Approximations for the Entropy Rate of a Hidden Markov Process

Erik Ordentlich* and Tsachy Weissman†[1],
*Hewlett-Packard Laboratories, Palo Alto, CA 94304, U.S.A., eord@hpl.hp.com
†Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA, 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. We present an approach to bounding the entropy rate of $\{Z_t\}$ by the construction and study of a related measure-valued Markov process. To illustrate its efficacy, we specialize it to the case of a BSC-corrupted binary Markov chain. The bounds obtained are sufficiently tight to characterize the behavior of the entropy rate in asymptotic regimes that exhibit a "concentration of the support". Examples include the 'high SNR', 'low SNR', 'rare spikes', and 'weak dependence' regimes. Our analysis also gives rise to a deterministic algorithm for approximating the entropy rate, achieving the best known precision-complexity tradeoff, for a significant subset of the process parameter space.**

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, i.e., the $|\mathcal{X}| \times |\mathcal{X}|$ matrix with entries $\mathcal{K}(x,x') = P(X_{t+1} = x'|X_t = x)$. $\mathcal{C}$ will denote the channel transition matrix, i.e., the $|\mathcal{X}| \times |\mathcal{Z}|$ matrix with entries $\mathcal{C}(x,z) = P(Z_t = z|X_t = x)$. $\{Z_t\}$ is known as a Hidden Markov Process (HMP). 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, and is our interest in this work.

Hidden Markov Processes (HMPs) arise naturally in many contexts, both as information sources and as noise (cf. [6] and references therein). Key questions in lossless and lossy compression [9], [10], and in channel coding [5], [16], [13], reduce to finding the entropy rate $\overline{H}(Z)$.

Let $\mathcal{M}(\mathcal{X})$ denote the simplex of distributions on $\mathcal{X}$ and β_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 β_t. We refer to $\{\beta_t\}$ as the "belief process", as it represents the "belief" of an observer of the HMP regarding the value of the underlying state. Conditional independence of X_{t+1} and $Z_{-\infty}^t$ given X_t implies that $P(X_{t+1} \in \cdot|Z_{-\infty}^t) = \beta_t \cdot \mathcal{K}$, in turn implying, by the memorylessness of the noise, $P(Z_{t+1} \in \cdot|Z_{-\infty}^t) = \beta_t \cdot \mathcal{K} \cdot \mathcal{C}$. With $H(Q)$ denoting the entropy of a distribution Q on $\mathcal{Z}$, we obtain

$$\overline{H}(Z) = H(Z_{t+1}|Z_{-\infty}^t) = EH\left(\beta_t \cdot \mathcal{K} \cdot \mathcal{C}\right). \tag{1}$$

Evidently, the distribution of β_t holds the key to the value of the entropy rate. This distribution, however, shown by Blackwell in [3] (cf. also [18, Claim 1]) to satisfy an integral equation, remains elusive to date even for the simplest HMPs. Additional perspectives by which the hardness of the problem can be appreciated pertain to Lyapunov exponents [1], [21], [12], [13] and to statistical physics [24], [15], [22].

Given the hardness of the problem, the predominant approach to the study of the entropy rate has been one of approximation, via both deterministic bounds [4, Section 4.4], and Monte Carlo simulation (cf. [12] and references therein).

Useful as these techniques may be from a numerical standpoint, they lack the capacity to resolve basic questions regarding the dependence of the entropy rate on the Markov transition kernel and the channel parameters. First steps towards the resolution of such questions have only recently been taken: Continuity of the entropy rate in the parameters was established by [12]. Expansions of the entropy rate for the BSC-corrupted binary Markov chain in the "high SNR" regime, where the channel crossover probability is small, have been obtained in [14], [18], [24], [19]. Initial results on the behavior in various additional asymptotic regimes such as "rare-spikes", "rare-bursts", "low SNR", and "almost memoryless", were obtained in [18]. In this submission we present tighter bounds and finer characterizations of the entropy rate, by refining the idea behind the bounds of [18].

