On the Entropy of a Hidden Markov Process

Philippe Jacquet,* Gadiel Seroussi,† and Wojciech Szpankowski‡

Abstract. We study the entropy rate of a binary hidden Markov process (HMP) defined by observing the output of a binary symmetric channel whose input is a first-order binary Markov process. Despite the simplicity of the models involved, the characterization of this entropy is a long standing open problem. By presenting the probability of a sequence under the model as a product of random matrices, we show that the entropy rate sought is a top Lyapunov exponent of the product, which explains the difficulty in its explicit computation. We apply the same product of random matrices to derive an explicit expression for a first order Taylor approximation of the entropy rate with respect to the parameter of the binary symmetric channel. The accuracy of the approximation is validated against empirical simulation results. We also extend our results to Rényi's entropy of any order.

1 Introduction

Let $X = \{X_k\}_{k\geq 1}$ be a first order stationary Markov process over a binary alphabet, with transition matrix $\mathbf{P} = \{\pi_{ab}\}$ such that $\pi_{ab} = P_X(X_k{=}b|X_{k-1}{=}a),\ a, b \in \{0, 1\}$. Consider also a Bernoulli (binary i.i.d.) *noise* process $E = \{E_k\}_{k\geq 1}$, independent of X, such that $P(E_i = 1) = \varepsilon$. Finally, define the process $Z = \{Z_k\}_{k\geq 1}$, with

$$Z_k = X_k \oplus E_k, \quad k \geq 1, \tag{1}$$

where $\oplus$ denotes addition modulo 2 (exclusive-or). One can view Z as the output of a *binary symmetric channel* with noise E, whose input is X. Notice that the process Z is completely characterized by the parameters π_{01}, π_{10}, and ε.

The process Z is, in a sense, one of the simplest examples of a *hidden Markov process* (HMP). In more generality, a HMP is a process resulting from observing any discrete-time, finite state homogeneous Markov chain through a discrete-time memoryless channel [1, 2, 3]. The chain and the channel, can be defined over arbitrary alphabets, discrete or continuous. HMPs have been studied extensively, and few other statistical tools have had such a wide range of applications in so many domains of science and technology. The applications, to cite just a few, include automatic character recognition [5], speech recognition [6, 7, 8], statistics [9], communications and information theory [10], DNA sequencing [11, 12], etc., with just a small sample of references given for each application. A comprehensive survey of HMP research and applications can be found in [4], including an extensive bibliography. HMPs are also referred to in the literature as *hidden Markov models* (HMM) (see, e.g., [7, 8]).

The simplicity of the definition of HMPs is misleading, and despite the extensive research on their properties and applications, some questions on fundamental

*INRIA, Rocquencourt, France, philippe.jacquet@inria.fr.

†HP Laboratories, Palo Alto, California, USA, seroussi@hpl.hp.com.

‡Purdue University, W. Lafayette, Indiana, USA, spa@cs.purdue.edu. This research was supported in part by NSF Grant CCR-0208709 and NIH Grant R01 GM068959-01.

properties of the processes remain open, even for the "simple" case (1). Some of these questions concern the performance of filtering [13, 14], denoising [15, 14], and compression [4] on hidden Markov sources. In all these cases, algorithms exist that achieve optimal performance (e.g., minimal residual noise or code length), even universally (without knowledge of the process parameters). However, in general, the optimal value of the performance of interest for each of the problems has not been explicitly characterized. In the case of compression, the problem of interest is the determination of the *entropy rate* $h(Z)$ of the process Z as an explicit ("single-letter") expression in the parameters of the process [4]. This is the question addressed in this paper. Entropy rates are important in data compression (for obvious reasons), but also in other algorithms such as searching and sorting. For example, it is known (cf. [16]) that the average search time in digital trees is asymptoticaly equal to $h^{-1}\log n$ where h is the entropy rate of the input sequence of symbols (assumed generated by a strongly mixing process of positive entropy rate), and n the number of sequences stored.

