetry of the support of  $Y_i$  implies that this support is contained in  $[-A, -B] \cup [-a, -b] \cup [b, a] \cup [B, A]$ . It also follows that  $Y_i$  falls in the interval, say  $[b, a]$ , if and only if both  $r_i = 1$  and  $s_i h(Y_{i-1}) < 0$ , i.e.,

$$\begin{aligned} P(Y_i \in [b, a]) &= P(r_i = 1, s_i h(Y_{i-1}) < 0) \\ &= P(r_i = 1)P(\{s_i = 1, h(Y_{i-1}) < 0\} \\ &\quad \cup \{s_i = -1, h(Y_{i-1}) > 0\}) \\ &= (1-\delta)[(1-\pi)\delta + \pi(1-\delta)] \\ &= (1-\delta)[\pi * \delta]. \end{aligned}$$

Using similar reasoning gives

$$P(Y_i \in [B, A]) = P(r_i = 1, s_i h(Y_{i-1}) > 0) = (1-\delta)[\pi * (1-\delta)],$$

$$P(Y_i \in [-a, -b]) = P(r_i = -1, s_i h(Y_{i-1}) > 0) = \delta[\pi * (1-\delta)],$$

and

$$P(Y_i \in [-A, -B]) = P(r_i = -1, s_i h(Y_{i-1}) < 0) = \delta[\pi * \delta]. \square$$

Specializing Theorem 1 to the case  $M = 2$  and combining with Lemma 1 gives:

**Theorem 2** *For all  $\delta \leq \frac{1}{2}(1 - \sqrt{\max\{1-4\pi, 0\}})$*

$$\begin{aligned} &\{(1-\delta)[\pi * (1-\delta)] + \delta[\pi * \delta]\} h_b \left( \frac{e^A}{1+e^A} * \pi * \delta \right) \\ &+ \{(1-\delta)[\pi * \delta] + \delta[\pi * (1-\delta)]\} h_b \left( \frac{e^a}{1+e^a} * \pi * \delta \right) \\ &\leq \bar{H}(Z) \\ &\leq \{(1-\delta)[\pi * (1-\delta)] + \delta[\pi * \delta]\} h_b \left( \frac{e^B}{1+e^B} * \pi * \delta \right) \\ &+ \{(1-\delta)[\pi * \delta] + \delta[\pi * (1-\delta)]\} h_b \left( \frac{e^b}{1+e^b} * \pi * \delta \right), \end{aligned} \quad (29)$$

where  $A, B, a, b$  are as specified in (25)-(28).

As can be expected, the bounds in Theorem 2 turn out to be considerably tighter, in various asymptotic regimes, than those based on Corollary 1.

For example, in the ‘‘high SNR’’ regime where  $\delta \downarrow 0$ , the analysis in [OW04a, Section 5] established that  $\bar{H}(Z) - h_b(\pi) = \Theta(\delta)$ , while Theorem 2 can be shown [OW04b] to yield

$$\bar{H}(Z) = h_b(\pi) + \left[ 2(1-2\pi) \log \frac{1-\pi}{\pi} \right] \cdot \delta + o(\delta), \quad (30)$$

which was also found in the recent [JSS04] via a different route.

In the ‘‘low SNR’’ regime, where  $\delta = \frac{1}{2} - \varepsilon$  and  $\varepsilon \rightarrow 0$ , it was found in [OW04a] that

$$c(\pi) \leq \liminf_{\varepsilon \rightarrow 0} \frac{1 - \bar{H}(Z)}{\varepsilon^4} \leq \limsup_{\varepsilon \rightarrow 0} \frac{1 - \bar{H}(Z)}{\varepsilon^4} \leq C(\pi), \quad (31)$$

with  $c(\cdot), C(\cdot)$  explicitly identified functions. Theorem 2 is shown in [OW04b] to considerably improve both the lower- and upper-bound in (31), though not entirely close the gap.

Refinements of the results of [OW04a, Section 5] using Theorem 2 in additional asymptotic regimes (e.g., ‘‘almost memoryless’’) will be detailed in [OW04b].