The gist of our approach is the following, which is an immediate consequence of (1):

Observation 1:

$$\min_{\beta \in \mathcal{S}} H\left(\beta \cdot \mathcal{K} \cdot \mathcal{C}\right) \le \overline{H}(Z) \le \max_{\beta \in \mathcal{S}} H\left(\beta \cdot \mathcal{K} \cdot \mathcal{C}\right),$$

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

Trivial as this observation may seem, it was seen in [18] to lead to useful bounds in cases where bounds on the support set $\mathcal{S}$ are obtainable, and this set is significantly smaller than the whole simplex $\mathcal{M}(\mathcal{X})$, a situation referred to as "concentration of the support". We now note that the bounds of Observation (1), which depend on the distribution of β_t through its support only, can be refined by partitioning $\mathcal{S}$ into subsets:

[1] This author is also with Hewlett–Packard Laboratories, Palo Alto, CA 94304, U.S.A.

Observation 2: For any countable collection $\{I_i\}$ of pairwise disjoint sets $I_i \subseteq \mathcal{M}(\mathcal{X})$ covering $\mathcal{S}$

$$\sum_i P(\beta_t \in I_i) \inf_{\beta \in I_i} H(\beta \cdot \mathcal{K} \cdot \mathcal{C}) \leq \overline{H}(Z) \leq$$

$$\sum_i P(\beta_t \in I_i) \sup_{\beta \in I_i} H(\beta \cdot \mathcal{K} \cdot \mathcal{C}). \quad (2)$$

Since the distribution of β_t is unknown, $P(\beta_t \in I_i)$ 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 process is directly and simply related to the distribution of β_t. The fraction of times that this new process visits the set I_i, for appropriately chosen I_i, is computable, a computation that can then be directly translated to give the value of $P(\beta_t \in I_i)$. Thus, in a nutshell, the two new ingredients that our approach involves are the bounds on the support of the belief process, and the construction of an alternative more tractable Markov process as a tool for analyzing the former.

For brevity in illustrating our main ideas, we restrict attention in this extended abstract to the case where $\{Z_t\}$ is a BSC-corrupted binary symmetric Markov chain. The more general case, as well as all details and proofs omitted below, are given in [20].

In Section II we start with a concrete description of the problem setting, and the evolution of the log-likelihood process (equivalent to the belief process but in a more convenient form). We then detail the construction of an alternative Markov process, and its relationship to the original log-likelihood process. Section III gives some details regarding the form the bounds in (2) assume in terms of the alternative Markov process. Using these bounds, we then derive the behavior of the entropy rate in various asymptotic regimes. In Section IV we describe a deterministic algorithm, inspired by the alternative Markov process, for approximating the entropy rate. We show that its guaranteed precision-complexity tradeoff is the best among the known deterministic schemes for approximation of the entropy rate, for a significant subset of the process parameter space. We close in Section V with some concluding remarks.

II. The Alternative Markov Process

Assume henceforth the case $\mathcal{X} = \mathcal{Z} = \{0, 1\}$, where the Markov transition matrix and the channel matrix are, respectively,

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

There is no loss of generality in assuming $\pi < 1/2$ since the argument in [18, 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 π. Defining $l_i = \log \frac{\beta_i(1)}{1-\beta_i(1)}$ as the log-likelihood process, the standard forward recursions [6] are readily verified to assume the form

$$l_i = (2Z_i - 1) \log \left[\frac{1-\delta}{\delta} \right] + f(l_{i-1}), \quad (4)$$

where $f(x) = \log \frac{e^x(1-\pi)+\pi}{e^x\pi+(1-\pi)}$. Note that f is contractive since

$$\sup_x |f'(x)| = f'(0) = 1 - 2\pi < 1, \quad (5)$$

a property that plays a key role in proofs of results such as Theorem 1 and Theorem 6 below.

In terms of the log-likelihood process, (1) in this setting becomes

$$\overline{H}(Z) = E h_b \left(\frac{e^{l_i}}{1 + e^{l_i}} * \pi * \delta \right), \quad (6)$$

where h_b and $*$ denote, respectively, the binary entropy function and binary convolution.

We now construct a Markov process which, as a process, is more tractable than the log-likelihood process $\{l_i\}$, but whose stationary distribution is closely and simply related to that of l_i. The benefit is that the entropy rate, which was expressed as the expectation in (6), will be expressible as a similar expectation involving the new process.