The question of computing the entropy of a HMP was studied as early as [17], where the analysis suggests the intrinsic complexity of the HMP entropy as a function of the process parameters.[1] The reference shows an expression of the entropy in terms of a measure Q, which solves an integral equation dependent on the parameters of the process. The measure is hard to extract from the equation in any explicit way. More recently, the problem of determining the residual noise of the best filter for a HMP was studied in [13], and explicit asymptotic formulas for the regime where $\pi_{ab} \to 0,\ a \neq b$ were obtained. Explicit formulas for the entropy rate for some asymptotic regimes of the parameters π_{ab} are obtained in [14]. In contrast, our study focuses on the regime where the channel parameter (noise) ε is small.

The rest of this extended summary is organized as follows: In Section 2 we present the probability of a HMP sequence Z_1^n as a product of random matrices. In Section 3 we rely on classical results on products of random matrices, and show that the entropy we seek is a top Lyapunov exponent of such a product. Lyapunov exponents are notoriously difficult to compute. In [19] it was proved that the Lyapunov exponent is not algorithmically approximable (i.e., there is no polynomial algorithm to compute it with any degree of approximation). This provides additional supporting evidence to the elusiveness of the HMP entropy. In Section 4 we symbolically manipulate the product of random matrices of Section 2, in a manner that allows us to obtain our main result: an explicit first-order Taylor expansion of $h(Z)$ near $\varepsilon = 0$, as a function of the parameters π_{ab}. We show that the linear term of the expansion has a pleasant information-theoretic flavor, and can be expressed as a Kullback-Liebler divergence between distributions of triplets related to the underlying Markov process. We extend our analysis to Rényi's entropies. Finally, in Section 5, we present results of empirical simulations of HMPs, and validate the entropy rate approximation.

[1] The setting in [17] is that of deterministic functions of Markov chains, which can be shown to be equivalent to the HMP setting in the finite-alphabet case [3, 18, 4].

2 A Product of Random Matrices

For any sequence $\{Y_k\}_{k\geq 1}$, we denote by Y_i^j the (sub-)sequence $Y_i, Y_{i+1} \ldots Y_j$, $j \geq i$. For a binary variable Y, we write $\bar{Y} = 1 \oplus Y$. Observe that $Z_i = X_i$ if $E_i = 0$ and $Z_i = \bar{X}_i$ if $E_i = 1$. We compute the probability $P(Z_1^n)$ of a vector Z_1^n emitted by the HMP, for some $n \geq 1$. Using elementary properties of probabilities, and our assumptions on the processes X and E, we have

$$\begin{aligned} P(Z_1^n, E_n) &= P(Z_1^n, E_{n-1}=0, E_n) + P(Z_1^n, E_{n-1}=1, E_n) \\ &= P(Z_1^{n-1}, Z_n, E_{n-1}=0, E_n) + P(Z_1^{n-1}, Z_n, E_{n-1}=1, E_n) \\ &= P(Z_n, E_n | Z_1^{n-1}, E_{n-1}=0) P(Z_1^{n-1}, E_{n-1}=0) + \\ &\quad\; P(Z_n, E_n | Z_1^{n-1}, E_{n-1}=1) P(Z_1^{n-1}, E_{n-1}=1) \\ &= P(E_n) P_X(Z_n \oplus E_n | Z_{n-1}) P(Z_1^{n-1}, E_{n-1}=0) \\ &+ P(E_n) P_X(Z_n \oplus E_n | \bar{Z}_{n-1}) P(Z_1^{n-1}, E_{n-1}=1) \end{aligned} \tag{2}$$

We shall write this probability in matrix form. We denote *row* vectors by bold lowercase letters, matrices by bold uppercase letters, and we let $\mathbf{1} = [1, 1]$; superscript t will denote transposition. Let

$$\mathbf{p}_n = [P(Z_1^n, E_n = 0), P(Z_1^n, E_n = 1)]$$

and

$$\mathbf{M}(Z_{n-1}, Z_n) = \begin{bmatrix} (1-\varepsilon) P_X(Z_n | Z_{n-1}) & \varepsilon P_X(\bar{Z}_n | Z_{n-1}) \\ (1-\varepsilon) P_X(Z_n | \bar{Z}_{n-1}) & \varepsilon P_X(\bar{Z}_n | \bar{Z}_{n-1}) \end{bmatrix} \tag{3}$$

where the expressions $P_X(Z_i | Z_{i-1})$ are the Markov transition probabilities computed on the components of the HMP Z.

