4

Bounds on the entropy rate of binary hidden Markov processes

ERIK ORDENTLICH
Hewlett-Packard Laboratories, 1501 Page Mill Rd., MS 1181, Palo Atto, CA 94304, USA
E-mail address: erik.ordentlich@hp.com

TSACHY WEISSMAN
Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Packard 256, Stanford, CA 94305, USA
E-mail address: tsachy@stanford.edu

Abstract. Let $\{X_i\}$ be a stationary finite-alphabet Markov chain and $\{Z_i\}$ denote its noisy version when corrupted by a discrete memoryless channel. Let $P(X_i \in \cdot|Z_{-\infty}^i)$ denote the conditional distribution of X_i given all past and present noisy observations, a simplex-valued random variable. We present an approach to bounding the entropy rate of $\{Z_i\}$ by approximating the distribution of this simplex-valued random variable. This approximation is facilitated by the construction and study of a Markov process whose stationary distribution determines the distribution of $P(X_i \in \cdot|Z_{-\infty}^i)$, while being more tractable than the latter. The bounds are particularly meaningful in situations where the support of $P(X_i \in \cdot|Z_{-\infty}^i)$ is significantly smaller than the whole simplex. To illustrate its efficacy, we specialize this approach 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 certain subsets of the process parameter space.

1 Introduction

1.1 The problem

Let $\{X_i\}$ be a stationary Markov chain and $\{Z_i\}$ denote its noisy version when corrupted by a discrete memoryless channel (DMC). The components of these processes take values, respectively, in the finite alphabets $\mathcal{X}$ and $\mathcal{Z}$. We let $\mathcal{K}$

Entropy of Hidden Markov Processes and Connections to Dynamical Systems: Papers from the Banff International Research Station Workshop, October 2007, ed. B. Marcus, K. Petersen, and T. Weissman. Published by Cambridge University Press.

denote the transition kernel of the Markov chain, i.e., the $|\mathcal{X}| \times |\mathcal{X}|$ matrix with entries

$$\mathcal{K}(x,x') = P(X_{i+1} = x' | X_i = x). \tag{1}$$

Let $\mathcal{C}$ denote the channel transition matrix, i.e., the $|\mathcal{X}| \times |\mathcal{Z}|$ matrix with entries

$$\mathcal{C}(x,z) = P(Z_i = z | X_i = x). \tag{2}$$

The process $\{Z_i\}$ 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.

1.2 Motivation

Hidden Markov processes (HMPs) arise naturally in many contexts, both as information sources and as noise (cf. [9] and references therein). Their entropy rate naturally arises in data compression and communications:

- Lossless compression: how many bits per source symbol are required to losslessly encode an HMP?
- Lossy compression: assume that $\mathcal{X} = \mathcal{Z}$ and that addition and subtraction of elements in this alphabet are well-defined. Assume further that the DMC relating $\{X_i\}$ to $\{Z_i\}$ is an additive noise channel with a distribution which is maximum entropy [5, Chapter 12] with respect to the per-letter additive distortion criterion $d(x,y) = \rho(x-y)$ for some nonnegative, real-valued function ρ. For example, if the DMC is symmetric, leaving the input symbol unchanged with a certain probability $p > 1/|\mathcal{X}|$, and flipping equiprobably to each of the remaining symbols, the corresponding distortion measure is Hamming loss. For this setting, the rate distortion function satisfies the Shannon lower bound [14, 13, 2, 34, 35], so is explicitly given by

$$R(D) = \overline{H}(Z) - \phi(D), \tag{3}$$

where $\phi(D)$ is the "single-letter" maximum-entropy function defined by

$$\phi(D) = \max\{H(X) : E\rho(X) \le D\},$$

with ρ as above and the maximum being over all $\mathcal{X}$-valued random variables X. Since $\phi(D)$ is readily obtainable in closed form, evaluation of the rate distortion function for the HMP reduces, by (3), to evaluation of its entropy rate.

- Channel coding: consider an additive noise channel of the form

$$Y_i = U_i + Z_i, \tag{4}$$

where $\{U_i\}$ is the transmitted channel input, $\{Y_i\}$ is the received channel output, the noise process $\{Z_i\}$ is the above-described HMP, and all process components are $|\mathcal{Z}|$-valued, with addition in (4) being in the finite-field mod-$|\mathcal{Z}|$ sense. For example, in the binary case this is the "Gilbert–Elliot" or "burst-noise" channel [10, 8, 25]. It is easy to show that the capacity of the channel in (4), for the case of no input constraints, is achieved by an i.i.d. uniform distribution on the input, implying that its capacity is given by

$$C = \log_2 |\mathcal{Z}| - \overline{H}(Z). \tag{5}$$

Evidently, key questions in lossless compression, lossy compression (3), and channel coding (5) reduce to finding the entropy rate $\overline{H}(Z)$.

1.3 On the hardness of determining $\overline{H}(Z)$

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

$$\beta_i(x) = P(X_i = x | Z_{-\infty}^i),$$

where $\beta_i(x)$ denotes the xth component of β_i. We denote this by

$$\beta_i = P(X_i \in \cdot | Z_{-\infty}^i).$$

We refer to $\{\beta_i\}$ 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_{i+1} and $Z_{-\infty}^i$ given X_i implies that $P(X_{i+1} \in \cdot | Z_{-\infty}^i) = \beta_i \cdot \mathcal{K}$, in turn implying, by the memorylessness of the noise, that

$$P(Z_{i+1} \in \cdot | Z_{-\infty}^i) = \beta_i \cdot \mathcal{K} \cdot \mathcal{C}, \tag{6}$$

where $\mathcal{K}$ and $\mathcal{C}$ are, respectively, the Markov and channel transition matrices defined in (1) and (2) (viewing elements of $\mathcal{M}(\mathcal{X})$ as row vectors). With $H(Q)$ denoting the entropy of a distribution Q on $\mathcal{Z}$,

$$H(Q) = \sum_{z \in \mathcal{Z}} Q(z) \log_2 \frac{1}{Q(z)},$$

we obtain

$$\overline{H}(Z) = H(Z_{i+1}|Z_{-\infty}^{i}) = EH\left(P(Z_{i+1} \in \cdot|Z_{-\infty}^{i})\right) = EH\left(\beta_i \cdot \mathcal{K} \cdot \mathcal{C}\right). \quad (7)$$

Evidently, the distribution of β_i holds the key to the value of the entropy rate. This distribution, however, shown by Blackwell in [4] (cf. also [29, Claim 1]) to satisfy an integral equation, remains elusive to date even for the simplest HMPs.

Another perspective by which the hardness of the problem can be appreciated is that developed in [21, 20] of Lyapunov exponents. Standard recursion for HMPs yields [9]

$$P(Z^n = z^n) = \mu_s^{\mathrm{T}} \left[\prod_{i=1}^{n} \left[\mathcal{K} \odot \mathcal{C}(\cdot, z_i)^{\mathrm{T}}\right]\right] \mathbf{1},$$

where μ_s is the stationary distribution of the underlying Markov chain (represented as a column vector), $\mathcal{K} \odot \mathcal{C}(\cdot, z_i)^{\mathrm{T}}$ denotes the $|\mathcal{X}| \times |\mathcal{X}|$ matrix whose xth row is given by the componentwise multiplication of the xth row of $\mathcal{K}$ by the row vector whose x'th component is $\mathcal{C}(x', z_i)$, and $\mathbf{1}$ denotes the "all-ones" column vector. The Shannon–McMillan–Breiman theorem [5] implies then that, with probability one,

$$\begin{aligned}\overline{H}(Z) &= \lim_{n\to\infty} -\frac{1}{n} \log_2 \mu_s^{\mathrm{T}} \left[\prod_{i=1}^{n} \left[K \odot C(\cdot, Z_i)^{\mathrm{T}}\right]\right] \mathbf{1} \\ &= \lim_{n\to\infty} -\frac{1}{n} \log_2 \left\| \prod_{i=1}^{n} \left[K \odot C(\cdot, Z_i)^{\mathrm{T}}\right] \right\|, \end{aligned} \quad (8)$$

where $\|\cdot\|$ denotes any matrix norm. In other words, up to sign, $\overline{H}(Z)$ is the (top) Lyapunov exponent (cf., e.g., [30]) associated with the square-matrix-valued process $\{K \odot C(\cdot, Z_i)^{\mathrm{T}}\}_{i\geq 1}$. Characterization of the Lyapunov exponent is an open question, even in the simplest cases of finite-valued i.i.d. matrices [30, 1]. In our case, $\{K \odot C(\cdot, Z_i)^{\mathrm{T}}\}_{i\geq 1}$ is not even a Markov process.

Yet another perspective on the problem is that of statistical physics. For simplicity, consider the HMP given by a binary symmetric channel (BSC) corrupted symmetric binary Markov source. The distribution of Z^n can be put in

the form [36, 24]

$$P(Z^n) = \sum_{x^n} P(x^n)P(Z^n|x^n) = \sum_{x^n} P(x_1)\prod_{i=1}^{n-1}\mathcal{K}(x_i,x_{i+1})\prod_{i=1}^{n}\mathcal{C}(x_i,Z_i)$$
$$= c_1 c_2^n \sum_{\tau^n} \exp\left(J\sum_{i=1}^{n-1}\tau_i\tau_{i+1} + K\sum_{i=1}^{n}\tau_i\sigma_i\right), \tag{9}$$

where the last equality is obtained by a change of variables $\tau_i = (-1)^{X_i}$ and $\sigma_i = (-1)^{Z_i}$, and c_1, c_2, J, K are explicit functions of the process parameters. Characterization of the entropy rate reduces then to that of the limit, in n, of

$$\frac{1}{n} E \log_2 \sum_{\tau^n} \exp\left(J\sum_{i=1}^{n-1}\tau_i\tau_{i+1} + K\sum_{i=1}^{n}\tau_i\sigma_i\right), \tag{10}$$

which is the expected density of the logarithm of the partition function associated with the Gibbs measure for (random) energy levels $E(\tau^n) = -\sum_{i=1}^{n-1}\tau_i\tau_{i+1} - K/J\sum_{i=1}^{n}\tau_i\sigma_i$ at temperature $1/J$. The asymptotic value, for large n, of this expected density is an open problem in statistical physics *even* for the case where the energy levels are i.i.d. [33].

1.4 Existing results

Given the hardness of the problem, the predominant approach to the study of the entropy rate, until relatively recently, has been one of approximation (cf. [25, 21, 7] and references therein). Indeed, what we refer to as the "Cover and Thomas" bounds

$$H(Z_0|Z_{-n}^{-1}, X_{-n-1}) \le \overline{H}(Z) \le H(Z_0|Z_{-n}^{-1}) \tag{11}$$

hold for every n, becoming arbitrarily tight with increasing n [5, Section 4.5]. We shall discuss these bounds in more detail in Section 5, where we suggest an alternative deterministic scheme for approximating the entropy rate. Another approach for approximating the entropy rate is via (8), which implies that simulating the HMP and evaluating $-\frac{1}{n}\log_2 \left\| \prod_{i=1}^{n} \left[K \odot C(\cdot, Z_i)^{\mathrm{T}}\right] \right\|$ gives an estimate, which, for large n, becomes arbitrarily precise with probability arbitrarily close to one (cf. [21] 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 were taken in [21], where continuity of the entropy rate in the parameters was established. Significant progress in this direction was made by Han and Marcus in a recent series of papers [15, 16, 17, 18], which not only established smoothness, but also characterized conditions for differentiability and analyticity of the entropy rate in the transition parameters.

Expansions of the entropy rate for the BSC-corrupted binary Markov chain in the "high SNR" regime, where the channel cross-over probability is small, have been obtained initially in [22, 29, 36, 27]. Initial results on the behavior in various additional asymptotic regimes such as "rare spikes", "rare bursts", "low SNR", and "almost memoryless" were obtained in [29]. We expand on and strengthen this line of results in subsequent sections by incorporating into the approach of [29] finer properties of the distribution of the belief process β_i, as summarized in the next subsection. More recent refinements and extensions were obtained in [15, 16, 17].

1.5 Our approach

An immediate consequence of (7) is

Observation 4.1.

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

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

Trivial as this observation may seem, it was shown in [29] to lead to useful bounds in cases where bounds on the support set $\mathcal{S}$ are obtainable, and these bounds are significantly smaller than the whole simplex $\mathcal{M}(\mathcal{X})$.

The bounds of Observation (4.1), which depend on the distribution of β_i through its support only, can be refined by covering $\mathcal{S}$ using several disjoint sets and considering also the probabilities of these sets.

Observation 4.2. For any countable collection $\{I_k\}$ of pairwise-disjoint sets $I_k \subseteq \mathcal{M}(\mathcal{X})$ covering $\mathcal{S}$ (i.e., for which $\mathcal{S} \subseteq \bigcup_k I_k$),

$$\sum_k P(\beta_i \in I_k) \inf_{\beta\in I_k} H\left(\beta\cdot\mathcal{K}\cdot\mathcal{C}\right) \le \overline{H}(Z) \le \sum_k P(\beta_i \in I_k) \sup_{\beta\in I_k} H\left(\beta\cdot\mathcal{K}\cdot\mathcal{C}\right). \quad (12)$$

Since the distribution of β_i is unknown, $P(\beta_i \in I_k)$ will also be unknown in general. However, for certain choices of $\{I_k\}$, and in certain regions of the space of parameters governing the HMP, the bounds in (12) can be either explicitly

evaluated or closely bounded. This is done by constructing a Markov process which is more tractable than the $\{\beta_i\}$ process. The stationary distribution of this process is directly and simply related to the distribution of β_i. The fraction of times that the process visits the set I_k, for appropriately chosen I_k, is computable, a computation that can then be directly translated to give the value of $P(\beta_i \in I_k)$.

For concreteness and simplicity in illustrating the idea, we concentrate on the case where $\{Z_i\}$ is a BSC-corrupted binary Markov chain. In this context, the two new ingredients of our approach relative to [29] involve covering the support of the belief process by multiple disjoint intervals I_k (as opposed to only two intervals in [29]) and the construction of an alternative, more tractable, Markov process as a tool for analyzing the probabilities $P(\beta_i \in I_k)$ of the belief process falling into these intervals. The incorporation of these finer properties of β_i is shown to lead to tighter characterizations of $\overline{H}(Z)$, in various asymptotic regimes, than were obtained in [29]. The alternative Markov process is also leveraged to obtain and analyze the aforementioned deterministic algorithm for numerically approximating $\overline{H}(Z)$. Several of the above results have appeared in preliminary form in our previous conference papers [27, 28].

We remark that while our approach is based, in part, on finite coverings of the support $\mathcal{S}$ of the belief process, little is known about $\mathcal{S}$, as a whole, beyond the observations in [4]. It is shown therein, via examples, that $\mathcal{S}$, in general, can be a finite set, a countable set, or an uncountable set with Lebesgue measure 0 (in a strong sense made precise in [4]). It is in fact conjectured in [4] that if the distribution of β_i is continuous (e.g., no point masses), it will be singular with respect to Lebesgue measure (again, in a strong sense made precise in [4]).

1.6 Remaining content

In Section 2, 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) for the case of the BSC-corrupted Markov chain. We then detail the construction of an alternative Markov process, and its relationship to the original log-likelihood (and, therefore, belief) process. Section 3 focuses on the case of a symmetric Markov chain, and details the form the bounds in (12) assume for this case, in terms of the alternative Markov process. Using these bounds, we then derive the behavior of the entropy rate in various asymptotic regimes. Section 4 follows a similar development for the case where the underlying binary Markov chain is not necessarily symmetric. In Section 5 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 all known deterministic schemes for approximation of the entropy rate, for certain subsets of the parameter space. This algorithm was preliminarily presented in [28] for the symmetric Markov chain case. More recently, a similar approach was taken in [23], again for the symmetric Markov chain case, but the details of the resulting algorithm and its analysis are different. Section 6 contains a summary of the paper, along with a discussion of some related directions.

2 The BSC-corrupted Binary Markov Chain

2.1 Setup and some notation

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

$$\mathcal{K}=\begin{pmatrix}1-\pi_{01} & \pi_{01}\\ \pi_{10} & 1-\pi_{10}\end{pmatrix},\quad \mathcal{C}=\begin{pmatrix}1-\delta & \delta\\ \delta & 1-\delta\end{pmatrix}. \tag{13}$$

Without loss of generality, we assume that $\delta\le 1/2$ and $\pi_{01}\le\pi_{10}$. To avoid trivialities, we also assume below that

1. $\pi_{01}+\pi_{10}\neq 1$ (otherwise the state process is i.i.d.).
2. Either $\pi_{01}\in(0,1)$ or $\pi_{10}\in(0,1)$ (otherwise the state process is essentially (up to its initial state) deterministic and $\overline{H}(Z)=h_b(\delta)$).