## V. NON-SYMMETRIC CASE

We now illustrate the use of the process  $(U_i, V_i)$  (defined via (20) and (21)) in the non-symmetric case. To this end, we confine attention to one particular example: the case where  $\pi_{10} = 1$ , in the ‘‘high SNR’’ regime. We will establish the following:

**Theorem 3** *For  $\pi_{10} = 1$ ,  $0 \leq \pi_{01} < 1$ , and  $\delta$  tending to 0,*

$$\bar{H}(Z) = \bar{H}(X) + \frac{\pi_{01}(2 - \pi_{01})}{1 + \pi_{01}} \delta \log \frac{1}{\delta} + O(\delta). \quad (32)$$

Interestingly, the first term in the expansion is of order  $\delta \log \frac{1}{\delta}$ , in contrast to that in (30) which is of order  $\delta$ . It can actually be shown, [JSS04, OW04a], that the order of  $\delta$  behavior reigns for all values of the pair  $(\pi_{10}, \pi_{01})$ , except when one of the two values equals 1 (in which case Theorem 3 asserts that the order is  $\delta \log \frac{1}{\delta}$ ). This case is left unresolved by the asymptotic expansion of [JSS04] and [ZKD04] which only hints at the above behavior in that the constant multiplying the order  $\delta$  term increases to infinity as either  $\pi_{01}$  or  $\pi_{10}$  tend to one. A variation on the proof of Theorem 3 appearing below is shown in [OW04b] to also recover the expansion of [JSS04] for the case  $\pi_{10} < 1, \pi_{01} < 1$ .

Note that in the case  $\pi_{10} = \pi_{01} = 1$ ,  $\overline{H}(Z) = h_b(\delta)$  while  $\overline{H}(X) = 0$  so  $\overline{H}(Z) = \overline{H}(X) + \delta \log \frac{1}{\delta} + O(\delta)$ , where the factor multiplying the  $\delta \log \frac{1}{\delta}$  term in (32) is  $1/2$  when  $\pi_{01} = 1$ . The reason for this is that just like there is a transition from order of  $\delta$  to order of  $\delta \log 1/\delta$  when going from  $\pi_{10} < 1$  to  $\pi_{10} = 1$ , a similar term that will be order of  $\delta$  in our analysis below (where we assume  $\pi_{01} < 1$ ) becomes order of  $\delta \log 1/\delta$  when going from  $\pi_{01} < 1$  to  $\pi_{01} = 1$ . This accounts for the doubling of the said factor from  $1/2$  to  $1$ .

Throughout the remainder of this section we assume that  $\pi_{10} = 1$  and  $\pi_{01} < 1$ , in which case  $h(x)$  simplifies to

$$h(x) = \log \frac{\pi_{01}}{\pi_{01} + e^x},$$

where  $\bar{x}$  denotes  $1 - x$ . Note that  $h(x)$  is decreasing in  $x$  and is upper bounded by  $h(-\infty) = \log \pi_{01}/\overline{\pi_{01}}$ . Define

$$f(x) = \log \frac{1-x}{x}.$$

Thus the upper bound on  $h(x)$  just stated is  $-f(\pi_{01})$ .

In the spirit of the developments in the previous subsection for bounding the support in the case  $M = 2$ , considering the alternative process constructed in Subsection A.2 of the previous section, we will show that the support of  $P_U = P(l_i|X_i = 0)$  (and  $P_Y = P(l_i|X_i = 1)$ , as they have identical supports) is contained in the union of four disjoint intervals on the real line whose boundary points and probabilities (under  $P_U$  and  $P_Y$ ) we characterize explicitly. We will then obtain upper and lower bounds on the entropy rate of  $\{Z_i\}$  in terms of the interval boundary points and probabilities, similarly as was done in the derivation of Theorem 2. The bounds thus obtained will be shown to lead to the asymptotic behavior of the entropy rate stated in Theorem 3.

The following lemma, which follows from elementary calculus, will be used throughout our analysis.

**Lemma 2** *Suppose  $p = p_0 + \delta p_1 + O(\delta^2)$ . If  $0 < p_0 \overline{\pi_{10}} + \overline{p_0} \pi_{01} < 1$  then*

$$\begin{aligned} h_b([p\overline{\pi_{10}} + \overline{p}\pi_{01}] * \delta) &= h_b(p_0\overline{\pi_{10}} + \overline{p_0}\pi_{01}) \\ &\quad - \delta [(p_0(1 - 2\pi_{01}) + p_0(2\pi_{10} - 1) \\ &\quad + p_1(1 - \pi_{01} - \pi_{10})) \log \frac{p_0\overline{\pi_{10}} + \overline{p_0}\pi_{01}}{p_0\pi_{10} + \overline{p_0}\pi_{01}}] + O(\delta^2). \end{aligned} \quad (33)$$