Theorem 1: Consider the 1st-order Markov process $\{Y_i\}_{i \geq 0}$ formed by letting $Y_0 = Y$, and $\{Y_i\}_{i \geq 1}$ evolve according to

$$Y_i = r_i \log \frac{1-\delta}{\delta} + s_i f(Y_{i-1}), \quad (7)$$

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

$$r_i = \begin{cases} -1 & \text{w.p. } \delta \\ 1 & \text{w.p. } 1-\delta, \end{cases} \quad s_i = \begin{cases} -1 & \text{w.p. } \pi \\ 1 & \text{w.p. } 1-\pi. \end{cases} \quad (8)$$

Then:

1) *[Existence and uniqueness of stationary distribution:]* There exists a unique (in distribution) random variable Y under which $\{Y_i\}_{i \geq 0}$ is stationary.
2) *[Connection to the original process:]* $\mathcal{L}(Y) = \mathcal{L}(l_i | X_i = 1)$, where $\mathcal{L}$ denotes the probability law.

The connection established in Theorem 1 between the HMP and the process $\{Y_i\}$, when combined with (6), can be shown to yield:

Theorem 2: For the process constructed in Theorem 1

$$\overline{H}(Z) = E h_b \left(\frac{e^{Y_i}}{1 + e^{Y_i}} * \pi * \delta \right). \quad (9)$$

The bottom line is that we have transformed the calculation of the entropy rate into an expectation of a simple function of the variable Y_i. It will be seen that the benefit in doing that is that information on the distribution of Y_i, which translates via Theorem 2 to bounds on the entropy rate, can be inferred by studying the dynamics of the process $\{Y_i\}$.

III. Bounds on the Entropy Rate

Theorem 2 can now be used to bound the entropy rate, by bounding the expectation on the right side of (9). The following gives the general form of the bounds one obtains in this way.

Theorem 3: Let $\{Y_i\}$ be the stationary Markov process whose evolution is given by (7). 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 the following are, respectively, lower and upper bounds on $\overline{H}(Z)$

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

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

Evidently, a bound of the type in Theorem 3 would be applicable only in situations where: 1) the support of Y_i is contained in a set of the form $\bigcup_{k=1}^M [a_k, b_k] \cup \bigcup_{k=1}^M [-b_k, -a_k]$ and 2) the probabilities $P\left(Y_i \in [-b_k, -a_k] \cup [a_k, b_k]\right)$ can be computed (or bounded from above and below). To get an appreciation for when this can happen, it is instructive to consider first the case $M = 1$, for which Theorem 3 yields

Corollary 1: Let $\{Y_i\}$ be the process in (7). Let $0 \leq b \leq A$ be such that $[-A, -b] \cup [b, A]$ contains the support of Y_i. Then

$$h_b\left(\frac{e^A}{1+e^A} * \pi * \delta\right) \leq \overline{H}(Z) \leq h_b\left(\frac{e^b}{1+e^b} * \pi * \delta\right). \tag{10}$$

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 (7), to be the solution to the equation $A = f(A) + \log \frac{1-\delta}{\delta}$, whose explicit form we omit. The obvious symmetry of the support of Y_i around 0 implies that $-A$ is the lower endpoint of the support of Y_i. In particular, this establishes that the support of Y_i is contained in the interval $[-A, A]$. 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. The value of this lower endpoint can similarly be read from the dynamics of the process in (7). By symmetry, $-b$ is the upper endpoint of the support of Y_i in the negative half. This implies then that the support of Y_i is contained in $[-A, -b] \cup [A, b]$, and that we can easily explicitly compute the smallest and largest values of A and b with this property. Crude as the bound of Corollary 1 may seem, it was shown in [18] to convey non-trivial information (when optimizing over the values of A and b).

Taking one step of refinement beyond Corollary 1, when specialized to the case $M = 2$, Theorem 3 can be shown to yield:

Corollary 2: For all $\delta \leq \frac{1}{2}\left(1 - \sqrt{\max\{1-4\pi, 0\}}\right)$, $\overline{H}(Z)$ is lower and upper bounded, respectively, by

$$\{(1-\delta)[\pi*(1-\delta)] + \delta[\pi*\delta]\}\, h_b\left(\frac{e^A}{1+e^A} * \pi * \delta\right) +$$
$$\{(1-\delta)[\pi*\delta] + \delta[\pi*(1-\delta)]\}\, h_b\left(\frac{e^a}{1+e^a} * \pi * \delta\right),$$
$$\{(1-\delta)[\pi*(1-\delta)] + \delta[\pi*\delta]\}\, h_b\left(\frac{e^B}{1+e^B} * \pi * \delta\right)$$
$$+ \{(1-\delta)[\pi*\delta] + \delta[\pi*(1-\delta)]\}\, h_b\left(\frac{e^b}{1+e^b} * \pi * \delta\right),$$

for all values of $b \leq a \leq B \leq A$ such that $[b, a] \cup [B, A]$ contains the support of $|Y_i|$.