For positive-valued functions f and g, $f(\varepsilon)\sim g(\varepsilon)$ will stand for $\lim_{\varepsilon\downarrow 0}\frac{f(\varepsilon)}{g(\varepsilon)}=1$ and $f(\varepsilon)\stackrel{\sim}{<}g(\varepsilon)$ will stand for $\limsup_{\varepsilon\downarrow 0}\frac{f(\varepsilon)}{g(\varepsilon)}\le 1$. $f(\varepsilon)=O(g(\varepsilon))$ will stand for $\limsup_{\varepsilon\downarrow 0}\frac{f(\varepsilon)}{g(\varepsilon)}<\infty$ and $f(\varepsilon)=\Omega(g(\varepsilon))$ will stand for $\liminf_{\varepsilon\downarrow 0}\frac{f(\varepsilon)}{g(\varepsilon)}>0$. $f(\varepsilon)\asymp g(\varepsilon)$ will stand for the statement that both $f(\varepsilon)=O(g(\varepsilon))$ and $f(\varepsilon)=\Omega(g(\varepsilon))$ hold. If R is a random variable, $\mathcal{L}(R)$ will denote its law. Similarly, if A is an event, $\mathcal{L}(R|A)$ will denote the law of R conditioned on A. Also, for R,S random variables and $0\le\alpha\le 1$, $\alpha\mathcal{L}(R)+(1-\alpha)\mathcal{L}(S)$ will denote the law of $BR+(1-B)S$, for $B\sim\text{Bernoulli}(\alpha)$ that is independent of R and S. We extend this interpretation to the combination of more than two laws, namely to $\sum_i\alpha_i\mathcal{L}(R_i)$ for $\alpha_i\ge 0$ summing to 1 and R_i random variables, in the obvious way. Throughout the article, $\log_2$ and log will denote the base-2 and natural logarithms, respectively, and all entropies and entropy rates are expressed in bits (underlying $\log_2$).

2.2 Evolution of the log-likelihood

The standard forward recursions [9] are readily shown (cf., e.g., [29]) to assume the form

$$\frac{\beta_i(0)}{1-\beta_i(0)} = \left[\frac{1-\delta}{\delta}\right]^{1-2Z_i} g\left(\frac{\beta_{i-1}(0)}{1-\beta_{i-1}(0)}\right), \tag{14}$$

where

$$g(x) = \frac{x(1-\pi_{01})+\pi_{10}}{x\pi_{01}+(1-\pi_{10})}. \tag{15}$$

Equivalently, this can be expressed as

$$l_i = (2Z_i - 1)\log\left[\frac{1-\delta}{\delta}\right] + f(l_{i-1}), \tag{16}$$

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

$$f(x) = \log\frac{\pi_{01}+e^x(1-\pi_{10})}{(1-\pi_{01})+e^x\pi_{10}}. \tag{17}$$

It follows from (7) that, in terms of the log-likelihood process, the entropy rate is given by

$$\begin{aligned}\overline{H}(Z) &= Eh_b([\beta_i(1)(1-\pi_{10})+(1-\beta_i(1))\pi_{01}]*\delta) \\ &= Eh_b\left(\left[\frac{e^{l_i}}{1+e^{l_i}}(1-\pi_{10})+\frac{1}{1+e^{l_i}}\pi_{01}\right]*\delta\right),\end{aligned} \tag{18}$$

where $*$ denotes binary convolution defined by $p*q = (1-p)q+(1-q)p$ and

$$h_b(x) = -x\log_2 x - (1-x)\log_2(1-x)$$

is the binary entropy function.

In the sequel, we shall make use of the following properties of the function f in (17). First, note that

$$f'(x) = \frac{e^x(1-\pi_{01}-\pi_{10})}{(1-\pi_{01})\pi_{01}+e^{2x}(1-\pi_{10})\pi_{10}+e^x(1-\pi_{01}-\pi_{10}+2\pi_{01}\pi_{10})}, \tag{19}$$

which has the sign of the numerator, so f is either strictly increasing or strictly decreasing according to whether $\pi_{01}+\pi_{10}<1$ or $\pi_{01}+\pi_{10}>1$. Another

important property of f is its contractiveness. To be sure, note that

$$\begin{aligned}\sup_x |f'(x)| &= \frac{|1-\pi_{01}-\pi_{10}|}{\min_{y>0}\left[(1-\pi_{01})\pi_{01}y^{-1}+(1-\pi_{10})\pi_{10}y+(1-\pi_{01}-\pi_{10}+2\pi_{01}\pi_{10})\right]} \\ &= \frac{|1-\pi_{01}-\pi_{10}|}{2\sqrt{(1-\pi_{01})\pi_{01}(1-\pi_{10})\pi_{10}}+(1-\pi_{01}-\pi_{10}+2\pi_{01}\pi_{10})} \\ &\triangleq c(\pi_{01},\pi_{10}), \end{aligned} \tag{20}$$

where for the second equality we have used the elementary fact that for nonnegative a,b,c

$$\min_{y>0}\left[ay^{-1}+by+c\right]=2\sqrt{ab}+c.$$

This implies that f is contractive since $c(\pi_{01},\pi_{10})<1$ (the denominator is obviously greater than the numerator if $\pi_{01}+\pi_{10}<1$ and is invariant under the transformation $(\pi_{01},\pi_{10})\to(1-\pi_{01},1-\pi_{10})$).

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

$$l_i=(2Z_i-1)\log\left[\frac{1-\delta}{\delta}\right]+f\left(l_{i-1}\right), \tag{21}$$

where $f(x)=\log\frac{e^x(1-\pi)+\pi}{e^x\pi+(1-\pi)}$. Specializing (18) for this case gives

$$\overline{H}(Z)=Eh_b\left(\frac{e^{l_i}}{1+e^{l_i}}*\pi*\delta\right). \tag{22}$$

In this symmetric case, (20) becomes

$$\sup_x |f'(x)|=1-2\pi. \tag{23}$$

Thus, as discussed in the introduction, the distribution of β_i (or, equivalently, of l_i) is key to the evaluation of the entropy rate. Although $\{\beta_i\}$ was shown to be a Markov process by Blackwell [4], 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 β_i.

2.3 An alternative Markov process

In this subsection, we 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 (18), or (22) in the symmetric case, will be expressible as a similar expectation involving the new process. Our alternative process is closely related to the joint state/belief process (X_i, β_i) studied in [11] and [32]. The former reference derived conditions for the geometric ergodicity (exponential convergence to the stationary distribution) of the joint process and the latter reference applied these conditions to give a simple proof of Birch's result [3] on the exponential decay (in n) of the difference between the upper and lower bounds on the entropy rate in (11) above.

2.3.1 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 where $\pi_{10} = \pi_{01} = \pi < 1/2$. There is no loss of generality in assuming that $\pi < 1/2$ since the argument in [29, 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.

Theorem 4.3. *Consider the first-order Markov process $\{Y_i\}_{i\geq 0}$ formed by letting $Y_0 = Y$ and $\{Y_i\}_{i\geq 1}$ evolve according to*

$$Y_i = r_i \log \frac{1-\delta}{\delta} + s_i f(Y_{i-1}), \tag{24}$$

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

$$r_i = \begin{cases} -1 & \text{w.p. } \delta, \\ 1 & \text{w.p. } 1-\delta, \end{cases} \quad s_i = \begin{cases} -1 & \text{w.p. } \pi, \\ 1 & \text{w.p. } 1-\pi. \end{cases} \tag{25}$$

In this theorem, "w.p." stands for "with probability".

1. [Then Existence and uniqueness of the 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] $\mathcal{L}(Y) = \mathcal{L}(l_i | X_i = 1)$.

Proof. It is evident from (24) and (25) that a distribution on Y is a stationary distribution for the process $\{Y_i\}$ if and only if it satisfies

$$\begin{aligned} \mathcal{L}(Y) = {} & \pi\delta \cdot \mathcal{L}\left(-\log\frac{1-\delta}{\delta} - f(Y)\right) + (1-\pi)\delta \cdot \mathcal{L}\left(-\log\frac{1-\delta}{\delta} + f(Y)\right) \\ & + \pi(1-\delta) \cdot \mathcal{L}\left(\log\frac{1-\delta}{\delta} - f(Y)\right) + (1-\pi)(1-\delta) \\ & \cdot \mathcal{L}\left(\log\frac{1-\delta}{\delta} + f(Y)\right). \end{aligned} \tag{26}$$

To prove uniqueness, assume first that there exists a distribution on Y satisfying (26), and let $\{\tilde{Y}_i\}$ denote the stationary process evolving according to (24), initiated at time 0 with an arbitrary stationary distribution (and arbitrarily jointly distributed with Y). Then, due to (23),

$$|\tilde{Y}_i - Y_i| = |f(\tilde{Y}_{i-1}) - f(Y_{i-1})| \leq (1-2\pi)|\tilde{Y}_{i-1} - Y_{i-1}|,$$

so

$$|\tilde{Y}_i - Y_i| \leq (1-2\pi)^i |\tilde{Y}_0 - Y_1|$$

and, in particular,

$$|\tilde{Y}_i - Y_i| \longrightarrow 0$$

as $i \to \infty$ (for all sample paths). This implies, when combined with the stationarity of both processes, that $\tilde{Y}_0 \stackrel{d}{=} Y_0$. To prove existence, as well as the second assertion of the theorem, it will suffice to establish the fact that taking $\mathcal{L}(Y) = \mathcal{L}(l_i|X_i = 1)$ satisfies (26). To see this, note first that for all $\alpha \in \mathbb{R}$,

$$\begin{aligned}
P(f(l_{i-1}) \leq \alpha | X_i = 1) &= \sum_j P(f(l_{i-1}) \leq \alpha, X_{i-1} = j | X_i = 1) \\
&= \sum_j P(X_{i-1} = j | X_i = 1) P(f(l_{i-1}) \leq \alpha | X_{i-1} = j) \\
&= \pi P(f(l_{i-1}) \leq \alpha | X_{i-1} = 0) + (1-\pi) P(f(l_{i-1}) \\
&\leq \alpha | X_{i-1} = 1) \\
&= \pi P(f(-l_{i-1}) \leq \alpha | X_{i-1} = 1) + (1-\pi) P(f(l_{i-1}) \\
&\leq \alpha | X_{i-1} = 1),
\end{aligned}$$

the last equality following since, by symmetry, $\mathcal{L}(l_{i-1}|X_{i-1} = 0) = \mathcal{L}(-l_{i-1}|X_{i-1} = 1)$. The strict increasing monotonicity of $f(\cdot)$ implies then that

$$\mathcal{L}(l_{i-1}|X_i = 1) = \pi \mathcal{L}(-l_i|X_i = 1) + (1-\pi)\mathcal{L}(l_i|X_i = 1). \tag{27}$$

Now, *conditioned on the event* $X_i = 1$, the two summands on the right-hand side of (16) are *independent*, with the first being distributed as

$$(2Z_i - 1) \log\left[\frac{1-\delta}{\delta}\right] = \begin{cases} \log \frac{1-\delta}{\delta} & \text{w.p. } 1-\delta, \\ -\log \frac{1-\delta}{\delta} & \text{w.p. } \delta. \end{cases} \tag{28}$$

Combined with (27), this implies that the distribution $\mathcal{L}(l_i|X_i = 1)$ satisfies (26). □

Henceforth, when referring to the process defined in Theorem 4.3, we assume that it was initiated by the stationary distribution.

Corollary 4.4. *For the process constructed in Theorem 4.3,*

$$\overline{H}(Z) = Eh_b\left(\frac{e^{Y_i}}{1+e^{Y_i}} * \pi * \delta\right). \tag{29}$$

Proof. We have

$$\begin{aligned}\overline{H}(Z) &= \frac{1}{2}E\left[h_b\left(\frac{e^{l_i}}{1+e^{l_i}} * \pi * \delta\right)|X_i = 1\right] + \frac{1}{2}E\left[h_b\left(\frac{e^{l_i}}{1+e^{l_i}} * \pi * \delta\right)|X_i = 0\right] \\ &= Eh_b\left(\frac{e^{Y_i}}{1+e^{Y_i}} * \pi * \delta\right),\end{aligned} \tag{30}$$

the first equality following from (22) and the second from the second item of Theorem 4.3 and the facts that $\mathcal{L}(l_i|X_i = 1) = \mathcal{L}(-l_i|X_i = 0)$ and that $h_b\left(\frac{e^y}{1+e^y} * \pi * \delta\right) = h_b\left(\frac{1}{2} - \frac{e^y}{1+e^y} * \pi * \delta\right) = h_b\left(\frac{e^{-y}}{1+e^{-y}} * \pi * \delta\right)$ for all y. □

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 so is that information on the distribution of Y_i, which translates via Corollary 4.4 to bounds on the entropy rate, can be inferred by studying the dynamics of the process $\{Y_i\}$.

2.3.2 The non-symmetric case

In the symmetric case, it sufficed to construct one process with real-valued components whose stationary distribution is $\mathcal{L}(l_i|X_i = 1)$, since this immediately conveyed also $\mathcal{L}(l_i|X_i = 0)$ as, by symmetry, $\mathcal{L}(l_i|X_i = 1) = \mathcal{L}(-l_i|X_i = 0)$. In the nonsymmetric case, we have the following theorem.

Theorem 4.5. *Define $\{(Y_i, U_i)\}_{i\geq 0}$, a Markov process with state space $\mathbb{R}^2$, by letting $(Y_0, U_0) = (Y, U)$ and, for $i \geq 1$,*

$$Y_i = r_i \log\frac{1-\delta}{\delta} + s_i f(U_{i-1}) + (1-s_i) f(Y_{i-1}) \tag{31}$$

and

$$U_i = q_i \log\frac{1-\delta}{\delta} + (1-t_i) f(U_{i-1}) + t_i f(Y_{i-1}), \tag{32}$$

where $\{q_i\}$, $\{r_i\}$, $\{s_i\}$, $\{t_i\}$ are independent i.i.d. sequences, independent of (Y, U), with

$$q_i = \begin{cases} 1 & \text{w.p. } \delta, \\ -1 & \text{w.p. } 1-\delta, \end{cases} \quad r_i = \begin{cases} -1 & \text{w.p. } \delta, \\ 1 & \text{w.p. } 1-\delta, \end{cases} \tag{33}$$

and $s_i \sim Bernoulli(\pi_{10})$, $t_i \sim Bernoulli(\pi_{01})$. *Then*

1. [Existence of a marginally stationary distribution] *There exists a distribution on the pair* (Y,U) *under which* $\{(Y_i,U_i)\}_{i\geq 0}$ *is marginally stationary in the sense that, for all* $i \geq 0$, $\mathcal{L}(Y_i)=\mathcal{L}(Y)$ *and* $\mathcal{L}(U_i)=\mathcal{L}(U)$.
2. [Uniqueness of marginals and connection to the original process] *Any distribution on* (Y,U) *giving rise to a process which is marginally stationary in the sense of the previous item satisfies* $\mathcal{L}(Y)=\mathcal{L}(l_i|X_i=1)$ *and* $\mathcal{L}(U)=\mathcal{L}(l_i|X_i=0)$.

Remark. We will refer to a distribution on (Y,U) that gives rise to a process $\{(Y_i,U_i)\}$ which is marginally stationary in the above sense as a "marginally stationary distribution". It is evident from the evolution equations (31) and (32) that $\mathcal{L}(Y_i)$ and $\mathcal{L}(U_i)$ depend on $\mathcal{L}(Y_{i-1},U_{i-1})$ only through the marginal distributions $\mathcal{L}(Y_{i-1})$ and $\mathcal{L}(U_{i-1})$. Thus, if a distribution on the pair (Y,U) gives rise to a marginally stationary process, then any other distribution with the same marginal distributions of U and Y has the same property. Conversely, the second item of the theorem implies that *all* distributions on (Y,U) giving rise to a marginally stationary process will share the same marginals, which, respectively, are given by $\mathcal{L}(l_i|X_i=1)$ and $\mathcal{L}(l_i|X_i=0)$. In particular, the marginal stationary distributions are unique.

Proof of Theorem 4.5. Conditioned on the event $X_i=1$, the two summands on the right-hand side of (16) are independent with

$$(2Z_i-1)\log\left[\frac{1-\delta}{\delta}\right]=\begin{cases}\log\frac{1-\delta}{\delta} & \text{w.p. } 1-\delta,\\ -\log\frac{1-\delta}{\delta} & \text{w.p. } \delta.\end{cases} \tag{34}$$

Furthermore, for any $\alpha\in\mathbb{R}$,

$$\begin{aligned}P(f(l_{i-1})\leq\alpha|X_i=1)&=\sum_j P(f(l_{i-1})\leq\alpha,X_{i-1}=j|X_i=1)\\&=\sum_j P(X_{i-1}=j|X_i=1)P(f(l_{i-1})\leq\alpha|X_{i-1}=j)\\&=\pi_{10}P(f(l_{i-1})\leq\alpha|X_{i-1}=0)+(1-\pi_{10})P(f(l_{i-1})\\&\quad\leq\alpha|X_{i-1}=1),\end{aligned}$$

implying with the strict monotonicity of f that

$$\mathcal{L}(l_{i-1}|X_i=1)=\pi_{10}\mathcal{L}(l_{i-1}|X_{i-1}=0)+(1-\pi_{10})\mathcal{L}(l_{i-1}|X_{i-1}=1). \tag{35}$$

Similarly, conditioned on the event $X_i = 0$, the two summands on the right-hand side of (16) are independent with

$$(2Z_i - 1)\log\left[\frac{1-\delta}{\delta}\right] = \begin{cases} \log\frac{1-\delta}{\delta} & \text{w.p. } \delta, \\ -\log\frac{1-\delta}{\delta} & \text{w.p. } 1-\delta, \end{cases} \tag{36}$$

and a computation similar to that leading to (35) gives

$$\mathcal{L}(l_{i-1}|X_i = 0) = \pi_{01}\mathcal{L}(l_{i-1}|X_{i-1} = 1) + (1-\pi_{01})\mathcal{L}(l_{i-1}|X_{i-1} = 0). \tag{37}$$

It follows from (16), (34)–(37), and the mentioned conditional independence that any distribution on (Y, U) where $\mathcal{L}(Y) = \mathcal{L}(l_i|X_i = 1)$ and $\mathcal{L}(U) = \mathcal{L}(l_i|X_i = 0)$ gives a marginally stationary process, establishing the first item.