If  $\pi_{10} = p_0 = 1$  and  $\pi_{01} < 1$  then

$$h_b([p\overline{\pi_{10}} + \overline{p}\pi_{01}] * \delta) = (1 - p_1\pi_{01})\delta \log \frac{1}{\delta} + O(\delta). \quad (34)$$

Define now the following four intervals on the real line, where an interval  $[a, b]$  is taken to be empty if  $a > b$ .

$$\begin{aligned} I_0 &= [-f(\delta) + h(f(\delta) - f(\pi_{01})), \\ &\quad -f(\delta) + h(f(\delta) + h(f(\delta) - f(\pi_{01})))], \\ I_1 &= [-f(\delta) + h(-f(\delta) - f(\pi_{01})), -f(\delta) - f(\pi_{01})] \\ I_2 &= [f(\delta) + h(f(\delta) - f(\pi_{01})), \\ &\quad f(\delta) + h(f(\delta) + h(f(\delta) - f(\pi_{01})))], \\ I_3 &= [f(\delta) + h(-f(\delta) - f(\pi_{01})), f(\delta) - f(\pi_{01})] \end{aligned} \quad (35)$$

We shall also rely on the following three lemmas. Their proofs, which we omit, are based on ideas similar to those used in the proof of Lemma 1.

**Lemma 3** *The intervals  $I_j$ ,  $j = 0, 1, 2, 3$  are non-empty (i.e. the left end points as specified above are smaller than the right end points).*

Given two intervals  $I$  and  $J$  let  $I < J$  express the fact that the right end point of  $I$  is (strictly) less than the left end point of  $J$ .

**Lemma 4** *For all sufficiently small  $\delta > 0$  the intervals  $I_j$ ,  $j = 0, 1, 2, 3$  satisfy  $I_0 < I_1 < I_2 < I_3$ .*

**Lemma 5** *The supports of both  $P_U$  and  $P_Y$  are contained in  $I_0 \cup I_1 \cup I_2 \cup I_3$ . For all sufficiently small  $\delta > 0$ , the probabilities of the intervals under  $P_U$  and  $P_Y$  are given by*

$I$	$P_Y(I)$	$P_U(I)$
$I_0$	$\delta^2$	$\overline{\delta}(\pi_{01} * \delta)$
$I_1$	$\delta \overline{\delta}$	$\overline{\delta}(\pi_{01} * \delta)$
$I_2$	$\delta \overline{\delta}$	$\delta(\pi_{01} * \delta)$
$I_3$	$\overline{\delta}^2$	$\delta(\pi_{01} * \overline{\delta})$

(36)

For any closed interval  $I$  on the real line let  $\ell(I)$  denote the smallest value in  $I$  (left end point) and  $u(I)$  denote the largest value (right end point). Let  $k_0 = 1/(1 + \pi_{01})$  and  $k_1 = \pi_{01}/(1 + \pi_{01})$  respectively denote the stationary probabilities of  $X_i = 0$  and  $X_i = 1$ . Also define

$$\beta(x) = \frac{e^x}{1 + e^x},$$

which maps  $x = \log Pr(1)/Pr(0)$  to  $Pr(1)$  (e.g. log-likelihood ratios to probabilities).

**Lemma 6** *For all sufficiently small  $\delta > 0$ , the entropy rate  $\overline{H}(Z)$  of the process  $\{Z_i\}$  satisfies*

$$\overline{H}(Z) \leq \sum_{j=0}^3 [k_0 P_U(I_j) + k_1 P_Y(I_j)] \max_{x \in I_j} h_b([\beta(x)\pi_{01}] * \delta) \quad (37)$$

and

$$\overline{H}(Z) \geq \sum_{j=0}^3 [k_0 P_U(I_j) + k_1 P_Y(I_j)] \min_{x \in I_j} h_b([\beta(x)\pi_{01}] * \delta). \quad (38)$$

**Proof:** That

$$\begin{aligned} \overline{H}(Z) &\geq \sum_{j=0}^3 [k_0 Pr(l_i \in I_j | X_i = 0) \\ &\quad + k_1 Pr(l_i \in I_j | X_i = 1)] \min_{x \in I_j} h_b([\beta(x)\pi_{01}] * \delta) \end{aligned}$$

follows from (22) and Lemma 5, once  $\delta$  is sufficiently small for Lemma 4 to imply that the  $I_j$  are disjoint. The upper bound follows similarly.  $\square$

**Proof of Theorem 3:** Lemma 5 shows that all but  $P_Y(I_3)$ ,  $P_U(I_0)$ , and  $P_U(I_1)$  are  $O(\delta)$ . Therefore, since  $h_b(\cdot)$  is bounded, the only terms in (37) and (38) that might be greater than  $O(\delta)$  are those involving  $P_Y(I_3)$ ,  $P_U(I_0)$ , and  $P_U(I_1)$ .