The optimal values of (b, a, B, A) (smallest A, a and largest B, b) can be obtained easily by observing the dynamics of $\{Y_i\}$ in (7), similarly as detailed in the context of Corollary 1. In fact, the optimum values of b and A are exactly those obtained for Corollary 1, so when moving to the approximation of order $M = 2$ from $M = 1$, it is only the new points a, B that need be computed.

As can be expected, the bounds in Corollary 2 are considerably tighter, in various asymptotic regimes, than those based on Corollary 1. As a first example, in the "high SNR" regime the analysis in [18, Section 5], which was based on Corollary 1, established $\overline{H}(Z) - h_b(\pi) = \Theta(\delta)$, while the bounds of Corollary 2 give, for $\pi \leq 1/2$ and $\delta \downarrow 0$,

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

a result first proved in [14], and subsequently derived in [19] and [24].

For the "almost memoryless" regime, the bounds of Corollary 2 lead to the following:

Theorem 4: For $0 \leq \delta \leq 1/2$, and $\pi = 1/2 - \varepsilon$, as $\varepsilon \downarrow 0$

$$1 - \overline{H}(Z) = \frac{2}{\ln 2}\varepsilon^2(1-2\delta)^4 + o(\varepsilon^3).$$

This is also a refinement of a result in [18, Section 5], which was based on Corollary 1. As a last example, in the "low SNR" regime, Corollary 2 leads to the following:

Theorem 5: For $1/4 \leq \pi \leq 1/2$, and $\delta = \frac{1}{2} - \varepsilon$,

$$c(\pi) \leq \liminf_{\varepsilon\to 0} \frac{1-\overline{H}(Z)}{\varepsilon^4} \leq \limsup_{\varepsilon\to 0} \frac{1-\overline{H}(Z)}{\varepsilon^4} \leq C(\pi), \tag{11}$$

where the constants $c(\pi)$ and $C(\pi)$ are explicitly identified as functions of π. In particular, both $c(\pi)$ and $C(\pi)$ behave as $\sim \frac{8}{\ln 2}(1-2\pi)^2$ for $\pi \to 1/2$, implying that $1 - \overline{H}(Z) \approx \frac{32}{\ln 2}(1/2-\delta)^4(1/2-\pi)^2$ for π and δ close to $1/2$.

We mention in this context that there are regimes in which the Cover and Thomas bounds [4, Section 4.4], of any order, will not capture the behavior of the entropy rate. For a simple example note that, for any n,

$$H(Z_0|Z_{-n+1}^{-1}, X_{-n}) \leq H(Z_0|X_{-n}) = h_b(\pi^{*n} * \delta), \tag{12}$$

where π^{*n} denotes binary convolution of π with itself n times. Thus, for example, in the "low SNR" regime where

π is fixed and $\delta = 1/2 - \varepsilon$, $H(Z_0|Z_{-n+1}^{-1}, X_{-n}) \leq h_b(\pi^{*n} * \delta) = h_b(1/2 - \varepsilon(1 - 2\pi^{*n}))$ and, in particular, $1 - H(Z_0|Z_{-n+1}^{-1}, X_{-n}) = \Omega(\varepsilon^2)$. In other words, using $H(Z_0|Z_{-n+1}^{-1}, X_{-n})$ to lower bound the entropy rate will give an upper bound on the left side of (11) of order ε^2, failing to provide the true ε^4 order established in Theorem 5.