For the second item, let $\{(Y_i, U_i)\}$ be the process initiated by a distribution on (Y, U) with $\mathcal{L}(Y) = \mathcal{L}(l_i|X_i = 1)$ and $\mathcal{L}(U) = \mathcal{L}(l_i|X_i = 0)$. Let $\{(\tilde{Y}_i, \tilde{U}_i)\}$ be the same process started at any other point. Then

$$\begin{aligned} |\tilde{Y}_i - Y_i| &\le s_i|f(\tilde{U}_{i-1}) - f(U_{i-1})| + (1-s_i)|f(\tilde{Y}_{i-1}) - f(Y_{i-1})| \\ &\le s_i c(\pi_{01},\pi_{10})|\tilde{U}_{i-1} - U_{i-1}| + (1-s_i)c(\pi_{01},\pi_{10})|\tilde{Y}_{i-1} - Y_{i-1}| \\ &\le c(\pi_{01},\pi_{10})\max\{|\tilde{U}_{i-1} - U_{i-1}|, |\tilde{Y}_{i-1} - Y_{i-1}|\}, \end{aligned} \tag{38}$$

where the inequality before last follows from (20) (with $c(\pi_{01},\pi_{10}) < 1$). We similarly obtain

$$|\tilde{U}_i - U_i| \le c(\pi_{01},\pi_{10})\max\{|\tilde{U}_{i-1} - U_{i-1}|, |\tilde{Y}_{i-1} - Y_{i-1}|\}$$

which, combined with (38), yields

$$\max\{|\tilde{U}_i - U_i|, |\tilde{Y}_i - Y_i|\} \le c(\pi_{01},\pi_{10})\max\{|\tilde{U}_{i-1} - U_{i-1}|, |\tilde{Y}_{i-1} - Y_{i-1}|\}. \tag{39}$$

Iterating gives

$$\|(\tilde{U}_i, \tilde{Y}_i) - (U_i, Y_i)\|_\infty \le c(\pi_{01},\pi_{10})^i \|(\tilde{U}_0, \tilde{Y}_0) - (U_0, Y_0)\|_\infty, \tag{40}$$

implying

$$|\tilde{U}_i - U_i| \longrightarrow 0, \quad |\tilde{Y}_i - Y_i| \longrightarrow 0. \tag{41}$$

But $\mathcal{L}(Y_i) = \mathcal{L}(Y)$ and $\mathcal{L}(U_i) = \mathcal{L}(U)$ for all i thus, if the 'tilded' process is marginally stationary, the only way this could be consolidated with (41) is if $\mathcal{L}(\tilde{Y}_i) = \mathcal{L}(Y)$ and $\mathcal{L}(\tilde{U}_i) = \mathcal{L}(U)$ for all i. □

Henceforth, when referring to the process constructed in Theorem 4.5, it will be implied that it was initiated with a marginally stationary distribution on (U,Y). When combined with (18), Theorem 4.5 implies

Corollary 4.6. *For the process constructed in Theorem 4.5*

$$\overline{H}(Z) = \frac{\pi_{01}}{\pi_{01}+\pi_{10}} Eh_b\left(\left[\frac{e^{Y_i}}{1+e^{Y_i}}(1-\pi_{10})+\frac{1}{1+e^{Y_i}}\pi_{01}\right]*\delta\right) + \frac{\pi_{10}}{\pi_{01}+\pi_{10}} Eh_b\left(\left[\frac{e^{U_i}}{1+e^{U_i}}(1-\pi_{10})+\frac{1}{1+e^{U_i}}\pi_{01}\right]*\delta\right). \quad (42)$$

3 Bounds on the Entropy Rate for the Symmetric Chain

In this section we use Corollary 4.4 to bound the entropy rate in the symmetric case, by bounding the expectation on the right side of (29). Assume throughout this section the case of a BSC-corrupted symmetric Markov chain with $0 < \pi_{10} = \pi_{01} = \pi \leq 1/2$. The following gives the general form of our bounds, and the condition under which they are valid.

Observation 4.7. Let $\{Y_i\}$ be the stationary Markov process whose evolution is given by (24). Let $\{a_i\}_{i=1}^M, \{b_i\}_{i=1}^M$ be strictly increasing sequences of nonnegative reals such that $b_k \leq a_k$ and $b_{k+1} > a_k$ (i.e., the intervals $[b_k, a_k]$ do not intersect). Assume further that $\bigcup_{k=1}^M [b_k,a_k] \cup \bigcup_{k=1}^M [-a_k,-b_k]$ contains the support of Y_i. Then

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

Proof. Immediate from Corollary 4.4 and the decreasing monotonicity of $h_b\left(\frac{e^y}{1+e^y} * \pi * \delta\right)$ in the absolute value of y. □

Evidently, a bound of the type in Observation 4.7 would be applicable only in situations where: 1) the support of Y_i is, in fact, contained in a set of the form $\bigcup_{k=1}^M [b_k,a_k] \cup \bigcup_{k=1}^M [-a_k,-b_k]$ and 2) the probabilities $P\left(Y_i \in [-a_k,-b_k]\cup[b_k,a_k]\right)$ 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 Observation 4.7 yields

Corollary 4.8. *Let* $\{Y_i\}$ *be the process in (24). Let* $0 \le b \le A$ *be such that* $[-A,-b]\cup[b,A]$ *contains the support of* Y_i*. Then*

$$h_b\left(\frac{e^A}{1+e^A} * \pi * \delta\right) \le \overline{H}(Z) \le h_b\left(\frac{e^b}{1+e^b} * \pi * \delta\right). \tag{43}$$

The lower bound of Corollary 4.8 is clearly optimized when taking A to be the upper end point of the support of Y_i. This point is readily seen, by observation of the dynamics of the process $\{Y_i\}$ in (24), to be the solution to the equation

$$A = f(A) + \log\frac{1-\delta}{\delta}, \tag{44}$$

namely

$$A = \log\frac{\alpha - 1 + (1-\alpha)\pi + \sqrt{4\alpha\pi^2 + (1-\alpha-(1-\alpha)\pi)^2}}{2\pi}, \tag{45}$$

where $\alpha = \frac{1-\delta}{\delta}$ (cf. Section 4 of [29]). The obvious symmetry of the support of Y_i around 0 implies that $-A$ is the lower end point of the support of Y_i. In particular, this establishes that the support of Y_i is contained in the interval $[-A,A]$, cf. Figure 1. Similarly, to optimize the upper bound, b should be taken as the lower end point of this support in the positive half of the real line. For the case where δ is small enough so that the first term on the right-hand side of (24) uniquely determines the sign of Y_i ("small enough", as will be made explicit below, is any value in the shaded region of Figure 4), the value of this lower end point can be read from the dynamics of the process in (24) (see also the proof of Lemma 4.9 below) to be given by

$$b = -f(A) + \log\frac{1-\delta}{\delta}. \tag{46}$$

Similarly, by symmetry, $-b$ is the upper end point of the support of Y_i in the negative half. This implies then that the support of Y_i is contained in $[-A,-b]\cup[b,A]$ (cf. Figure 2) and that A and b of (45) and (46) are, respectively, the smallest and largest values with this property. Crude as the bounds of Corollary 4.8 may seem, they were shown in [29] (obtained therein directly from the likelihood process (21)) to convey nontrivial information when optimized by substituting the values of A and b from (45) and (46). In particular, this led to varying degrees of precision in characterizing the entropy rate in various asymptotic regimes. Examples include (letting $\overline{H}(\pi,\delta)$ stand for $\overline{H}(Z)$ when π and δ are, respectively, the chain and channel transition probabilities):

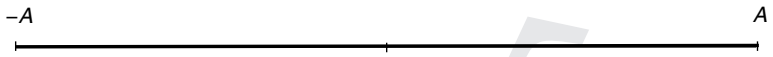

Figure 1 Smallest interval containing the support of Y_i. A is given by (45).

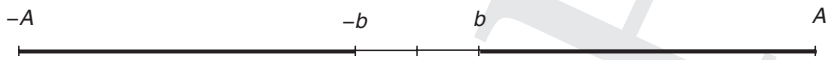

Figure 2 Smallest set of the form $[-A,-b]\cup[b,A]$ containing the support of Y_i. A and b are given, respectively, in (45) and (46).

Figure 3 Smallest set of the form $[-A,-B]\cup[-a,-b]\cup[b,a]\cup[B,A]$ containing the support of Y_i. A,b,a, and B are given, respectively, in (45), (46), (52), and (53). Probabilities of the four intervals are stated in Lemma 4.9.

"High SNR": for $0 \le \pi \le 1/2$, as $\delta \to 0$,

$$\overline{H}(\pi,\delta) - h_b(\pi) \asymp \delta. \tag{47}$$

"Almost memoryless": for $0 \le \delta \le 1/2$, as $\varepsilon \to 0$,

$$\frac{1-\overline{H}\left(\frac{1}{2}-\varepsilon,\delta\right)}{\varepsilon^2} \sim \frac{2}{\log 2}(1-2\delta)^4. \tag{48}$$

"Low SNR": for $1/4 \le \pi < 1/2$, as $\varepsilon \to 0$,

$$1-\overline{H}\left(\pi,\frac{1}{2}-\varepsilon\right) \asymp \varepsilon^4. \tag{49}$$

It is instructive to compare with the implications of the bounds of Cover and Thomas [5, Section 4.5] for these regimes. In our setting, a simple calculation shows that $H(Z_0|X_{-1}) = h_b(\pi * \delta)$ and $H(Z_0|Z_{-1}) = h_b(\pi * \delta * \delta)$, so the first-order ($n=1$) bounds are

$$h_b(\pi * \delta) \le \overline{H}(\pi,\delta) \le h_b(\pi * \delta * \delta), \tag{50}$$

which implies (47) (but no more), recovers the ε^2 behavior in (48) but without the constant, and does not recover the ε^4 behavior in (49). In fact, as was mentioned in [29], there are regimes in which the bounds of [5, Section 4.5],

of any order, will not capture the behavior of the entropy rate. For a simple example note that, in our binary symmetric setting, for any n,

$$H(Z_0|Z_{-n+1}^{-1}, X_{-n}) \leq H(Z_0|X_{-n}) = h_b(\pi^{*n} * \delta), \tag{51}$$

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-hand side of (49) of order ε^2, failing to provide the true ε^4 order in (49) (and, a fortiori, its refinements we derive below).

Let us now take one step of refinement beyond Corollary 4.8, to study the form of the bounds of Observation 4.7 in the case $M = 2$, and their implications in some asymptotic regimes. Define, in addition to A and b in (44) and (46),

$$a = -f(b) + \log \frac{1-\delta}{\delta} \tag{52}$$

and

$$B = f(b) + \log \frac{1-\delta}{\delta}. \tag{53}$$

Lemma 4.9. *Assume that 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 that $\delta \leq \frac{1}{2}\left(1 - \sqrt{\max\{1-4\pi, 0\}}\right)$ (cf. Figure 4). Then A, b, a, and B (defined in (45), (46), (52), and (53)) 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 in $[-A, -B] \cup [-a, -b] \cup [b, a] \cup [B, A]$.*

Proof. That the A solving (44) is the upper end point of the support of Y_i and, by symmetry, $-A$ its lower end point, is evident from (24). It was shown in [29, 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 $b = \log \frac{1-\delta}{\delta} - f(A)$. 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-hand side of (24) is negative, in which case the most (least negative) it can be is $-f(b)$, implying that in this case $Y_i \leq \log \frac{1-\delta}{\delta} - f(b) = a$. The second possibility is that this second term is positive, in which case the least it can be is $f(b)$, implying that $Y_i \geq \log \frac{1-\delta}{\delta} + f(b) = B$. 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 $s_i f(Y_{i-1}) < 0$, i.e.,

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

Using similar reasoning gives

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

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

and

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

□

Specializing Observation 4.7 to the case $M = 2$ and combining with Lemma 4.9 gives the following lemma.

Lemma 4.10. *For all* $\delta \le \frac{1}{2}\left(1 - \sqrt{\max\{1-4\pi, 0\}}\right)$,

$$\begin{aligned} &\{(1-\delta)[\pi * (1-\delta)] + \delta[\pi * \delta]\} h_b\left(\frac{e^A}{1+e^A} * \pi * \delta\right) \\ &\quad + \{(1-\delta)[\pi * \delta] + \delta[\pi * (1-\delta)]\} h_b\left(\frac{e^a}{1+e^a} * \pi * \delta\right) \\ &\le \overline{H}(Z) \\ &\le \{(1-\delta)[\pi * (1-\delta)] + \delta[\pi * \delta]\} h_b\left(\frac{e^B}{1+e^B} * \pi * \delta\right) \\ &\quad + \{(1-\delta)[\pi * \delta] + \delta[\pi * (1-\delta)]\} h_b\left(\frac{e^b}{1+e^b} * \pi * \delta\right), \end{aligned} \tag{54}$$

where $A, B, a,$ *and* b *are as specified in (45), (46), (52), and (53).*

As can be expected, the bounds in Lemma 4.10, which are based on the support bound depicted in Figure 3, are considerably tighter, in various asymptotic regimes, than those based on Corollary 4.8, which use the coarser support bound of Figure 2. As a first example, recall that in the "high SNR" regime the analysis in [29, Section 5], which was based on Corollary 4.8, established that

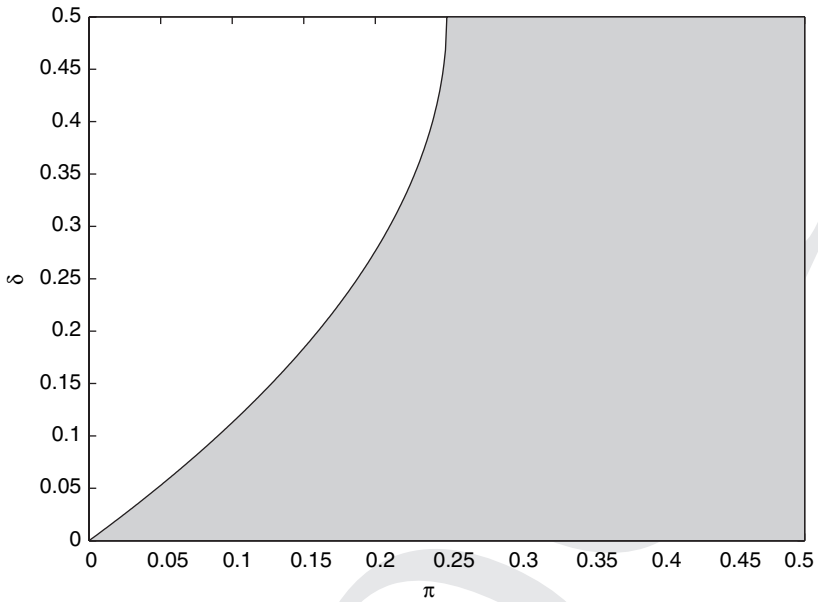

Figure 4 Shaded area below the curve $\frac{1}{2}\left(1-\sqrt{\max\{1-4\pi,0\}}\right)$ is the region in the $\pi-\delta$ plane where the sign of r_i determines the sign of Y_i in (24).

$\overline{H}(Z)-h_b(\pi)\asymp\delta$ (recall (47)), while, as we now show, the bounds of Lemma 4.10 recover the constant.

Theorem 4.11. *For $\pi\leq 1/2$ and $\delta\downarrow 0$,*

$$\overline{H}(Z)=h_b(\pi)+\left[2(1-2\pi)\log_2\frac{1-\pi}{\pi}\right]\cdot\delta+o(\delta).$$

The result of Theorem 4.11 was first established in [22], and subsequently derived in [27] and [36]. We give a simple proof of this result in Appendix A, via the bounds of Lemma 4.10.

For the "almost memoryless" regime, the bounds of Lemma 4.10 are tight enough to imply that the term following that characterized in (48) is $o(\varepsilon^3)$. More specifically, by evaluating the bounds of Lemma 4.10 for this regime, the following theorem is proved in Appendix B.

Theorem 4.12. *For $0\leq\delta\leq 1/2$ and $\pi=1/2-\varepsilon$, as $\varepsilon\downarrow 0$,*

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

For the "low SNR" regime, the bounds of Lemma 4.10 are shown in Appendix C to imply the following theorem.

Theorem 4.13. *For* $1/4 \le \pi \le 1/2$ *and* $\delta = \frac{1}{2} - \varepsilon$,

$$\begin{aligned}\frac{2(1-2\pi)^2(1-12\pi+48\pi^2-64\pi^3+32\pi^4)}{\pi^2\log 2} &\le \liminf_{\varepsilon\to 0}\frac{1-\overline{H}(Z)}{\varepsilon^4}\\ &\le \limsup_{\varepsilon\to 0}\frac{1-\overline{H}(Z)}{\varepsilon^4}\\ &\le \frac{2(1-2\pi)^2(1-4\pi+16\pi^2-32\pi^3+32\pi^4)}{\pi^2\log 2}.\end{aligned} \tag{55}$$

To see why Theorem 4.13 covers only the range $1/4 \le \pi \le 1/2$, note that Lemma 4.10, on which it relies, applies only in the shaded region of Figure 4. Clearly, for $\pi < 1/4$ and $\delta = 1/2 - \varepsilon$, the point (π, δ) will be outside the shaded region for ε small enough.