First we consider the terms involving  $P_U(I_0)$ , and  $P_U(I_1)$ . It follows from elementary calculus that  $h_b([\overline{p}\pi_{01}] * \delta)$  is maximized at  $\max\{0, (\pi_{01} - 1/2)/(\delta + \pi_{01})\}$ . This fact together with the concavity of  $h_b([\overline{p}\pi_{01}] * \delta)$  in  $p$ , and the fact that both end points of both  $I_0$  and  $I_1$  are tending to  $-\infty$  imply that

for all sufficiently small  $\delta > 0$ ,  $\min_{x \in I_j} h_b([\overline{\beta(x)}\pi_{01}] * \delta)$  and  $\max_{x \in I_j} h_b([\overline{\beta(x)}\pi_{01}] * \delta)$  are achieved either at  $\ell(I_j)$  or  $u(I_j)$  for  $j = 0, 1$ . It is not difficult to see that for  $j = 0, 1$  both  $\beta(\ell(I_j))$  and  $\beta(u(I_j))$  are ratios of polynomials in  $\delta$ . In particular, they will be of the form  $p_0 + \delta p_1 + O(\delta^2)$  with  $p_0 = 0$ . Therefore, using Lemmas 2 and 5

$$\begin{aligned} & \sum_{j=0}^1 k_0 P_U(I_j) \max_{x \in I_j} h_b([\overline{\beta(x)}\pi_{01}] * \delta) \\ &= k_0 \pi_{01} h_b(\pi_{01}) + k_0 \overline{\pi_{01}} h_b(\pi_{01}) + O(\delta) \\ &= \overline{H}(X) + O(\delta). \end{aligned} \quad (39)$$

Similarly,

$$\sum_{j=0}^1 k_0 P_U(I_j) \min_{x \in I_j} h_b([\overline{\beta(x)}\pi_{01}] * \delta) = \overline{H}(X) + O(\delta). \quad (40)$$

Next we focus on the terms involving  $P_Y(I_3)$ . In these cases, the above properties (concavity and extremal) of  $h_b([\overline{p}\pi_{01}] * \delta)$  viewed as a function of  $p$ , and the fact that the left end point of  $I_3$  is greater than the maximizing  $p$ , imply that

$$\min_{x \in I_3} h_b([\overline{\beta(x)}\pi_{01}] * \delta) = h_b([\overline{\beta(u(I_3))}\pi_{01}] * \delta),$$

and

$$\max_{x \in I_3} h_b([\overline{\beta(x)}\pi_{01}] * \delta) = h_b([\overline{\beta(\ell(I_3))}\pi_{01}] * \delta).$$

From (35) we see (some algebraic manipulations omitted) that

$$\beta(u(I_3)) = \frac{\overline{\delta}\pi_{01}}{\delta\overline{\pi_{01}} + \overline{\delta}\pi_{01}} \quad (41)$$

$$= 1 - \frac{\delta\overline{\pi_{01}}}{\delta\overline{\pi_{01}} + \overline{\delta}\pi_{01}} \quad (42)$$

$$= 1 - \delta \frac{\overline{\pi_{01}}}{\pi_{01}} + O(\delta^2), \quad (43)$$

and

$$\beta(\ell(I_3)) = \frac{\overline{\delta}^2 \pi_{01} \overline{\pi_{01}}}{\pi_{01}^2 \delta^2 + \overline{\delta} \delta \overline{\pi_{01}}^2 + \overline{\delta}^2 \pi_{01} \overline{\pi_{01}}} \quad (44)$$

$$= 1 - \frac{\pi_{01}^2 \delta^2 + \overline{\delta} \delta \overline{\pi_{01}}^2}{\pi_{01}^2 \delta^2 + \overline{\delta} \delta \overline{\pi_{01}}^2 + \overline{\delta}^2 \pi_{01} \overline{\pi_{01}}} \quad (45)$$

$$= 1 - \delta \frac{\overline{\pi_{01}}}{\pi_{01}} + O(\delta^2). \quad (46)$$

Lemma 2, (43), and (46), then imply that

$$\min_{x \in I_3} h_b([\overline{\beta(x)}\pi_{01}] * \delta) = (1 + \overline{\pi_{01}}) \delta \log \frac{1}{\delta} + O(\delta), \quad (47)$$

and

$$\max_{x \in I_3} h_b([\overline{\beta(x)}\pi_{01}] * \delta) = (1 + \overline{\pi_{01}}) \delta \log \frac{1}{\delta} + O(\delta). \quad (48)$$