The results above were given as examples for the kind of results that are obtained via evaluation of the bounds of Corollary 2 in the respective regimes. Corollary 2 is nothing but a specialization of Theorem 3 to the case $M = 2$, optimizing over the choice of constants a_k and b_k. Moving from the bounds corresponding to $M = 1$ to those of $M = 2$ results, for various asymptotic regimes, in characterization of higher order terms, and refinement of constants. The larger M one takes, the finer will the bounds become, leading also to finer characterizations in the various asymptotic regimes. The development we have detailed for the case $M = 2$ scales to any larger value of M. For example, $M = 3$ will correspond to an outer bound on the support of Y_i obtained by excluding a subinterval from each of the four intervals associated with Corollary 2 (namely $[b, a]$, $[B, A]$, $[-a, -b]$, $[-B, -A]$). The choice of these subintervals will be optimized analogously as for Corollary 2 quite simply, via the dynamics of the process $\{Y_i\}$ in (7). Note that it is only the endpoints of the new subintervals that need be computed, the remaining endpoints being identical to those evaluated for $M = 2$. More generally, moving from the approximation corresponding to a value of M to the value $M+1$ corresponds to discarding a subinterval from each of the intervals constituting the outer bound of the support obtained at the M-th level. Only the endpoints of the subintervals that are being discarded need be computed, the remaining ones coinciding with those already obtained in the previous stage.

IV. A Deterministic Approximation Algorithm

In this section we propose a new entropy rate approximation scheme, which is based on approximating the stationary distribution of the alternative Markov process constructed in Section II. Throughout 'operations' refers to arithmetic operations.

Since $|f| \leq \log(1 - \pi)/\pi$, from (7) it is clear that the support of Y_i is contained in the interval $\left[-\log\frac{(1-\pi)(1-\delta)}{\pi\delta}, \log\frac{(1-\pi)(1-\delta)}{\pi\delta}\right]$. Let Q be an M-level quantizer with the property that

$$\max_{x\in\left[-\log\frac{(1-\pi)(1-\delta)}{\pi\delta}, \log\frac{(1-\pi)(1-\delta)}{\pi\delta}\right]} |Q(x) - x| \leq \varepsilon. \tag{13}$$

For example, a uniform quantizer with $M \geq \frac{1}{\varepsilon}\log\frac{(1-\pi)(1-\delta)}{\pi\delta}$ levels has this property. Consider now the finite-state Markov process (with M states) evolving with the process in (7) according to

$$\tilde{Y}_i = Q\left(r_i \log\frac{1-\delta}{\delta} + s_i f(\tilde{Y}_{i-1})\right), \tag{14}$$

initialized at time $i = 0$ jointly with Y_0 such that the joint process $\{(Y_i, \tilde{Y}_i)\}$ is stationary. Then

$$\begin{aligned}
|\tilde{Y}_i - Y_i| &\leq \varepsilon + |f(\tilde{Y}_{i-1}) - f(Y_{i-1})| \\
&\leq \varepsilon + (1 - 2\pi)|\tilde{Y}_{i-1} - Y_{i-1}| \qquad (15)\\
&\leq \varepsilon + (1 - 2\pi)[\varepsilon + |f(\tilde{Y}_{i-2}) - f(Y_{i-2})|] \\
&\vdots \\
&\leq \varepsilon\sum_{j=0}^{i}(1 - 2\pi)^j + (1 - 2\pi)^i|\tilde{Y}_0 - Y_0| \\
&\leq \frac{\varepsilon}{2\pi} + (1 - 2\pi)^i|\tilde{Y}_0 - Y_0|, \qquad (16)
\end{aligned}$$

where (15) follows from (5). Since this is true for all i, stationarity implies

$$E|Y_i - \tilde{Y}_i| \leq \frac{\varepsilon}{2\pi}. \tag{17}$$

This motivates the following approximation algorithm:

Algorithm 4.1:

Input: M, π, δ

1) Let Q denote the M-level uniform quantizer of the interval $\left[-\log\frac{(1-\pi)(1-\delta)}{\pi\delta}, \log\frac{(1-\pi)(1-\delta)}{\pi\delta}\right]$ and $q_1, \ldots q_M$ denote the quantization levels. Let P_M be the $M \times M$ stochastic matrix whose (i, j)-th entry is given as

$$\begin{cases} (1-\delta)(1-\pi) & \text{if } q_j = Q\left(\log\frac{1-\delta}{\delta} + f(q_i)\right) \\ \delta(1-\pi) & \text{if } q_j = Q\left(-\log\frac{1-\delta}{\delta} + f(q_i)\right) \\ (1-\delta)\pi & \text{if } q_j = Q\left(\log\frac{1-\delta}{\delta} - f(q_i)\right) \\ \delta\pi & \text{if } q_j = Q\left(-\log\frac{1-\delta}{\delta} - f(q_i)\right) \\ 0 & \text{otherwise .} \end{cases} \tag{18}$$