Theorem 4.13 should be compared with Corollary 6 of [29], which gives

$$\begin{aligned}\frac{2}{\log 2}\left[\frac{(4\pi-1)(1-2\pi)}{\pi}\right]^2 \le \liminf_{\varepsilon\to 0}\frac{1-\overline{H}(Z)}{\varepsilon^4} &\le \limsup_{\varepsilon\to 0}\frac{1-\overline{H}(Z)}{\varepsilon^4}\\ &\le \frac{2}{\log 2}\left[\frac{1-2\pi}{\pi}\right]^2\end{aligned} \tag{56}$$

(and on the basis of which (49) was stated). Figure 5 plots the bounds of (56) (the lower and upper curves) and those of Theorem 4.13 (the two internal curves), as a function of π. The bounds become increasingly tight as π approaches $1/2$, all converging to 0. Furthermore, both lower and upper bounds in (56) (and, a fortiori, in (55)) behave as $\sim \frac{8}{\log 2}(1-2\pi)^2$ for $\pi \to 1/2$, implying that $1-\overline{H}(Z) \approx \frac{32}{\log 2}(1/2-\delta)^4(1/2-\pi)^2$ for π and δ close to $1/2$.

Theorems 4.11 through 4.13 were obtained via evaluation of the bounds of Lemma 4.10 in the respective regimes. In turn, Lemma 4.10 is nothing but a specialization of Observation 4.7 to the case $M = 2$, optimizing over the choice of constants a_k and b_k. It was seen that these constants define a region of the form depicted in Figure 3, and optimizing their values amounts to finding the smallest region of that form containing the support of Y_i (the alternative Markov process). This optimization was easy to do (in Lemma 4.9), by observation of the dynamics of the process $\{Y_i\}$, as given in Theorem 4.3.

Evidently, moving from the bounds corresponding to $M = 1$ (which lead to (47) – (49)) to those of $M = 2$ results, for various asymptotic regimes, in characterization of higher order terms, and refinement of constants. The larger

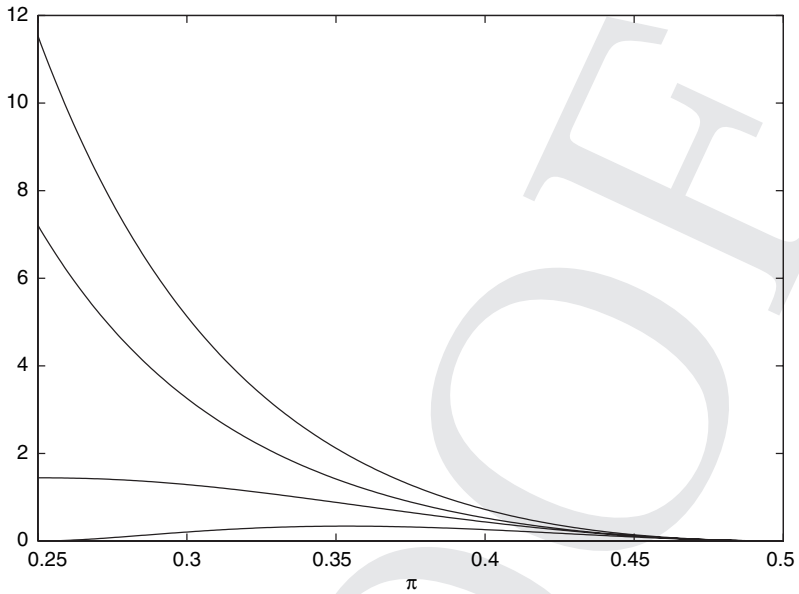

Figure 5 Upper and lower bounds associated with (49) and Theorem 4.13.

M one takes, the finer will the bounds become, leading, in particular, to finer characterizations in the respective 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 of Figure 3. The choice of these subintervals will be optimized analogously as in Lemma 4.9 quite simply, via the dynamics of the process $\{Y_i\}$ in (24). Note that it is only the end points of the new subintervals that need be computed, the remaining end points 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 Mth level. Only the end points of the subintervals that are being discarded need be computed, the remaining ones coinciding with those already obtained in the previous stage.

4 Non-symmetric case

In this section, we illustrate the use of the process (U_i, V_i) (defined via (31) and (32)) for obtaining bounds on the entropy rate in the nonsymmetric case. More

specifically, we use the dynamics of the process (U_i, V_i) detailed in Theorem 4.5 to obtain bounds on the support of (U_i, V_i), which, in turn, we translate to bounds on the expression for the entropy rate given in Corollary 4.6. In particular, paralleling the previous section, we develop bounds corresponding to Observation 4.7 in the case $M = 2$. As a representative example, we use this to obtain the first term in the expansion of the entropy rate in the "high SNR" regime, with the implication that expansions in other regimes, as well as higher-order terms (by using $M > 2$), are obtained similarly.

Bounds for regimes ("high SNR", "almost memoryless", "low SNR") mentioned in the previous section that were obtained via an approximation corresponding to $M = 1$ in [29] were obtained also for the nonsymmetric case. Additional regimes arising in the nonsymmetric setting include the "rare-spikes" and "rare-bursts" regimes. For example, it was shown (see [29, Theorem 5]) that for $0 \le \delta \le 1/2$ and any function $a(\cdot)$ satisfying $0 < a(\varepsilon) \le \varepsilon$, as $\varepsilon \to 0$,

$$\frac{\overline{H}(a(\varepsilon), 1-\varepsilon, \delta) - h_b(\delta)}{a(\varepsilon)} = \frac{\overline{H}(1-\varepsilon, a(\varepsilon), \delta) - h_b(\delta)}{a(\varepsilon)} \sim (1-2\delta)\log_2 \frac{1-\delta}{\delta}. \tag{57}$$

The bounds we develop below are applicable, e.g., also for refining the characterization in (57).

4.1 The case $\pi_{10} = 1$

The first example we consider is the case where $\pi_{10} = 1$, in the "high SNR" regime. We will establish the following theorem.

Theorem 4.14. *For $\pi_{10} = 1$, $0 \le \pi_{01} < 1$, and δ tending to 0,*

$$\overline{H}(Z) = \overline{H}(X) + \frac{\pi_{01}(2-\pi_{01})}{1+\pi_{01}} \delta \log_2 \frac{1}{\delta} + O(\delta). \tag{58}$$

Interestingly, the first term in the expansion is of order $\delta \log_2 \frac{1}{\delta}$, in contrast to that in Theorem 4.11 which is of order δ. As was first shown in [22], and we show in the next subsection in detail, the order of δ behavior in fact reigns for all values of the pair (π_{10}, π_{01}), except when one of the two values equals 1 (in which case Theorem 4.14 asserts that the order is $\delta \log_2 \frac{1}{\delta}$). This case is left unresolved by the asymptotic expansion of [22], which only hints at the above behavior in that the constant multiplying the order δ term increases to infinity as either π_{01} or π_{10} tends to one. A variation on the (second) proof of Theorem 4.14 appearing below is shown in the next subsection to also recover the expansion of [22] 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 $\overline{H}(Z) = \overline{H}(X) + \delta \log_2 \frac{1}{\delta} + O(\delta)$, where the factor multiplying the $\delta \log_2 \frac{1}{\delta}$ term in (58) is $1/2$ when $\pi_{01} = 1$. The reason for this is that just like there is a transition from order of δ to order of $\delta \log_2 1/\delta$ when going from $\pi_{10} < 1$ to $\pi_{10} = 1$, a similar term that will be order of δ in our analysis below (where we assume that $\pi_{01} < 1$) becomes order of $\delta \log_2 1/\delta$ when going from $\pi_{01} < 1$ to $\pi_{01} = 1$. This accounts for the doubling of the said factor from $1/2$ to 1.

As it turns out, Theorem 4.14 is provable via the Cover and Thomas bounds [5, Section 4.5] of order $n = 2$. We detail this proof in Appendix D. In this subsection, we give an alternative proof via our support bounds approach. Throughout the remainder of this subsection, we assume that $\pi_{10} = 1$ and $\pi_{01} < 1$, in which case $f(x)$ simplifies to

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

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

$$r(x) = \log \frac{1-x}{x}, \tag{59}$$

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

In the spirit of the developments in the previous section for bounding the support in the case $M = 2$, considering the alternative process constructed in Theorem 4.5, we will show that the support of $\mathcal{L}(U) = \mathcal{L}(l_i | X_i = 0)$ (and $\mathcal{L}(Y) = \mathcal{L}(l_i | X_i = 1)$, 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 4.10. The bounds thus obtained will be shown to lead to the asymptotic behavior of the entropy rate stated in Theorem 4.14.

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

Lemma 4.15. *Suppose that* $p = p_0 + \delta p_1 + O(\delta^2)$. *If* $0 < p_0\overline{\pi_{10}} + \overline{p_0}\pi_{01} < 1$, *then*

$$\begin{aligned} &h_b([p\overline{\pi_{10}} + \bar{p}\pi_{01}] * \delta) = h_b(p_0\overline{\pi_{10}} + \overline{p_0}\pi_{01}) \\ &-\delta \left[(\overline{p_0}(1-2\pi_{01}) + p_0(2\pi_{10}-1) + p_1(1-\pi_{01}-\pi_{10})) \log_2 \frac{p_0\overline{\pi_{10}} + \overline{p_0}\pi_{01}}{p_0\pi_{10} + \overline{p_0}\overline{\pi_{01}}} \right] \\ &+O(\delta^2). \end{aligned}$$

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

$$h_b([p\overline{\pi_{10}}+\bar{p}\pi_{01}]*\delta)=(1-p_1\pi_{01})\delta\log_2\frac{1}{\delta}+O(\delta). \tag{60}$$

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

$$\begin{aligned}
I_0&=[-r(\delta)+f(r(\delta)-r(\pi_{01})),-r(\delta)+f(r(\delta)+f(r(\delta)-r(\pi_{01})))],\\
I_1&=[-r(\delta)+f(-r(\delta)-r(\pi_{01})),-r(\delta)-r(\pi_{01})],\\
I_2&=[r(\delta)+f(r(\delta)-r(\pi_{01})),r(\delta)+f(r(\delta)+f(r(\delta)-r(\pi_{01})))],\\
I_3&=[r(\delta)+f(-r(\delta)-r(\pi_{01})),r(\delta)-r(\pi_{01})].
\end{aligned} \tag{61}$$

We shall also rely on the following three lemmas. Their proofs, which we defer to Appendix E, are based on ideas similar to those used in the proof of Lemma 4.9.

Lemma 4.16. *The intervals* $I_j, j=0,1,2,3$*, are nonempty (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 4.17. *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 4.18. *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 sufficently 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	δ^2	$\bar{\delta}(\pi_{01}*\delta)$
I_1	$\delta\bar{\delta}$	$\bar{\delta}(\pi_{01}*\bar{\delta})$
I_2	$\delta\bar{\delta}$	$\delta(\pi_{01}*\delta)$
I_3	$\bar{\delta}^2$	$\delta(\pi_{01}*\bar{\delta})$

(62)

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

$$\overline{H}(Z) \le \sum_{j=0}^{3} \left[\frac{1}{1+\pi_{01}} P_U(I_j) + \frac{\pi_{01}}{1+\pi_{01}} P_Y(I_j) \right] \max_{x \in I_j} h_b([\overline{\beta(x)}\pi_{01}] * \delta) \tag{63}$$

and

$$\overline{H}(Z) \ge \sum_{j=0}^{3} \left[\frac{1}{1+\pi_{01}} P_U(I_j) + \frac{\pi_{01}}{1+\pi_{01}} P_Y(I_j) \right] \min_{x \in I_j} h_b([\overline{\beta(x)}\pi_{01}] * \delta). \tag{64}$$

Proof.

$$\overline{H}(Z) \ge \sum_{j=0}^{3} \left[\frac{1}{1+\pi_{01}} Pr(l_i \in I_j | X_i = 0) + \frac{\pi_{01}}{1+\pi_{01}} Pr(l_i \in I_j | X_i = 1) \right] \min_{x \in I_j} h_b([\overline{\beta(x)}\pi_{01}] * \delta)$$

follows from (42) and Lemma 4.18, once δ is sufficiently small for Lemma 4.17 to imply that the I_j are disjoint. The upper bound follows similarly. □

Proof of Theorem 4.14. Lemma 4.18 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 (63) and (64) 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 at either $\ell(I_j)$ 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 δ. In particular, they will be of the form $p_0 + \delta p_1 + O(\delta^2)$ with $p_0 = 0$. Therefore, using Lemmas 4.15 and 4.18,

$$\begin{aligned} &\sum_{j=0}^{1} \frac{1}{1+\pi_{01}} P_U(I_j) \max_{x \in I_j} h_b([\overline{\beta(x)}\pi_{01}] * \delta) \\ &= \frac{\pi_{01}}{1+\pi_{01}} h_b(\pi_{01}) + \frac{\overline{\pi_{01}}}{1+\pi_{01}} h_b(\pi_{01}) + O(\delta) \\ &= \overline{H}(X) + O(\delta). \end{aligned} \tag{65}$$

Similarly,

$$\sum_{j=0}^{1} \frac{1}{1+\pi_{01}} P_U(I_j) \min_{x\in I_j} h_b([\overline{\beta(x)}\pi_{01}]*\delta) = \overline{H}(X) + O(\delta). \tag{66}$$

Next, we focus on the terms involving $P_Y(I_3)$. In these cases, the above properties (concavity and extremal) of $h_b([\bar{p}\pi_{01}]*\delta)$ 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 (61), we see (some algebraic manipulations omitted) that

$$\beta(u(I_3)) = \frac{\bar{\delta}\pi_{01}}{\delta\overline{\pi_{01}} + \bar{\delta}\pi_{01}} \tag{67}$$

$$= 1 - \frac{\delta\overline{\pi_{01}}}{\delta\overline{\pi_{01}} + \bar{\delta}\pi_{01}} \tag{68}$$

$$= 1 - \delta\frac{\overline{\pi_{01}}}{\pi_{01}} + O(\delta^2) \tag{69}$$

and

$$\beta(\ell(I_3)) = \frac{\bar{\delta}^2\pi_{01}\overline{\pi_{01}}}{\pi_{01}^2\delta^2 + \bar{\delta}\delta\overline{\pi_{01}}^2 + \bar{\delta}^2\pi_{01}\overline{\pi_{01}}} \tag{70}$$

$$= 1 - \frac{\pi_{01}^2\delta^2 + \bar{\delta}\delta\overline{\pi_{01}}^2}{\pi_{01}^2\delta^2 + \bar{\delta}\delta\overline{\pi_{01}}^2 + \bar{\delta}^2\pi_{01}\overline{\pi_{01}}} \tag{71}$$

$$= 1 - \delta\frac{\overline{\pi_{01}}}{\pi_{01}} + O(\delta^2). \tag{72}$$

Lemma 4.15, (69) and (72) then imply that

$$\min_{x\in I_3} h_b([\overline{\beta(x)}\pi_{01}]*\delta) = (1+\overline{\pi_{01}})\delta\log_2\frac{1}{\delta} + O(\delta) \tag{73}$$

and

$$\max_{x\in I_3} h_b([\overline{\beta(x)}\pi_{01}]*\delta) = (1+\overline{\pi_{01}})\delta\log_2\frac{1}{\delta} + O(\delta). \tag{74}$$

Equations (65), (66), (73), (74), and the expression for $P_Y(I_3)$ from (62) demonstrate that the combined contribution of the terms involving $P_Y(I_3)$, $P_U(I_0)$, and $P_U(I_1)$ to (63) and (64) is

$$\begin{aligned}&\overline{H}(X)+\frac{\pi_{01}}{1+\pi_{01}}(1+\overline{\pi_{01}})\delta\log_2\frac{1}{\delta}+O(\delta)\\&=\overline{H}(X)+\frac{\pi_{01}(2-\pi_{01})}{1+\pi_{01}}\delta\log_2\frac{1}{\delta}+O(\delta)\end{aligned}$$

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

4.2 The case $0<\pi_{01},\pi_{10}<1$

The analysis of this case is dependent on whether $\pi_{01}+\pi_{10}$ is smaller or greater than 1. If $\pi_{01}+\pi_{10}\le 1$, then, as shown in Section 2, $f(x)$ (defined in (17)) is nondecreasing and is upper bounded by $f(\infty)=r(\pi_{10})$ and lower bounded by $f(-\infty)=-r(\pi_{01})$, where $r(x)$ is defined in (59). If, on the other hand, $\pi_{01}+\pi_{10}>1$, then $f(x)$ is decreasing and is upper bounded by $f(-\infty)=-r(\pi_{01})$ and lower bounded by $f(\infty)=r(\pi_{10})$.