Equations (39), (40), (47), (48), and the expression for  $P_Y(I_3)$  from (36) demonstrate that the combined contributions of the terms involving  $P_Y(I_3)$ ,  $P_U(I_0)$ , and  $P_U(I_1)$  to (37) and (38) is

$$\begin{aligned} & \overline{H}(X) + k_1(1 + \overline{\pi_{01}}) \delta \log \frac{1}{\delta} + O(\delta) \\ &= \overline{H}(X) + \frac{\pi_{01}(2 - \pi_{01})}{1 + \pi_{01}} \delta \log \frac{1}{\delta} + O(\delta) \end{aligned}$$

in both cases. The theorem is proved since, as noted above, all the other terms are  $O(\delta)$ .  $\square$

## VI. CONCLUSIONS

We have presented an approach to approximating the entropy rate of a hidden Markov process via approximations of the stationary distribution of a related Markov process. In its crudest form, involving only bounds on the support of this distribution, this approach was seen in [OW04a] to lead to new bounds on the entropy rate for a BSC-corrupted binary Markov chain. Here we have presented the approach in more generality, both for non-binary alphabets and for more refined approximations of the said distribution. We then illustrated how it can be applied for characterizing the behavior of the entropy rate in some asymptotic regimes. It was seen that a slight refinement of the bounding technique in [OW04a] which considered only the support, to a partition of the support into a small number of non-overlapping regions with easily computed probabilities, can lead to significantly tighter bounds and finer characterizations of the asymptotics. The bounds can be further tightened by further refining this partition leading, in some asymptotic regimes, to characterization of higher-order terms [OW04b].

## REFERENCES

- [Bir62] J. J. Birch. Approximations for the Entropy for Functions of Markov Chains *Ann. Math. Stat.*, vol. 33, pp. 930-938, 1962.
- [Bla57] D. Blackwell. The entropy of functions of finite-state markov chains. *Trans. First Prague Conf. Inf. Th., Statistical Decision Functions, Random Processes*, pages 13–20, 1957.
- [CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
- [EBTBH04] S. Egner, V. B. Balakirsky, L. M. G. M. Tolhuizen, S. P. M. J. Baggen, and H. D. L. Hollmann. On the Entropy Rate of a Hidden Markov Model. *Int. Symp. Inf. Th.*, p. 12, Chicago, IL, June-July 2004.
- [EM02] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theory*, vol. IT-48, no. 6, pp. 1518-1569, June 2002.
- [HJ99] B. M. Hochwald and P. R. Jelenković. State Learning and Mixing in Entropy of Hidden Markov Processes and the Gilbert–Elliott Channel. *IEEE Trans. Inform. Theory*, vol. IT-45, no. 1, pp. 128-138, January 1999.
- [HGG03] T. Holliday, P. Glynn, and A. Goldsmith. Capacity of Finite State Markov Channels with General Inputs. *Int. Symp. Inf. Th.*, p. 289, Yokohama, Japan, June-July 2003.
- [JSS04] P. Jacquet, G. Seroussi, and W. Szpankowski. On the Entropy of a Hidden Markov Process. *Int. Symp. Inf. Th.*, p. 10, Chicago, IL, June-July 2004.
- [MBD89] M. Mushkin and I. Bar-David. Capacity and Coding for the Gilbert-Elliott Channel. *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 1277-1290, November 1989.
- [OW04a] E. Ordentlich and T. Weissman. On the Optimality of Symbol by Symbol Filtering and Denoising. Submitted to *IEEE Trans. Inform. Theory*. See also *Int. Symp. Inf. Th.*, p. 200, Chicago, IL, June-July 2004.
- [OW04b] E. Ordentlich and T. Weissman. First-Order Asymptotics for the Entropy Rate of Hidden Markov Processes. *In preparation*.
- [ZKD04] O. Zuk, I. Kanter, and E. Domany. Asymptotics of the Entropy Rate for a Hidden Markov Process. Submitted to SSPIT, 2004.