2) Compute stationary distribution of P_M, i.e., the M-dimensional row vector $\mathbf{a}_M$ solving $\mathbf{a}_M \cdot P_M = \mathbf{a}_M$.
3) Compute entropy estimate

$$\hat{H} = \sum_{i=1}^{M} \mathbf{a}_M(i) \cdot h_b\left(\frac{e^{q_i}}{1 + e^{q_i}} * \pi * \delta\right). \tag{19}$$

Output: $\hat{H}$.

Note that $\hat{H}$ in (19) is nothing but the expression $Eh_b\left(\frac{e^{\tilde{Y}_i}}{1+e^{\tilde{Y}_i}} * \pi * \delta\right)$, where $\{\tilde{Y}_i\}$ is the quantized process defined in (14) (initialized at its stationary distribution). One brute force method [8] for finding the stationary distribution of an $M \times M$ stochastic matrix is via Gaussian elimination ($2M^3/3$ operations), back substitution (M^2 operations), and normalization (M operations). Since the remaining steps in the algorithm require $O(M)$ operations, the overall number of operations required is $O(M^3)$. The bound in (17) and the fact that $h_b(\varepsilon) \sim \varepsilon\log(1/\varepsilon)$ can be seen to imply that the resulting precision is $O\left(\frac{\log M}{M}\right)$. In summary, we have established the following:

Theorem 6: For fixed π, δ, Algorithm 4.1 requires $O(M^3)$ operations and guarantees precision of $O\left(\frac{\log M}{M}\right)$. In other words, N operations buy precision $O\left(N^{-1/3}\log N\right)$.

Theorem 6 was derived via a rather rough analysis. Two ingredients that may significantly improve the bound on the

approximation-precision tradeoff are: 1) Using a non-uniform quantizer, with finer resolution near 0 (where f is least contractive) and coarser resolution towards the endpoints of the quantized interval (where f is highly contractive). 2) Capitalizing on the special structure of P_M, a very sparse matrix with the same 4 non-zero entries in each row, to simplify the scheme for obtaining its stationary distribution.

Theorem 6 as is, however, suffices to make our main point, which is the improved dependence of the bound on the precision order on the process parameters (in this case π and δ), relative to the best known precision-complexity tradeoff among deterministic approximation schemes that are obtained via the Cover and Thomas bounds. Specifically, the difference between the upper and lower bounds in the Cover and Thomas approximation, conveniently expressed as the mutual information $I(Z_0; X_{-n-1}|Z_{-n}^{-1})$, is known since [2] to decay exponentially with n. The best known bounds have been obtained in [11] and are in the form

$$I(Z_0; X_{-n-1}|Z_{-n}^{-1}) \leq C(\pi,\delta)\rho(\pi,\delta)^n,$$

where $C(\pi,\delta), \rho(\pi,\delta)$ are positive constants with $\rho(\pi,\delta) < 1$. On the other hand, the number of operations required to compute the Cover and Thomas bounds is exponential in n (the exponential rate depending on the size of the alphabet). When combined, these bounds imply precision $O(N^{-\eta})$, for $\eta = \eta(\pi,\delta) > 0$. However, $\eta(\pi,\delta)$ is arbitrarily small for appropriate values of the parameters, since in the known bound (IV) $\rho(\pi,\delta)$ is arbitrarily close to 1 for appropriate values of the parameters.

V. Conclusions and Discussion

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. It was then illustrated how the approach is applied for characterizing the behavior of the entropy rate in various asymptotic regimes. We have also derived a deterministic algorithm for approximating the entropy rate of the HMP. This scheme, based on approximating the stationary distribution of the related Markov process, was shown to achieve the best known precision-complexity tradeoff for a significant subset of the process parameter space.

A key ingredient in the bounds developed in this work is bounding the support of the belief process. As such, the asymptotic regimes characterized via these bounds are ones that exhibit a "concentration of the support", meaning that the conditional distribution of the state given the past and present HMP components lies, with probability one, in a very small subset of the simplex of possible distributions. For example, in the 'high SNR' regime, this belief falls, with probability one, in a region of the simplex corresponding to very high certainty (that the value is either 0 or 1, depending primarily on the present observation and very weakly on the remaining ones from the past). In the 'low SNR' regime, as another example, the belief falls, with probability one, in a small region of the simplex corresponding to very low certainty.