We define the four intervals J_j, $j=0,\ldots,3$, as

$$J_0=[-r(\delta)-r(\pi_{01}),-r(\delta)+f(-r(\delta)+r(\pi_{10}))], \tag{75}$$

$$J_1=[-r(\delta)+f(r(\delta)-r(\pi_{01})),-r(\delta)+r(\pi_{10})], \tag{76}$$

$$J_2=[r(\delta)-r(\pi_{01}),r(\delta)+f(-r(\delta)+r(\pi_{10}))], \tag{77}$$

$$J_3=[r(\delta)+f(r(\delta)-r(\pi_{01})),r(\delta)+r(\pi_{10})] \tag{78}$$

and the four intervals K_j as

$$K_0=[-r(\delta)+r(\pi_{10}),-r(\delta)+f(r(\delta)+r(\pi_{10}))], \tag{79}$$

$$K_1=[-r(\delta)+f(-r(\delta)-r(\pi_{01})),-r(\delta)-r(\pi_{01})], \tag{80}$$

$$K_2=[r(\delta)+r(\pi_{10}),r(\delta)+f(r(\delta)+r(\pi_{10}))], \tag{81}$$

$$K_3=[r(\delta)+f(-r(\delta)-r(\pi_{01})),r(\delta)-r(\pi_{01})]. \tag{82}$$

As before, the intervals $\{J_j\}$ and $\{K_j\}$ are respectively disjoint for sufficiently small δ when $\pi_{01}+\pi_{10}\le 1$ and $\pi_{01}+\pi_{10}\ge 1$. Additionally, the following analogue of Lemma 4.18 holds.

Lemma 4.20. *If $\pi_{01}+\pi_{10}\le 1$, the supports of both P_U and P_Y are contained in $J_0\cup J_1\cup J_2\cup J_3$. For all sufficiently small $\delta>0$, the probabilities of the intervals*

$\{J_j\}$ under P_U and P_Y are given by

I	$P_Y(I)$	$P_U(I)$
J_0	$\delta(\pi_{10} * \delta)$	$\bar{\delta}(\pi_{01} * \bar{\delta})$
J_1	$\delta(\pi_{10} * \bar{\delta})$	$\bar{\delta}(\pi_{01} * \delta)$
J_2	$\bar{\delta}(\pi_{10} * \delta)$	$\delta(\pi_{01} * \bar{\delta})$
J_3	$\bar{\delta}(\pi_{10} * \bar{\delta})$	$\delta(\pi_{01} * \delta)$

(83)

If $\pi_{01} + \pi_{10} > 1$, the supports of both P_U and P_Y are contained in $K_0 \cup K_1 \cup K_2 \cup K_3$. For all sufficiently small $\delta > 0$, the probabilities of the intervals $\{K_j\}$ under P_U and P_Y are given by

I	$P_Y(I)$	$P_U(I)$
K_0	$\delta(\pi_{10} * \bar{\delta})$	$\bar{\delta}(\pi_{01} * \delta)$
K_1	$\delta(\pi_{10} * \delta)$	$\bar{\delta}(\pi_{01} * \bar{\delta})$
K_2	$\bar{\delta}(\pi_{10} * \bar{\delta})$	$\delta(\pi_{01} * \delta)$
K_3	$\bar{\delta}(\pi_{10} * \delta)$	$\delta(\pi_{01} * \bar{\delta})$

(84)

The proof of Lemma 4.20 is similar to that of Lemma 4.18 with (A.64) replaced by

I	Y_i	U_i
J_0	$\{(r_i, s_i, r_{i-1}) = (-1,0,-1)\} \cup$ $\{(r_i, s_i, q_{i-1}) = (-1,1,-1)\}$	$\{(q_i, t_i, q_{i-1}) = (-1,0,-1)\} \cup$ $\{(q_i, t_i, r_{i-1}) = (-1,1,-1)\}$
J_1	$\{(r_i, s_i, r_{i-1}) = (-1,0,1)\} \cup$ $\{(r_i, s_i, q_{i-1}) = (-1,1,1)\}$	$\{(q_i, t_i, q_{i-1}) = (-1,0,1)\} \cup$ $\{(q_i, t_i, r_{i-1}) = (-1,1,1)\}$
J_2	$\{(r_i, s_i, r_{i-1}) = (1,0,-1)\} \cup$ $\{(r_i, s_i, q_{i-1}) = (1,1,-1)\}$	$\{(q_i, t_i, q_{i-1}) = (1,0,-1)\} \cup$ $\{(q_i, t_i, r_{i-1}) = (1,1,-1)\}$
J_3	$\{(r_i, s_i, r_{i-1}) = (1,0,1)\} \cup$ $\{(r_i, s_i, q_{i-1}) = (1,1,1)\}$	$\{(q_i, t_i, q_{i-1}) = (1,0,1)\} \cup$ $\{(q_i, t_i, r_{i-1}) = (1,1,1)\}$

(85)

for the case that $\pi_{01} + \pi_{10} \leq 1$. For the other case, the events for the intervals K_0, K_1, K_2, and K_3 coincide respectively with those for J_1, J_0, J_3, and J_2, given above.

Additionally, we have the following minor variation on Lemma 4.19.

Lemma 4.21. *For $\pi_{01}+\pi_{10}\le 1$ and all sufficiently small $\delta>0$, the entropy rate $\overline{H}(Z)$ of the process $\{Z_i\}$ satisfies*

$$\overline{H}(Z)\le\sum_{j=0}^{3}\left[\frac{\pi_{10}}{\pi_{01}+\pi_{10}}P_U(J_j)+\frac{\pi_{01}}{\pi_{01}+\pi_{10}}P_Y(J_j)\right]\times\max_{x\in J_j}h_b([\overline{\beta(x)}\pi_{01}+\beta(x)\overline{\pi_{10}}]*\delta) \tag{86}$$

and

$$\overline{H}(Z)\ge\sum_{j=0}^{3}\left[\frac{\pi_{10}}{\pi_{01}+\pi_{10}}P_U(J_j)+\frac{\pi_{01}}{\pi_{01}+\pi_{10}}P_Y(J_j)\right]\times\min_{x\in J_j}h_b([\overline{\beta(x)}\pi_{01}+\beta(x)\overline{\pi_{10}}]*\delta). \tag{87}$$

For $\pi_{01}+\pi_{10}>1$ and all sufficiently small $\delta>0$, the entropy rate satisfies (86) and (87) with J_j replaced by K_j.

Theorem 4.22. *For $0<\pi_{01},\pi_{10}<1$ and δ tending to 0,*

$$\overline{H}(Z)=\overline{H}(X)+\delta\left(\frac{1}{\pi_{01}+\pi_{10}}\left[(\pi_{01}+\pi_{10}-4\pi_{01}\pi_{10})\log_2\frac{\overline{\pi_{01}}\,\overline{\pi_{10}}}{\pi_{01}\pi_{10}}+(\pi_{10}-\pi_{01})\log_2\frac{\overline{\pi_{01}}}{\overline{\pi_{10}}}\right]\right)+O(\delta^2). \tag{88}$$

Note that (88) applies regardless of the value of $\pi_{01}+\pi_{10}$ even though our proof treats the cases when this sum is smaller or greater than one differently. The expression (88) was first obtained in [22] using a different technique. The factor multiplying δ can be shown to equal $D(P_{X_0,X_1,X_2}\|P_{X_0,\overline{X}_1,X_2})$ (where $D(P\|Q)$ denotes the relative entropy or Kullback–Leibler divergence between distributions P and Q), which is the form given in [22].

Proof of Theorem 4.22. When $0<\pi_{01},\pi_{10}<1$, straightforward calculus shows that the maximum of $h_b([\overline{p}\pi_{01}+p\overline{\pi_{10}}]*\delta)$ over p is bounded away from 0 and 1. Note also that both end points of J_0, J_1, K_0, and K_1 tend to $-\infty$ while the end points of J_2, J_3, K_2, and K_3 tend to ∞. Therefore, for δ sufficiently small, the concavity of $h_b([\overline{p}\pi_{01}+p\overline{\pi_{10}}]*\delta)$ in p implies that $\max_{x\in I}h_b([\overline{\beta(x)}\pi_{01}+\beta(x)\overline{\pi_{10}}]*\delta)$ is achieved at $x=u(I)$ for $I\in\{J_0,J_1,K_0,K_1\}$ and at $x=\ell(I)$ for

$I \in \{J_2, J_3, K_2, K_3\}$ and that $\min_{x \in I} h_b([\overline{\beta(x)}\pi_{01} + \beta(x)\overline{\pi_{10}}] * \delta)$ is achieved at $x = \ell(I)$ for $I \in \{J_0, J_1, K_0, K_1\}$ and at $x = u(I)$ for $I \in \{J_2, J_3, K_2, K_3\}$. Thus, by Lemma 4.21, for $\pi_{01} + \pi_{10} \le 1$,

$$\begin{aligned}
\overline{H}(Z) \le & \sum_{j=0}^{1} \left[\frac{\pi_{10}}{\pi_{01}+\pi_{10}} P_U(J_j) + \frac{\pi_{01}}{\pi_{01}+\pi_{10}} P_Y(J_j) \right] \\
& \times h_b([\overline{\beta(u(J_j))}\pi_{01} + \beta(u(J_j))\overline{\pi_{10}}] * \delta) \qquad (89) \\
& + \sum_{j=2}^{3} \left[\frac{\pi_{10}}{\pi_{01}+\pi_{10}} P_U(J_j) + \frac{\pi_{01}}{\pi_{01}+\pi_{10}} P_Y(J_j) \right] \\
& \times h_b([\overline{\beta(\ell(J_j))}\pi_{01} + \beta(\ell(J_j))\overline{\pi_{10}}] * \delta) \qquad (90)
\end{aligned}$$

and

$$\begin{aligned}
\overline{H}(Z) \ge & \sum_{j=0}^{1} \left[\frac{\pi_{10}}{\pi_{01}+\pi_{10}} P_U(J_j) + \frac{\pi_{01}}{\pi_{01}+\pi_{10}} P_Y(J_j) \right] \\
& \times h_b([\overline{\beta(\ell(J_j))}\pi_{01} + \beta(\ell(J_j))\overline{\pi_{10}}] * \delta) \qquad (91) \\
& + \sum_{j=2}^{3} \left[\frac{\pi_{10}}{\pi_{01}+\pi_{10}} P_U(J_j) + \frac{\pi_{01}}{\pi_{01}+\pi_{10}} P_Y(J_j) \right] \\
& \times h_b([\overline{\beta(u(J_j))}\pi_{01} + \beta(u(J_j))\overline{\pi_{10}}] * \delta). \qquad (92)
\end{aligned}$$

Corresponding expressions hold for $\pi_{01} + \pi_{10} > 1$, with $\{K_j\}$ replacing $\{J_j\}$.

The next step is to show that $\beta(u(I)) - \beta(\ell(I)) = O(\delta^2)$ for I equal to each of $\{J_j\}$ and $\{K_j\}$ and to express $\beta(u(I))$ (and hence $\beta(\ell(I))$) using the asymptotic approximation $p_0 + \delta p_1 + O(\delta^2)$, where p_0 and p_1 depend on I. We give the details for $I = J_3$ with the other cases following similarly. For $I = J_3$, we have

$$\begin{aligned}
\beta(u(J_3)) &= \frac{\overline{\delta}\overline{\pi_{10}}}{\delta\pi_{10} + \overline{\delta}\overline{\pi_{10}}} && (93) \\
&= 1 - \frac{\delta\pi_{10}}{\delta\pi_{10} + \overline{\delta}\overline{\pi_{10}}} && (94) \\
&= 1 - \delta\frac{\pi_{10}}{\overline{\pi_{10}}} + O(\delta^2) && (95)
\end{aligned}$$

and

$$\beta(\ell(J_3)) = \frac{\overline{\delta}\delta\pi_{01} + \overline{\delta}^2(\overline{\pi_{10}}\pi_{01}/\overline{\pi_{01}})}{\delta^2\overline{\pi}_{01} + \delta\overline{\delta}(\pi_{10}\pi_{01}/\overline{\pi_{01}}) + \overline{\delta}\delta\pi_{01} + \overline{\delta}^2(\overline{\pi_{10}}\pi_{01}/\overline{\pi_{01}})} \tag{96}$$

$$= 1 - \frac{\delta^2\overline{\pi}_{01} + \delta\overline{\delta}(\pi_{10}\pi_{01}/\overline{\pi_{01}})}{\delta^2\overline{\pi}_{01} + \delta\overline{\delta}(\pi_{10}\pi_{01}/\overline{\pi_{01}}) + \overline{\delta}\delta\pi_{01} + \overline{\delta}^2(\overline{\pi_{10}}\pi_{01}/\overline{\pi_{01}})} \tag{97}$$

$$= 1 - \delta\frac{(\pi_{10}\pi_{01}/\overline{\pi_{01}})}{(\overline{\pi_{10}}\pi_{01}/\overline{\pi_{01}})} + O(\delta^2) \tag{98}$$

$$= 1 - \delta\frac{\pi_{10}}{\overline{\pi_{10}}} + O(\delta^2). \tag{99}$$

The asymptotic expressions for the intervals $\{J_j\}$ are given by (100). The expressions for K_0, K_1, K_2, and K_3 coincide respectively with those for J_1, J_0, J_3, and J_2.

I	p_0	p_1
J_0	0	$\pi_{01}/\overline{\pi_{01}}$
J_1	0	$\overline{\pi_{10}}/\pi_{10}$
J_2	1	$-\overline{\pi_{01}}/\pi_{01}$
J_3	1	$-\pi_{10}/\overline{\pi_{10}}$

(100)

The asymptotic expression (88) for the entropy rate is then obtained by substituting the expressions (100) into (90) and (91), invoking Lemma 4.15, substituting the interval probabilities (83) and (84), and combining terms. It is easy to see that the two cases $\pi_{01} + \pi_{10} \leq 1$ and $\pi_{01} + \pi_{10} > 1$ should result in the same expression, since the asymptotic expressions for the interval end points are permuted in the same manner as the interval probabilities. □

5 A Deterministic Approximation Algorithm

In this section, we present and analyze an entropy rate approximation scheme, which is based on approximating the stationary distribution of the alternative Markov processes constructed in Section 2. We remark that a somewhat similar scheme has previously been described, though not analyzed, in [32]. Throughout, "operations" refers to arithmetic operations.

5.1 The symmetric case

Assume without loss of generality that $\pi < 1/2$. Since $|f| \le \log(1-\pi)/\pi$, from (24) it is clear that the support of Y_i is contained in the interval

$$I_{\pi,\delta} \stackrel{\triangle}{=} \left[-\log \frac{(1-\pi)(1-\delta)}{\pi\delta}, \log \frac{(1-\pi)(1-\delta)}{\pi\delta} \right].$$

Let Q be an M-level quantizer of $I_{\pi,\delta}$ (i.e., a mapping from $I_{\pi,\delta}$ to M real numbers) with the property that

$$\max_{x \in I_{\pi,\delta}} |Q(x) - x| \le \varepsilon. \tag{101}$$

For example, a uniform quantizer of $I_{\pi,\delta}$ with $M \ge \frac{1}{\varepsilon} \log \frac{(1-\pi)(1-\delta)}{\pi\delta}$ levels has this property. Consider now the finite-state Markov process (with M states) evolving with the process in (24) according to

$$\tilde{Y}_i = Q\left(r_i \log \frac{1-\delta}{\delta} + s_i f(\tilde{Y}_{i-1}) \right) \tag{102}$$

and initiated (at time $i = 0$) with its stationary distribution (say independently of Y_0). Then

$$\begin{aligned}
|\tilde{Y}_i - Y_i| &\le \varepsilon + |f(\tilde{Y}_{i-1}) - f(Y_{i-1})| && (103)\\
&\le \varepsilon + (1-2\pi)|\tilde{Y}_{i-1} - Y_{i-1}| && (104)\\
&\le \varepsilon + (1-2\pi)[\varepsilon + |f(\tilde{Y}_{i-2}) - f(Y_{i-2})|] && (105)\\
&\vdots && (106)\\
&\le \varepsilon \sum_{j=0}^{i} (1-2\pi)^j |\tilde{Y}_0 - Y_0| && (107)\\
&\le \frac{\varepsilon}{2\pi} |\tilde{Y}_0 - Y_0| && (108)\\
&\le \varepsilon \frac{1}{\pi} \log \frac{(1-\pi)(1-\delta)}{\pi\delta} && (109)\\
&\le \varepsilon \frac{1}{\pi} \log \frac{1}{\pi\delta}, && (110)
\end{aligned}$$

where (104) follows from (23), and (108) from the fact that both $\tilde{Y}_0$ and Y_0 belong to the interval $I_{\pi,\delta}$. Let

$$\lambda_{\pi,\delta} = \max_{y \in I_{\pi,\delta}} \left| \frac{\partial h_b\left(\frac{e^y}{1+e^y} * \pi * \delta\right)}{\partial y} \right|.$$

Note that $\lambda_{\pi,\delta} < \infty$ when π or δ are bounded away from 0 and 1. Combining the fact that

$$\left| h_b\left(\frac{e^y}{1+e^y} * \pi * \delta\right) - h_b\left(\frac{e^{y'}}{1+e^{y'}} * \pi * \delta\right) \right| \leq \lambda_{\pi,\delta} |y - y'|$$

for $y, y' \in I_{\pi,\delta}$ with Corollary 4.4, we obtain

$$\left| \overline{H}(Z) - E h_b\left(\frac{e^{\tilde{Y}_i}}{1+e^{\tilde{Y}_i}} * \pi * \delta\right) \right| \leq \lambda_{\pi,\delta} \varepsilon \frac{1}{\pi} \log \frac{1}{\pi\delta}. \tag{111}$$

In particular, for the M-level uniform quantizer mentioned above, $\varepsilon \leq \frac{1}{M} \log \frac{(1-\pi)(1-\delta)}{\pi\delta}$, so (111) implies that

$$\left| \overline{H}(Z) - E h_b\left(\frac{e^{\tilde{Y}_i}}{1+e^{\tilde{Y}_i}} * \pi * \delta\right) \right| \leq \frac{1}{M} \frac{\lambda_{\pi,\delta}}{\pi} \left[\log \frac{1}{\pi\delta}\right]^2. \tag{112}$$

Thus, for a given precision ϵ we would need to take M such that $\frac{1}{M} \frac{\lambda_{\pi,\delta}}{\pi} \left[\log \frac{1}{\pi\delta}\right]^2 \leq \epsilon$, find the M-dimensional stationary distribution vector, and use it to compute $E h_b\left(\frac{e^{\tilde{Y}_i}}{1+e^{\tilde{Y}_i}} * \pi * \delta\right)$. More specifically, we have the following approximation algorithm.

Algorithm 4.25.

Input: M, π, δ

(1) Let Q denote the M-level uniform quantizer of the interval $I_{\pi,\delta}$ and $q_1, \ldots, q_M$ denote the quantization levels. Let P_M be the $M \times M$ stochastic matrix given by

$$\begin{aligned} P_M(i,j) = {} & [(1-\delta)(1-\pi)]\, 1\left(q_j = Q\left(\log \tfrac{1-\delta}{\delta} + f(q_i)\right)\right) + \\ & [\delta(1-\pi)]\, 1\left(q_j = Q\left(-\log \tfrac{1-\delta}{\delta} + f(q_i)\right)\right) + \\ & [(1-\delta)\pi]\, 1\left(q_j = Q\left(\log \tfrac{1-\delta}{\delta} - f(q_i)\right)\right) + \\ & [\delta\pi]\, 1\left(q_j = Q\left(-\log \tfrac{1-\delta}{\delta} - f(q_i)\right)\right), \end{aligned} \tag{113}$$

where $1(\cdot)$ is the indicator function of the condition in the argument (i.e., $1(\cdot)=1$ if the condition in the argument is true and $1(\cdot)=0$ otherwise).

(2) Compute the 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 the entropy estimate

$$\hat{H} = \sum_{i=1}^{M} \mathbf{a}_M(i) \cdot h_b\left(\frac{e^{q_i}}{1+e^{q_i}} * \pi * \delta\right). \tag{114}$$

Output: $\hat{H}$.

Note that $\hat{H}$ in (114) is nothing but the expression $Eh_b\left(\frac{e^{\tilde{Y}_i}}{1+e^{\tilde{Y}_i}} * \pi * \delta\right)$ that appears in (112), where $\{\tilde{Y}_i\}$ is the quantized process defined in (102) (initiated at its stationary distribution). One possible brute force method 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) [12]. Since the remaining steps in the algorithm require $O(M)$ operations, the overall number of operations required is $O(M^3)$. From (112), it follows that the resulting precision is $O\left(\frac{1}{M}\right)$. In summary, we have established the following theorem.

Theorem 4.24. *For fixed π, δ, Algorithm 4.25 requires $O(M^3)$ operations and guarantees precision of $O\left(\frac{1}{M}\right)$. In other words, N operations buy precision $O\left(N^{-1/3}\right)$.*

Theorem 4.24 was derived via a rather rough analysis. Two ingredients that are likely to significantly improve the bound on the approximation–precision tradeoff are:

(1) Using a nonuniform quantizer, with finer resolution near 0 (where f is least contractive) and coarser resolution towards the end points of the quantized interval (where f is highly contractive).

(2) The main part of the computational burden is finding the stationary distribution of the stochastic matrix P_M given in (113). The upper bound of $O(M^3)$ that was used on the number of operations that this requires holds for any $M \times M$ stochastic matrix. This does not use the particular structure of P_M, a very sparse matrix with the same four nonzero entries in each row.

Thus, a nonuniform quantization followed by an efficient procedure for finding the stationary distribution of P_M should result in an implementation of Algorithm 4.25 with higher precision than that guaranteed in the theorem. Theorem 4.24, however, suffices to make our main point, which is the independence of

the bound on the precision order on the process parameters (in this case π and δ). This should be contrasted with the hitherto best known precision–complexity tradeoff among deterministic approximation schemes obtained via (11). Specifically, the difference between the upper and lower bounds in (11), conveniently expressed as the mutual information $I(Z_0; X_{-n-1}|Z_{-n}^{-1})$, is known since [3] to decay exponentially with n. The best known bounds have been obtained in [19] and are of the form

$$I(Z_0; X_{-n-1}|Z_{-n}^{-1}) \leq C(\pi,\delta)\rho(\pi,\delta)^n, \tag{115}$$

where $C(\pi,\delta), \rho(\pi,\delta)$ are positive constants and $\rho(\pi,\delta) < 1$. On the other hand, the number of operations required to compute the Cover and Thomas bounds (11) 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 (115) $\rho(\pi,\delta)$ is arbitrarily close to 1 for appropriate values of the parameters.

5.2 The nonsymmetric case

Let us now derive a similar algorithm for the nonsymmetric chain. Parallelling the development of the previous section, the idea is to couple the process pair $\{(Y_i, U_i)\}$ with a quantized process pair $\{(\tilde{Y}_i, \tilde{U}_i)\}$ evolving as

$$\tilde{Y}_i = Q\left(r_i \log \frac{1-\delta}{\delta} + s_i f(\tilde{U}_{i-1}) + (1-s_i) f(\tilde{Y}_{i-1}) \right) \tag{116}$$

and

$$\tilde{U}_i = Q\left(q_i \log \frac{1-\delta}{\delta} + (1-t_i) f(\tilde{U}_{i-1}) + t_i f(\tilde{Y}_{i-1}) \right), \tag{117}$$

where $\{q_i\}$, $\{r_i\}$, $\{s_i\}$, and $\{t_i\}$ are as defined in Theorem 4.5. We have

$$|\tilde{Y}_i - Y_i| \leq \varepsilon + s_i |f(\tilde{U}_{i-1}) - f(U_{i-1})| + (1-s_i)|f(\tilde{Y}_{i-1}) - f(Y_{i-1})| \tag{118}$$

$$\leq \varepsilon + \max\{|f(\tilde{U}_{i-1}) - f(U_{i-1})|, |f(\tilde{Y}_{i-1}) - f(Y_{i-1})|\} \tag{119}$$

$$\leq \varepsilon + c(\pi_{01}, \pi_{10}) \max\{|\tilde{U}_{i-1} - U_{i-1}|, |\tilde{Y}_{i-1} - Y_{i-1}|\}, \tag{120}$$

with $c(\pi_{01}, \pi_{10})$ defined in (20). Since the right-hand side will similarly also bound $|\tilde{U}_i - U_i|$, we have

$$\max\{|\tilde{U}_i - U_i|, |\tilde{Y}_i - Y_i|\} \leq \varepsilon + c(\pi_{01}, \pi_{10}) \max\{|\tilde{U}_{i-1} - U_{i-1}|, |\tilde{Y}_{i-1} - Y_{i-1}|\}. \tag{121}$$

Iterating gives, similarly as in (110),

$$\max\{|\tilde{U}_i - U_i|, |\tilde{Y}_i - Y_i|\} \le \frac{\varepsilon}{1 - c(\pi_{01}, \pi_{10})} \times \max\{|\tilde{U}_0 - U_0|, |\tilde{Y}_0 - Y_0|\} \le \varepsilon \tilde{c}(\pi_{01}, \pi_{10}, \delta), \tag{122}$$

where $\tilde{c}(\pi_{01}, \pi_{10}, \delta) = \frac{1}{1-c(\pi_{01},\pi_{10})} \max_{x \in \text{support}(U_i) \cup \text{support}(Y_i)} |Q(x) - x|$. It follows similarly as in (111) that by letting $\hat{H}(\pi_{01}, \pi_{10}, \delta)$ denote the expectation on the right-hand side of (42), with $\tilde{Y}_i$ replacing Y and $\tilde{U}_i$ replacing U, if the quantizer used has resolution ε, then

$$\left| \overline{H}(Z) - \hat{H}(\pi_{01}, \pi_{10}, \delta) \right| \le O(\varepsilon). \tag{123}$$

Similarly as in the symmetric case, most of the burden in computing $\hat{H}(\pi_{01}, \pi_{10}, \delta)$ is in computing the stationary distribution for $(\tilde{Y}_i, \tilde{U}_i)$, which is a Markov chain with state space of size M^2 and transition kernel with the same 16 nonzero entries per line, regardless of (sufficiently large) M. Done brute force (without exploiting the structure of the transition matrix), this requires no more than $(M^2)^3 = M^6$ operations. Since $\varepsilon = O(1/M)$, we get by (123) precision $O\left(\frac{1}{M}\right)$ for $O(M^6)$ operations, or similarly as in Theorem 4.24, precision $O\left(\frac{1}{N^{1/6}}\right)$ for N operations. As in the previous subsection, the analysis can be refined to improve the order of this polynomial dependence. Beyond possible refinements that were already mentioned for the symmetric case, further simplification is possible for the nonsymmetric case by noting that rather than the joint distribution of $(\tilde{Y}_i, \tilde{U}_i)$, it is only its two marginals that are needed. The analysis given, however, suffices to make the main point, which is the independence of the order of the polynomial on the process parameters.

5.3 Larger alphabet sizes

Extending the above approximation algorithm and its analysis to larger state and observation alphabet sizes is nontrivial and we leave it for future work. Briefly, however, such an extension would first require an extension of the alternative Markov process, the components of which would be $(|\mathcal{X}| - 1)$-dimensional vectors, corresponding to the conditional probability distribution of the underlying state variables conditioned on current and past observations. The components of the alternative Markov process would then capture the

evolution of this conditional probability distribution conditioned on the possible values of the underlying state variable. An approximation algorithm could then be obtained by applying vector quantization to the vector-valued components of the process, in analogy to the scalar quantization step above. As noted for a similar algorithm in [32], the complexity of such an algorithm for a given approximation accuracy would scale rather adversely with the state alphabet size $|\mathcal{X}|$ since, by vector quantization theory (also rate distortion theory), the number of quantization points required for a given level of quantization error per dimension of the Markov process components increases exponentially with the dimension, which is $|\mathcal{X}| - 1$. The complexity scaling, however, is much more benign in the observation alphabet size $|\mathcal{Z}|$.

Interestingly, the reverse seems to be true for the Cover and Thomas approximation bounds. The complexity of computing these bounds for a given conditioning history n, assuming that observation sequence probabilities are computed efficiently using the well-known "forward" recursion, is easily seen to be proportional to $|\mathcal{Z}|^n(n|\mathcal{X}|^2)$. Thus, the required computation increases considerably faster with the observation alphabet size than with the state alphabet size, for even moderate n.

The above considerations suggest that, among deterministic algorithms, a quantization-based approach, along the lines we have presented, may be the best choice for approximating the entropy rate when $|\mathcal{X}|$ is small, while the Cover and Thomas bounds may be the better choice when $|\mathcal{X}|$ is large but $|\mathcal{Z}|$ is small. A rigorous comparison of the precision–complexity tradeoff of these two approaches for larger alphabet sizes is left for future work.

6 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 illustrated how the approach leads to characterization of the entropy rate in various asymptotic regimes. It was seen that a refinement of the bounding technique in [29], whereby the support is partitioned into a small number of nonoverlapping regions with easily computed probabilities, can lead to significantly tighter bounds and finer characterizations of the asymptotics. It was argued that the bounds derived can be further tightened by further refining this partition leading, in various asymptotic regimes, to characterization of higher-order terms. Finally, a deterministic algorithm for approximating the entropy rate of the HMP was derived. This scheme, based on approximating

the stationary distribution of the related Markov process, was shown to achieve the best known precision–complexity tradeoff.

Though focus was on the binary case, the approach developed for bounding the entropy rate, for asymptotic characterizations, and for approximation are applicable in the general finite-alphabet case.

Results for some of the asymptotic regimes (e.g., the "high SNR" regime when $\pi_{10} = 1$) were shown to be derivable also via the Cover and Thomas bounds [5, Section 4.5]. On the other hand, for other regimes (e.g., the "low SNR" one), our bounds were shown to yield precise characterizations of the asymptotics, while the Cover and Thomas bounds of arbitrarily high order were shown to fall short of implying such characterizations.

A key ingredient in the bounds developed 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 was seen to fall, 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, the belief falls, with probability one, in a small region of the simplex corresponding to very low certainty. In the "almost memoryless" regime, as a final example, the belief falls in a small region, concentrated near the belief of a "singlet filter" [6] (which in the binary case consists of two point masses).

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 processes of Subsection 2.3. One such example is the "rare transitions" regime[1] considered in [26], and recently conclusively characterized in [31]. As is argued in [26], this regime is another example of one whose asymptotics are not captured by the Cover and Thomas bounds of arbitrarily high order.

[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.

Appendices

A Proof of Theorem 4.11

We first argue that the bounds of Lemma 4.10 continue to hold (and are relaxed) when A, B, a, b are replaced, respectively, by $\tilde{A}, \tilde{B}, \tilde{a}, \tilde{b}$ that have the following simpler expressions:

$$\tilde{A} = \log \frac{1-\pi}{\pi} + \log \frac{1-\delta}{\delta}, \tag{A.1}$$

$$\tilde{b} = -f(\tilde{A}) + \log \frac{1-\delta}{\delta}, \tag{A.2}$$

$$\tilde{a} = -f(\tilde{b}) + \log \frac{1-\delta}{\delta}, \tag{A.3}$$

and

$$\tilde{B} = f(\tilde{b}) + \log \frac{1-\delta}{\delta}. \tag{A.4}$$

To see this, note that $\tilde{A} = f(\infty) + \log \frac{1-\delta}{\delta} \geq f(A) + \log \frac{1-\delta}{\delta} = A$, implying by the monotonicity of f that $\tilde{b} \leq b$, in turn implying that both $\tilde{a} \geq a$ and $\tilde{B} \leq B$. Combined with the decreasing monotonicity of $h_b\left(\frac{e^x}{1+e^x} * \pi * \delta\right)$ for $x > 0$, this implies that indeed substituting the tilded quantities in (54) increases the upper bound and decreases the lower bound. Noting now that

$$(1-\delta)[\pi * (1-\delta)] + \delta[\pi * \delta] = 1 - \pi - \delta(2-4\pi) + \delta^2 2(1-2\pi), \tag{A.5}$$

$$(1-\delta)[\pi * \delta] + \delta[\pi * (1-\delta)] = \pi + \delta(2-4\pi) - \delta^2 2(1-2\pi), \tag{A.6}$$

and

$$\frac{1}{1+e^{\tilde{A}}} \sim \frac{\pi}{1-\pi}\delta,$$

we obtain

$$\begin{aligned} h_b\left(\frac{e^{\tilde{A}}}{1+e^{\tilde{A}}} * \pi * \delta\right) &= h_b\left(\frac{1}{1+e^{\tilde{A}}} * \pi * \delta\right) \\ &= h_b\left(\left[\frac{\pi}{1-\pi}\delta(1+o(1))\right] * \pi * \delta\right) \end{aligned}$$

$$
\begin{aligned}
&= h_b\left(\left[\left(\frac{\pi}{1-\pi}+1\right)\delta(1+o(1))\right]*\pi\right)\\
&= h_b\left(\pi+\frac{1-2\pi}{1-\pi}\delta+o(\delta)\right)\\
&= h_b(\pi)+h_b'(\pi)\frac{1-2\pi}{1-\pi}\delta+o(\delta). \qquad \text{(A.7)}
\end{aligned}
$$

Similarly,

$$
\frac{1}{1+e^{\tilde{b}}}\sim\frac{1}{e^{\tilde{b}}}\sim\frac{e^{\tilde{A}}(1-\pi)+\pi}{e^{\tilde{A}}\pi+(1-\pi)}\frac{\delta}{1-\delta}\sim\frac{(1-\pi)}{\pi}\delta,
$$

$$
\frac{1}{1+e^{\tilde{a}}}\sim\frac{1}{e^{\tilde{a}}}\sim\frac{e^{\tilde{b}}(1-\pi)+\pi}{e^{\tilde{b}}\pi+(1-\pi)}\frac{\delta}{1-\delta}\sim\frac{(1-\pi)}{\pi}\delta,
$$

and

$$
\frac{1}{1+e^{\tilde{B}}}\sim\frac{1}{e^{\tilde{B}}}\sim\frac{e^{\tilde{b}}\pi+(1-\pi)}{e^{\tilde{b}}(1-\pi)+\pi}\frac{\delta}{1-\delta}\sim\frac{\pi}{1-\pi}\delta.
$$