Asymptotics of the entropy rate can be obtained also in regimes that lack this concentration property via a more delicate study of the dynamics of the alternative Markov process constructed in Section II. One such example is the 'rare transitions' regime[1] considered in [17].

References

[1] L. Arnold, L. Demetrius, and M. Gundlach. Evolutionary formalism for products of positive random matrices. *Annals of Applied Probability*, (4):859–901, 1994.

[2] J.J. Birch. Approximations for the entropy for functions of Markov chains. *Ann. Math. Stat.*, 33:930–938, 1962.

[3] 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.

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

[5] E. O. Elliot. Estimates of error rates for codes on burst-noise channels. *Bell Syst. Tech. J.*, 42:1977–1997, September 1963.

[6] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theory*, 48(6):1518–1569, June 2002.

[7] E. N. Gilbert. Capacity of a burst-noise channel. *Bell Syst. Tech. J.*, 39:1253–1265, September 1960.

[8] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Ed. Johns Hopkins, third edition, 1996.

[9] R. M. Gray. Information rates of autoregressive processes. *IEEE Trans. Inform. Theory*, 16(2):412–421, July 1970.

[10] R. M. Gray. Rate distortion functions for finite-state finite-alphabet markov sources. *IEEE Trans. Inform. Theory*, 17(2):127–134, March 1971.

[11] 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*, 45(1):128–138, January 1999.

[12] T. Holliday, P. Glynn, and A. Goldsmith. On entropy and Lyapunov exponents for finite state channels. *Preprint*.

[13] T. Holliday, P. Glynn, and A. Goldsmith. Capacity of finite state markov channels with general inputs. In *Int. Symp. Inf. Th.*, page 289, Yokohama, Japan, June-July 2003.

[14] P. Jacquet, G. Seroussi, and W. Szpankowski. On the entropy of a hidden Markov process. In *Proc. Data Compression Conference*, pages 362–371, Snowbird, Utah, USA, June 2004.

[15] D.J.C. MacKay. Equivalence of Boltzmann chains and hidden Markov models. *Neural Computation*, 8:178–181, January 1996.

[16] M. Mushkin and I. Bar-David. Capacity and coding for the Gilbert-Elliot channel. *IEEE Trans. Inform. Theory*, 35:1277–1290, 1989.

[17] C. Nair, E. Ordentlich, and T. Weissman. On asymptotic filtering and entropy rate for a hidden Markov process in the rare transitions regime. Submitted.

[18] E. Ordentlich and T. Weissman. On the optimality of symbol by symbol filtering and denoising. *IEEE Trans. Inform. Theory*. To appear.

[19] E. Ordentlich and T. Weissman. New bounds on the entropy of hidden Markov processes. In *Proc. IEEE Information Theory workshop*, San Antonio, Texas, USA, October 2004.

[20] E. Ordentlich and T. Weissman. Bounds on the Entropy Rate of Binary Hidden Markov Processes. Submitted.

[21] Y. Peres. Analytic dependence of Lyapunov exponents on transition probabilities. Number 1486 in Proceedings of Oberwolfack Conference, Lecture Notes in Math, pages 64–80. Springer Verlag Press, 1991.

[22] M. Talagrand. The Sherrington-Kirkpatrick model: a challenge to mathematicians. *Prob. Th. Relat. Fields*, 110:109–176, 1998.

[23] T. Weissman and E. Ordentlich. The empirical distribution of rate-constrained source codes. *IEEE Trans. Inform. Theory*. To appear.

[24] O. Zuk, I. Kanter, and E. Domany. Asymptotics of the entropy rate for a hidden Markov process. Preprint.

[1] In this regime, "most" of the time, between the transitions, there is high certainty regarding the value of the underlying state yet, every once in a while, around the occurrences of state transitions, an observer of the HMP will be uncertain regarding the exact location of the transition and, hence, the state values in the neighborhood of these transitions. Consequently, the support of the belief process is a large part of the simplex, which includes regions corresponding to varying degrees of certainty.