Thus, we get also

$$
\begin{aligned}
h_b\left(\frac{e^{\tilde{a}}}{1+e^{\tilde{a}}}*\pi*\delta\right) &= h_b\left(\frac{1}{1+e^{\tilde{a}}}*\pi*\delta\right)\\
&= h_b\left(\left[\frac{1-\pi}{\pi}\delta(1+o(1))\right]*\pi*\delta\right)\\
&= h_b\left(\left[\frac{1}{\pi}\delta(1+o(1))\right]*\pi\right)\\
&= h_b\left(\pi+\frac{1-2\pi}{\pi}\delta+o(\delta)\right)\\
&= h_b(\pi)+h_b'(\pi)\frac{1-2\pi}{\pi}\delta+o(\delta) \qquad \text{(A.8)}
\end{aligned}
$$

and similarly obtain

$$
h_b\left(\frac{e^{\tilde{b}}}{1+e^{\tilde{b}}}*\pi*\delta\right)=h_b(\pi)+h_b'(\pi)\frac{1-2\pi}{\pi}\delta+o(\delta) \qquad \text{(A.9)}
$$

and

$$
h_b\left(\frac{e^{\tilde{B}}}{1+e^{\tilde{B}}}*\pi*\delta\right)=h_b(\pi)+h_b'(\pi)\frac{1-2\pi}{1-\pi}\delta+o(\delta). \qquad \text{(A.10)}
$$

Combining Lemma 4.10 (with $\tilde{A},\tilde{B},\tilde{a},\tilde{b}$ replacing A,B,a,b) with (A.5), (A.6), (A.7), (A.8), (A.9), and (A.10) gives

$$\begin{aligned}
\overline{H}(Z) &= [(1-\pi)-\delta(2-4\pi)+o(\delta)]\left[h_b(\pi)+h_b'(\pi)\frac{1-2\pi}{1-\pi}\delta+o(\delta)\right] \\
&\quad +[\pi+\delta(2-4\pi)+o(\delta)]\left[h_b(\pi)+h_b'(\pi)\frac{1-2\pi}{\pi}\delta+o(\delta)\right] \\
&= h_b(\pi)+\delta\big[h_b'(\pi)(1-2\pi)-(2-4\pi)h_b(\pi)+h_b'(\pi)(1-2\pi) \\
&\quad +(2-4\pi)h_b(\pi)\big]+o(\delta) \\
&= h_b(\pi)+\delta 2h_b'(\pi)(1-2\pi)+o(\delta) \\
&= h_b(\pi)+\delta 2(1-2\pi)\log_2\frac{1-\pi}{\pi}+o(\delta). \qquad \square
\end{aligned}$$

B Proof of Theorem 4.12

It is easily checked that in this regime $f\left(\log\frac{1-\delta}{\delta}+o(1)\right)=4\varepsilon(1-2\delta)+o(\varepsilon)$, implying that

$$A = f(A)+\log\frac{1-\delta}{\delta} = \log\frac{1-\delta}{\delta}+4\varepsilon(1-2\delta)+o(\varepsilon), \tag{A.11}$$

$$b = -f(A)+\log\frac{1-\delta}{\delta} = \log\frac{1-\delta}{\delta}-4\varepsilon(1-2\delta)+o(\varepsilon), \tag{A.12}$$

$$a = -f(b)+\log\frac{1-\delta}{\delta} = \log\frac{1-\delta}{\delta}-4\varepsilon(1-2\delta)+o(\varepsilon), \tag{A.13}$$

and

$$B = f(b)+\log\frac{1-\delta}{\delta} = \log\frac{1-\delta}{\delta}+4\varepsilon(1-2\delta)+o(\varepsilon). \tag{A.14}$$

It follows that

$$\frac{1}{1+e^A} = \delta+4\delta(1-\delta)(1-2\delta)\varepsilon+o(\varepsilon), \tag{A.15}$$

$$\frac{1}{1+e^B} = \delta+4\delta(1-\delta)(1-2\delta)\varepsilon+o(\varepsilon), \tag{A.16}$$

$$\frac{1}{1+e^a} = \delta-4\delta(1-\delta)(1-2\delta)\varepsilon+o(\varepsilon), \tag{A.17}$$

and

$$\frac{1}{1+e^b} = \delta-4\delta(1-\delta)(1-2\delta)\varepsilon+o(\varepsilon). \tag{A.18}$$

Now

$$(1-\delta)[\pi*(1-\delta)]+\delta[\pi*\delta] = (1-\delta)[(1/2-\varepsilon)*(1-\delta)]+\delta[(1/2-\varepsilon)*\delta]$$
$$= \frac{1}{2}+\varepsilon(1-2\delta)^2 \tag{A.19}$$

and

$$(1-\delta)[\pi*\delta]+\delta[\pi*(1-\delta)] = \frac{1}{2}-\varepsilon(1-2\delta)^2. \tag{A.20}$$

Now

$$\frac{1}{1+e^A}*\pi*\delta = [\delta+4\delta(1-\delta)(1-2\delta)\varepsilon+o(\varepsilon)]*\left(\frac{1}{2}-\varepsilon\right)*\delta$$
$$= \frac{1}{2}-\varepsilon(1-2\delta)^2+\varepsilon^2 8\delta(1-\delta)(1-2\delta)^2+o(\varepsilon^2) \tag{A.21}$$

and, similarly,

$$\frac{1}{1+e^B}*\pi*\delta = [\delta+4\delta(1-\delta)(1-2\delta)\varepsilon+o(\varepsilon)]*\left(\frac{1}{2}-\varepsilon\right)*\delta$$
$$= \frac{1}{2}-\varepsilon(1-2\delta)^2+\varepsilon^2 8\delta(1-\delta)(1-2\delta)^2+o(\varepsilon^2), \tag{A.22}$$
$$\frac{1}{1+e^a}*\pi*\delta = [\delta+4\delta(1-\delta)(1-2\delta)\varepsilon+o(\varepsilon)]*\left(\frac{1}{2}-\varepsilon\right)*\delta$$
$$= \frac{1}{2}-\varepsilon(1-2\delta)^2-\varepsilon^2 8\delta(1-\delta)(1-2\delta)^2+o(\varepsilon^2), \tag{A.23}$$

and

$$\frac{1}{1+e^b}*\pi*\delta = [\delta+4\delta(1-\delta)(1-2\delta)\varepsilon+o(\varepsilon)]*\left(\frac{1}{2}-\varepsilon\right)*\delta$$
$$= \frac{1}{2}-\varepsilon(1-2\delta)^2-\varepsilon^2 8\delta(1-\delta)(1-2\delta)^2+o(\varepsilon^2). \tag{A.24}$$

It follows now from Lemma 4.10, (A.49), the above displays and

$$h_b\left(\frac{1}{2}-\varepsilon\right)=1+\frac{1}{2}h_b''(1/2)\varepsilon^2+\frac{1}{4!}h_b^{(4)}(1/2)\varepsilon^4+O(\varepsilon^6)=1-\frac{2}{\log 2}\varepsilon^2+O(\varepsilon^4) \tag{A.25}$$

(note that h_b is symmetric around $1/2$, so odd-ordered terms annihilate) that

$$\begin{aligned}
\overline{H}(Z)=&\left[\frac{1}{2}+\varepsilon(1-2\delta)^2\right]\\
&\times h_b\left(\frac{1}{2}-\varepsilon(1-2\delta)^2+\varepsilon^2 8\delta(1-\delta)(1-2\delta)^2+o(\varepsilon^2)\right) &\text{(A.26)}\\
&+\left[\frac{1}{2}-\varepsilon(1-2\delta)^2\right]\\
&\times h_b\left(\frac{1}{2}-\varepsilon(1-2\delta)^2-\varepsilon^2 8\delta(1-\delta)(1-2\delta)^2+o(\varepsilon^2)\right) &\text{(A.27)}\\
=&\left[\frac{1}{2}+\varepsilon(1-2\delta)^2\right]\\
&\times\left\{1-\frac{2}{\log 2}\left[\varepsilon(1-2\delta)^2-\varepsilon^2 8\delta(1-\delta)(1-2\delta)^2\right]^2+o(\varepsilon^3)\right\} &\text{(A.28)}\\
&+\left[\frac{1}{2}-\varepsilon(1-2\delta)^2\right]\\
&\times\left\{1-\frac{2}{\log 2}\left[\varepsilon(1-2\delta)^2+\varepsilon^2 8\delta(1-\delta)(1-2\delta)^2\right]^2+o(\varepsilon^3)\right\}. &\text{(A.29)}
\end{aligned}$$

So,

$$\begin{aligned}
1-\overline{H}(Z)=&\frac{2}{\log 2}\left\{\left[\frac{1}{2}+\varepsilon(1-2\delta)^2\right]\right.\\
&\times\left[\varepsilon(1-2\delta)^2-\varepsilon^2 8\delta(1-\delta)(1-2\delta)^2\right]^2 &\text{(A.30)}\\
&\left.+\left[\frac{1}{2}-\varepsilon(1-2\delta)^2\right]\left[\varepsilon(1-2\delta)^2+\varepsilon^2 8\delta(1-\delta)(1-2\delta)^2\right]^2\right\}\\
&+o(\varepsilon^3) &\text{(A.31)}\\
=&\frac{2}{\log 2}\varepsilon^2(1-2\delta)^4+o(\varepsilon^3). &\text{(A.32)}
\end{aligned}$$

C Proof of Theorem 4.13

Proof. By (A.5) and (A.6),

$$(1-\delta)[\pi*(1-\delta)]+\delta[\pi*\delta] = 1-\pi-\left(\frac{1}{2}-\varepsilon\right)(2-4\pi)+\left(\frac{1}{2}-\varepsilon\right)^2$$

$$2(1-2\pi) = \frac{1}{2}+\varepsilon^2 2(1-2\pi) \tag{A.33}$$

and

$$(1-\delta)[\pi*\delta]+\delta[\pi*(1-\delta)] = \pi+\left(\frac{1}{2}-\varepsilon\right)(2-4\pi)-\left(\frac{1}{2}-\varepsilon\right)^2$$

$$2(1-2\pi) = \frac{1}{2}-\varepsilon^2 2(1-2\pi). \tag{A.34}$$

Now, recalling (45),

$$A = \log\left[(\alpha-1)\frac{1-\pi}{2\pi}+\sqrt{\alpha+\left[(\alpha-1)\frac{1-\pi}{2\pi}\right]^2}\right],$$

where $\alpha \stackrel{\triangle}{=} \frac{1-\delta}{\delta} = \frac{1+2\varepsilon}{1-2\varepsilon} = 1+4\varepsilon+o(\varepsilon)$, so

$$A = \log\left[(4\varepsilon+o(\varepsilon))\frac{1-\pi}{2\pi}+\sqrt{1+4\varepsilon+o(\varepsilon)+\left[(4\varepsilon+o(\varepsilon))\frac{1-\pi}{2\pi}\right]^2}\right] \tag{A.35}$$

$$= \log\left[(4\varepsilon+o(\varepsilon))\frac{1-\pi}{2\pi}+\sqrt{1+4\varepsilon+o(\varepsilon)}\right] \tag{A.36}$$

$$= \log\left[(4\varepsilon+o(\varepsilon))\frac{1-\pi}{2\pi}+1+2\varepsilon+o(\varepsilon)\right] \tag{A.37}$$

$$= \log\left[1+\varepsilon\left(4\cdot\frac{1-\pi}{2\pi}+2\right)+o(\varepsilon)\right] \tag{A.38}$$

$$= \log\left[1+\varepsilon\cdot\frac{2}{\pi}+o(\varepsilon)\right] \tag{A.39}$$

$$= \varepsilon\cdot\frac{2}{\pi}+o(\varepsilon), \tag{A.40}$$

implying that

$$\frac{1}{1+e^A} = \frac{1}{2+\varepsilon\cdot\frac{2}{\pi}+o(\varepsilon)} = \frac{1}{2}-\frac{1}{2\pi}\varepsilon+o(\varepsilon). \tag{A.41}$$

Now, recalling that $f(0)=0$ and using (19),

$$b=-f(A)+\log\frac{1-\delta}{\delta}=-f'(0)A+o(A)+\log[1+4\varepsilon+o(\varepsilon)]$$
$$=-(1-2\pi)\varepsilon\cdot\frac{2}{\pi}+4\varepsilon+o(\varepsilon)=(8-2/\pi)\varepsilon+o(\varepsilon), \tag{A.42}$$

implying that

$$\frac{1}{1+e^b}=\frac{1}{2+\varepsilon\cdot(8-2/\pi)+o(\varepsilon)}=\frac{1}{2}-\left(2-\frac{1}{2\pi}\right)\varepsilon+o(\varepsilon). \tag{A.43}$$

Moving to a, we have

$$a=-f(b)+\log\frac{1-\delta}{\delta}=-f'(0)b+o(b)+\log[1+4\varepsilon+o(\varepsilon)]$$
$$=-(1-2\pi)(8-2/\pi)\varepsilon+4\varepsilon+o(\varepsilon)=(16\pi+2/\pi-8)\varepsilon+o(\varepsilon), \tag{A.44}$$

implying that

$$\frac{1}{1+e^a}=\frac{1}{2+\varepsilon\cdot(16\pi+2/\pi-8)+o(\varepsilon)}=\frac{1}{2}-\left(4\pi+\frac{1}{2\pi}-2\right)\varepsilon+o(\varepsilon). \tag{A.45}$$

Finally,

$$B=f(b)+\log\frac{1-\delta}{\delta}=f'(0)b+o(b)+\log[1+4\varepsilon+o(\varepsilon)]$$
$$=(1-2\pi)(8-2/\pi)\varepsilon+4\varepsilon+o(\varepsilon)=(-16\pi-2/\pi+16)\varepsilon+o(\varepsilon), \tag{A.46}$$

implying that

$$\frac{1}{1+e^B}=\frac{1}{2+\varepsilon\cdot(-16\pi-2/\pi+16)+o(\varepsilon)}=\frac{1}{2}-\left(-4\pi-\frac{1}{2\pi}+4\right)\varepsilon+o(\varepsilon). \tag{A.47}$$

Using the easily verified identity

$$\left(\frac{1}{2}-c\varepsilon\right)*\pi*\left(\frac{1}{2}-\varepsilon\right)=\frac{1}{2}-2c(1-2\pi)\varepsilon^2, \tag{A.48}$$

the Taylor expansion

$$h_b\left(\frac{1}{2}-\varepsilon\right)=1+\frac{1}{2}h_b''(1/2)\varepsilon^2+o(\varepsilon^2)=1-\frac{2}{\log 2}\varepsilon^2+o(\varepsilon^2), \tag{A.49}$$

and combining with (A.33), (A.34), (A.41), and (A.45) gives

$$\begin{aligned}
&\{(1-\delta)[\pi*(1-\delta)]+\delta[\pi*\delta]\}h_b\left(\frac{e^A}{1+e^A}*\pi*\delta\right)\\
&+\{(1-\delta)[\pi*\delta]+\delta[\pi*(1-\delta)]\}h_b\left(\frac{e^a}{1+e^a}*\pi*\delta\right)\\
&=\left[\frac{1}{2}+\varepsilon^2 2(1-2\pi)\right]h_b\left(\frac{1}{2}-\frac{1-2\pi}{\pi}\varepsilon^2+o(\varepsilon^2)\right) &\text{(A.50)}\\
&\quad+\left[\frac{1}{2}-\varepsilon^2 2(1-2\pi)\right]\\
&\quad\times h_b\left(\frac{1}{2}-2\left(4\pi+\frac{1}{2\pi}-2\right)(1-2\pi)\varepsilon^2+o(\varepsilon^2)\right) &\text{(A.51)}\\
&=\left[\frac{1}{2}+\varepsilon^2 2(1-2\pi)\right]\left\{1-\frac{2}{\log 2}\left[\frac{1-2\pi}{\pi}\varepsilon^2\right]^2+o(\varepsilon^4)\right\}\\
&\quad+\left[\frac{1}{2}-\varepsilon^2 2(1-2\pi)\right] &\text{(A.52)}\\
&\quad\times\left\{1-\frac{2}{\log 2}\left[2\left(4\pi+\frac{1}{2\pi}-2\right)(1-2\pi)\varepsilon^2\right]^2+o(\varepsilon^4)\right\} &\text{(A.53)}\\
&=1-\frac{1}{\log 2}\left\{\left[\frac{1-2\pi}{\pi}\right]^2+\left[2\left(4\pi+\frac{1}{2\pi}-2\right)(1-2\pi)\right]^2\right\}\varepsilon^4+o(\varepsilon^4) &\text{(A.54)}\\
&=1-\frac{2(1-2\pi)^2(1-4\pi+16\pi^2-32\pi^3+32\pi^4)}{\pi^2\log 2}\varepsilon^4+o(\varepsilon^4). &\text{(A.55)}
\end{aligned}$$

Similarly, using (A.43) and (A.47) in lieu of (A.41) and (A.45), we obtain

$$\begin{aligned}
&\{(1-\delta)[\pi*(1-\delta)]+\delta[\pi*\delta]\}h_b\left(\frac{e^B}{1+e^B}*\pi*\delta\right)\\
&\quad+\{(1-\delta)[\pi*\delta]+\delta[\pi*(1-\delta)]\}h_b\left(\frac{e^b}{1+e^b}*\pi*\delta\right)\\
&\quad=\left[\frac{1}{2}+\varepsilon^2 2(1-2\pi)\right]\\
&\qquad\times\left\{1-\frac{2}{\log 2}\left[2\left(-4\pi-\frac{1}{2\pi}+4\right)(1-2\pi)\varepsilon^2\right]^2+o(\varepsilon^4)\right\} &\text{(A.56)}
\end{aligned}$$

$$+\left[\frac{1}{2}-\varepsilon^2 2(1-2\pi)\right]\left\{1-\frac{2}{\log 2}\left[2\left(2-\frac{1}{2\pi}\right)(1-2\pi)\varepsilon^2\right]^2+o(\varepsilon^4)\right\}$$

$$=1-\frac{1}{\log 2}\left\{\left[2\left(-4\pi-\frac{1}{2\pi}+4\right)(1-2\pi)\right]^2 \tag{A.57}$$

$$+\left[2\left(2-\frac{1}{2\pi}\right)(1-2\pi)\right]^2\right\}\varepsilon^4$$

$$+o(\varepsilon^4) \tag{A.58}$$

$$=1-\frac{2(1-2\pi)^2(1-12\pi+48\pi^2-64\pi^3+32\pi^4)}{\pi^2\log 2}\varepsilon^4+o(\varepsilon^4). \tag{A.59}$$

Combining (A.55) and (A.59) with Lemma 4.10 completes the proof. □

D Proof of Theorem 4.14 via the Cover and Thomas bounds

In this section, we show that the result of Subsection 4.1 can be obtained using the upper and lower bounds on the entropy rate of an HMM given in the book of Cover and Thomas. These upper and lower bounds are $H(Z_k|Z^k)$ and $H(Z_k|Z_2^k,X_1)$, respectively, and are valid for any k. We analyze the asymptotic behavior of these bounds for $k=3$ for the setting of Section 4.1, and show that the factor multiplying the $\delta\log_2 1/\delta$ term in both the lower and upper bounds agrees with that given in Theorem 4.14.

We first treat the upper bound $H(Z_3|Z_2,Z_1)$ and expand it as $H(Z_3,Z_2,Z_1)-H(Z_2,Z_1)$. In each resulting joint entropy the $-p(\cdot)\log_2 p(\cdot)$ terms contributing to the $\delta\log_2 1/\delta$ factor are those for which $p(z_3,z_2,z_1)$ and $p(z_2,z_1)$ tend to zero no faster than δ. The remaining $-p(\cdot)\log_2 p(\cdot)$ terms contribute to higher-order asymptotics and we ignore these. The $\Omega(\delta)$ probabilities arise from those sequences (z_3,z_2,z_1) and (z_2,z_1) that have zero probability under the Markov chain distribution $P_{X_3X_2X_1}$, and differ from at least one nonzero probability sequence, again under the Markov chain distribution, in precisely one position. The factor contributed to the $\delta\log_2 1/\delta$ term by any such zero probability sequence is then the probability, under the Markov chain distribution, of the set of sequences at Hamming distance 1. The sequences at Hamming distance 2 or greater contribute terms of $\delta^2\log_2 1/\delta$ or smaller.

Let $m_0=Pr(X_i=0)=\pi_{10}/(\pi_{01}+\pi_{10})$ and $m_1=Pr(X_i=1)=1-m_0$. For the $H(Z_2,Z_1)$ term, $(1,1)$ is the only zero probability sequence and the probability of the Hamming distance 1 sequences $(1,0)$ and $(0,1)$ is $m_0\pi_{01}+m_1$. For the $H(Z_3,Z_2,Z_1)$ term, there are three zero probability sequences $(1,1,1)$,

$(1,1,0)$, and $(0,1,1)$. For $(1,1,1)$, the only Hamming distance 1 sequence with non-zero probability is $(1,0,1)$ and its probability is $m_1\pi_{01}$. For $(1,1,0)$, the non-zero probability Hamming distance 1 sequences are $(0,1,0)$ and $(1,0,0)$ with a combined probability of $m_0\pi_{01}+m_0\overline{\pi_{01}}\pi_{01}$. Similarly, for $(0,1,1)$, the contributing sequences are $(0,1,0)$ and $(0,0,1)$ with a combined probability of $m_0\pi_{01}+m_1\overline{\pi_{01}}$. Thus, the overall factor multiplying the $\delta\log_2 1/\delta$ term is

$$\begin{aligned} &m_1\pi_{01}+m_0\pi_{01}+m_0\overline{\pi_{01}}\pi_{01}+m_0\pi_{01}+m_1\overline{\pi_{01}}-m_0\pi_{01}-m_1 \\ &\quad = m_0\overline{\pi_{01}}\pi_{01}+m_0\pi_{01} \qquad \text{(A.60)} \\ &\quad = \frac{\pi_{01}}{1+\pi_{01}}(2-\pi_{01}). \qquad \text{(A.61)} \end{aligned}$$

The lower bound $H(Z_3|Z_2,X_1)$ is similarly expanded to $H(Z_3,Z_2,X_1)-H(Z_2,X_1)$ and the two joint entropies are analyzed as above. In this case, the Hamming distance 1 sequences must differ from the zero probability sequences (z_3,z_2,x_1) and (z_2,x_1) only in the z_i positions. For the $H(Z_2,X_1)$ term, again $(1,1)$ is the only zero probability sequence, and the only allowed nonzero probability Hamming distance 1 sequence under the new restriction is $(0,1)$, the probability of which is m_1. For the $H(Z_3,Z_2,X_1)$ term, the three zero probability sequences are again $(1,1,1)$, $(1,1,0)$, and $(0,1,1)$. The contributions from $(1,1,1)$ and $(1,1,0)$ are as above. For $(0,1,1)$, the only contributing Hamming distance 1 sequence is $(0,0,1)$ (since the sequence $(0,1,0)$ differs in the x position), and its probability is $m_1\overline{\pi_{01}}$. The overall factor multiplying $\delta\log_2 1/\delta$ for the lower bounds is

$$\begin{aligned} &m_1\pi_{01}+m_0\pi_{01}+m_0\overline{\pi_{01}}\pi_{01}+m_1\overline{\pi_{01}}-m_1 = m_0\overline{\pi_{01}}\pi_{01}+m_0\pi_{01} \qquad \text{(A.62)} \\ &\quad \frac{\pi_{01}}{1+\pi_{01}}(2-\pi_{01}). \qquad \text{(A.63)} \end{aligned}$$

The claim of Theorem 4.14 then follows from the agreement of the upper and lower bound factors. □

E Proofs of lemmas used in proving theorem 4.14

Proof of Lemma 4.16. As noted, $-r(\pi_{01}) = f(-\infty) \geq f(r(\delta)-r(\pi_{01}))$. This fact together with the fact that $f(x)$ is decreasing shows that I_0 and I_2 are non-empty. The other two intervals are handled by similarly noting that $-r(\pi_{01}) \geq f(-r(\delta)-r(\pi_{01}))$. □

Proof of Lemma 4.17. For x large and positive the difference between $f(x)$ and $-x+\log\pi_{01}$ converges to 0. Therefore, δ tends to 0, the right end point

of I_0 behaves like $-r(\delta)+f(r(\pi_{01})+\log\pi_{01})=-r(\delta)+f(\log\overline{\pi_{01}})=-r(\delta)-r(\pi_{01})-\log 2$, while the left end point of I_1 behaves like $-r(\delta)-f(-\infty)=-r(\delta)-r(\pi_{01})$. It thus follows that $I_0<I_1$ for all sufficiently small $\delta>0$. The right end point of I_1 tends to $-\infty$, while the left end point of I_2, based on the preceding observations, tends to $r(\pi_{01})+\log\pi_{01}=\log\overline{\pi_{01}}$. Consequently, $I_1<I_2$ for all sufficiently small $\delta>0$. Finally, the right end point of I_2 (similarly to the right end point of I_0) behaves like $r(\delta)-r(\pi_{01})-\log 2$, while the left end point of I_3 behaves like $r(\delta)-r(\pi_{01})$, implying here as well that $I_2<I_3$ for all sufficiently small $\delta>0$. □

Proof of Lemma 4.18. As a consequence of the above upper bound on $f(x)$, it follows from the transitions of $\{(U_i,Y_i)\}$ that the supports of P_U and P_Y are bounded from above by

$$u=\log\frac{\overline{\delta}}{\delta}+\log\frac{\pi_{01}}{\overline{\pi_{01}}}=r(\delta)-r(\pi_{01}).$$

Furthermore, since $f(x)\geq f(u)$ for $x\leq u$, the U_i and Y_i are also bounded from below by

$$\ell=-r(\delta)+f(u).$$

Next, we argue that the values of q_i, t_i, r_i, q_{i-1}, and r_{i-1}, appearing in the Markov chain transitions, determine intervals among $I_j, j=0,1,2,3$, into which U_i and Y_i must fall and do so according to the table below.

I	Y_i	U_i
I_0	$\{(r_i,q_{i-1})=(-1,1)\}$	$\{(q_i,t_i,q_{i-1})=(-1,0,1)\}\cup\{(q_i,t_i,r_{i-1})=(-1,1,1)\}$
I_1	$\{(r_i,q_{i-1})=(-1,-1)\}$	$\{(q_i,t_i,q_{i-1})=(-1,0,-1)\}\cup\{(q_i,t_i,r_{i-1})=(-1,1,-1)\}$
I_2	$\{(r_i,q_{i-1})=(1,1)\}$	$\{(q_i,t_i,q_{i-1})=(1,0,1)\}\cup\{(q_i,t_i,r_{i-1})=(1,1,1)\}$
I_3	$\{(r_i,q_{i-1})=(1,-1)\}$	$\{(q_i,t_i,q_{i-1})=(1,0,-1)\}\cup\{(q_i,t_i,r_{i-1})=(1,1,-1)\}$

(A.64)

The entries in the row corresponding to interval I_j and columns corresponding to Y_i and U_i specify the values of the above variables that force Y_i and U_i, respectively, to fall in I_j. The table is derived by inspecting the transitions of $\{(U_i,Y_i)\}$. As a representative example of how the table is filled, we consider the entry in row I_2 and column U_i and show how $\{(q_i,t_i,q_{i-1})=(1,0,1)\}\cup\{(q_i,t_i,r_{i-1})=(1,1,1)\}$ implies that $U_i\in I_2$. In the case that $\{(q_i,t_i,q_{i-1})=(1,0,1)\}$, the

transitions of $\{(U_i, Y_i)\}$ imply that

$$U_i \in [\min_{x\in[\ell,u]} r(\delta)+f(r(\delta)+f(x)), \max_{x\in[\ell,u]} r(\delta)+f(r(\delta)+f(x))] \quad \text{(A.65)}$$

$$\overset{(a)}{\subset} [r(\delta)+f(r(\delta)+f(-\infty)), r(\delta)+f(r(\delta)+f(u))] \quad \text{(A.66)}$$

$$= [r(\delta)+f(r(\delta)-r(\pi_{01})), r(\delta)+f(r(\delta)+f(r(\delta)-r(\pi_{01})))], \quad \text{(A.67)}$$

where (a) follows from the fact noted above that $f(x)$ is decreasing in x. The case of $\{(q_i, t_i, r_{i-1}) = (1,1,1)\}$ and the other entries of the table are obtained in a similar fashion.

Lemma 4.17 guarantees that for all sufficiently small $\delta > 0$, the intervals I_j, $j = 0,1,2,3$, are disjoint. From this and the fact that the events in each column of the above table are exhaustive, we can conclude that for all sufficiently small $\delta > 0$, the probabilities of the interval I_j under P_{Y_i} and P_{U_i}, for all $i > 2$, and hence under P_Y and P_U, coincide with the probabilities of the corresponding events in the respective columns of Table A.64. The entries of Table 62 are simply the probabilities of the events in Table A.64 as specified by the Markov chain transitions. □

References

[1] L. Arnold, L. Demetrius, and M. Gundlach. *Evolutionary formalism for products of positive random matrices*. Ann. Appl. Prob., **4** (1994) 859–901

[2] T. Berger and J. D. Gibson. *Lossy source coding*. IEEE Trans. Inf. Theory, **44**(6) (1998) 2693–2723

[3] J. J. Birch. *Approximations for the entropy for functions of Markov chains*. Ann. Math. Statist., **33** (1962) 930–938

[4] D. Blackwell. *The entropy of functions of finite-state Markov chains.* In Trans. First Prague Conf. Information Theory, Statistical Decision Functions, Random Processes, 1957, pp. 13–20

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory, 2nd edn.*. Wiley, New Jersey, 2006

[6] J. L. Devore. *A note on the observation of a Markov source through a noisy channel*. IEEE Trans. Inf. Theory **20** (1974) 762–764

[7] 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*. In Proc. Int. Symp. Information Theory, Chicago, IL, June 2004, p. 12

[8] E. O. Elliott. *Estimates of error rates for codes on burst-noise channels*. Bell Syst. Tech. J. **42** (1963) 1977–1997

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

[10] E. N. Gilbert. *Capacity of a burst-noise channel*. Bell Syst. Tech. J. **39** (1960) 1253–1265
[11] F. Le Gland and L. Mevel. *Exponential forgetting and geometric ergodicity in hidden Markov models*. Math. Control Signals Syst. **13**(1) (2000) 63–93
[12] G. H. Golub and C. F. Van Loan. *Matrix Computations*, third edn. Johns Hopkins, Baltimore, MD, 1996
[13] R. M. Gray. *Information rates of autoregressive processes*. IEEE Trans. Inf. Theory **16**(2) (1970) 412–421
[14] R. M. Gray. *Rate distortion functions for finite-state finite-alphabet Markov sources*. IEEE Trans. Inf. Theory **17**(2) (1971) 127–134
[15] G. Han and B. Marcus. *Analyticity of entropy rate of hidden Markov chains*. IEEE Trans. Inf. Theory **52**(12) (2006) 5251–5266
[16] G. Han and B. Marcus. *Derivatives of entropy rate in special families of hidden Markov chains*. IEEE Trans. Inf. Theory **53**(7) (2007) 2642–2652
[17] G. Han and B. Marcus. *Asymptotics of noisy constrained channel capacity*. Ann. Appl. Prob. **19**(3) (2009) 1063–1091
[18] G. Han, B. Marcus, and Y. Peres. *A note on a complex Hilbert metric with application to domain of analyticity for entropy rate of hidden Markov processes*. This volume, 2011
[19] B. M. Hochwald and P. R. Jelenković. *State learning and mixing in entropy of hidden Markov processes and the Gilbert–Elliott channel*. IEEE Trans. Inf. Theory **45**(1) (1999) 128–138
[20] T. Holliday, P. Glynn, and A. Goldsmith. *Capacity of finite state Markov channels with general inputs*. In Int. Symp. Information Theory, Yokohama, Japan, June–July 2003, p. 289
[21] T. Holliday, P. Glynn, and A. Goldsmith. *Capacity of finite state channels based on Lyapunov exponents of random matrices*. IEEE Trans. Inf. Theory **52**(8) (2006) 3509–3532
[22] P. Jacquet, G. Seroussi, and W. Szpankowski. *On the entropy of a hidden Markov process*. In Proc. Data Compression Conf., Snowbird, UT, June 2004, pp. 362–371
[23] J. Luo and D. Guo. *On the entropy rate of hidden Markov processes observed through arbitrary memoryless channels*. IEEE Trans. Inf. Theory **55**(4) (2009) 1460–1467
[24] D. J. C. MacKay. *Equivalence of Boltzmann chains and hidden Markov models*. Neural Comput., **8** (1996) 178–181
[25] M. Mushkin and I. Bar-David. *Capacity and coding for the Gilbert–Elliott channel*. IEEE Trans. Inf. Theory **35** (1989) 1277–1290
[26] C. Nair, E. Ordentlich, and T. Weissman. *On asymptotic filtering and entropy rate for a hidden Markov process in the rare transitions regime*. In Int. Symp. Information Theory Adelaide, Australia, September 2005, pp. 1838–1842
[27] E. Ordentlich and T. Weissman. *New bounds on the entropy of hidden Markov processes*. In *Proc. IEEE Information Theory Workshop*, San Antonio, TX October 2004, pp. 117–122
[28] E. Ordentlich and T. Weissman. *Approximations for the entropy rate of a hidden Markov process*. In Int. Symp. Information Theory, Adelaide, Australia, September 2005, pp. 2198–2202

[29] E. Ordentlich and T. Weissman. *On the optimality of symbol by symbol filtering and denoising*. IEEE Trans. Information Theory **52**(1) (2006) 19–40
[30] Y. Peres. *Analytic dependence of Lyapunov exponents on transition probabilities*. In Proc. Oberwolfack Conf., Lecture Notes in Mathematics **1486**. Springer, Berlin, 1991, pp. 64–80
[31] Y. Peres and A. Quas. *Entropy rate for hidden Markov chains with rare transitions*. This volume, 2011
[32] H. Pfister. *On the Capacity of Finite State-Channels and the Analysis of Convolutional Accumulate-m Codes*. PhD thesis, University of California, San Diego, CA, 2003
[33] M. Talagrand. *The Sherrington–Kirkpatrick model: a challenge to mathematicians*. Prob. Theory Relat. Fields **110** (1998) 109–176
[34] T. Weissman and N. Merhav. *On competitive predictability and its relation to rate-distortion theory*. IEEE Trans. Inf. Theory, **49** (2003) 3185–3193
[35] T. Weissman and E. Ordentlich. *The empirical distribution of rate-constrained source codes*. IEEE Trans. Inf. Theory, **51**(11) (2005) 3718–3733
[36] O. Zuk, I. Kanter, and E. Domany. *Asymptotics of the entropy rate for a hidden Markov process*. In Proc. Data Compression Conf., Snowbird, UT 2005 pp. 173–182