One concludes from (2) that

$$\mathbf{p}_n = \mathbf{p}_{n-1} \mathbf{M}(Z_{n-1}, Z_n). \tag{4}$$

Since $P(Z_1^n) = \mathbf{p}_n \mathbf{1}^t = P(Z_1^n, E_n = 0) + P(Z_1^n, E_n = 1)$, after iterating (4), we obtain

$$P(Z_1^n) = \mathbf{p}_1 \mathbf{M}(Z_1, Z_2) \cdots \mathbf{M}(Z_{n-1}, Z_n) \mathbf{1}^t. \tag{5}$$

Notice that the $\mathbf{M}(Z_{i-1}, Z_i)$ are random matrices, since the conditionals $P_X(Z_i | Z_{i-1})$ are random variables.

3 Entropy Rate as a Lyapunov Exponent

The joint distribution $P(Z_1^n)$ of the HMP, presented in (5), has the form of a product of random matrices. It is easy to apply a subadditive ergodic theorem to see that $\mathbf{E}[\log P(Z_1^n)]$ must converge to a constant known as the *top Lyapunov exponent* of the random product [20, 21, 19]). Furstenberg and Kesten [20] (see also [21]) proved the following result.

Theorem 1 (Furstenberg and Kesten, 1960) *Let* $\mathbf{M}_1, \ldots, \mathbf{M}_n$ *form a stationary ergodic sequence and* $\mathbf{E}[\log^+ ||\mathbf{M}_1||] < \infty$ *Then*

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{E}[\log ||\mathbf{M}_1 \cdots \mathbf{M}_n||] = \lim_{n\to\infty} \frac{1}{n} \log ||\mathbf{M}_1 \cdots \mathbf{M}_n|| = \mu \quad \text{a.s..} \tag{6}$$

An immediate corollary to Theorem 1 and (5) is that the entropy of the HMP is equal to a Lyapunov exponent.

Corollary 1 *Consider the HMP* Z *as defined above. The entropy rate* [2]

$$h(Z) = \lim_{n\to\infty} \mathbf{E}[-\frac{1}{n}\log P(Z_1^n)] = \lim_{n\to\infty} \frac{1}{n}\mathbf{E}[-\log\left(\mathbf{p}_1\mathbf{M}(Z_1,Z_2)\cdots\mathbf{M}(Z_{n-1},Z_n)\mathbf{1}^t\right)]$$

is a top Lyapunov exponent of $\mathbf{M}(Z_1,Z_2)\cdots\mathbf{M}(Z_{n-1},Z_n)$.

Proof. To verify the claim, it suffices to show that $\mathbf{p}_1\mathbf{A}\mathbf{1}^t$, of (5) (where we define $\mathbf{A} = \mathbf{M}(Z_1,Z_2)\cdots\mathbf{M}(Z_{n-1},Z_n)$) is a norm. Certainly this is the case since it satisfies the three properties of the norm (since $\mathbf{M}_i$ is a nonnegative matrix). □

We note that the connection between discrete-time, finite state Markov chains and Lyapunov exponents, based also on Theorem 1, is studied in [22], where the dual problem of a memoryless signal going through a Markov channel is considered. The results in [22] also link other interesting parameters like mutual information and channel capacity to Lyapunov exponents.

Unfortunately, it is notoriously difficult to compute top Lyapunov exponents as proved in [19]. Therefore, in the next section we derive an explicit asymptotic expansion of the entropy rate $h(Z)$, which does not depend on direct computation of Lyapunov exponents.

4 Asymptotic Expansion

We derive an expansion of the entropy rate $h(Z)$ for the HMP Z as a function of ε, for small values of ε.

For a given binary n-tuple z_1^n, let

$$\begin{aligned}
\mathbf{M}_i &= \mathbf{M}(z_i, z_{i+1}) = \begin{bmatrix} (1-\varepsilon)P_X(z_{i+1}|z_i) & \varepsilon P_X(\bar{z}_{i+1}|z_i) \\ (1-\varepsilon)P_X(z_{i+1}|\bar{z}_i) & \varepsilon P_X(\bar{z}_{i+1}|\bar{z}_i) \end{bmatrix} \\
&= \begin{bmatrix} P_X(z_{i+1}|z_i) & 0 \\ P_X(z_{i+1}|\bar{z}_i) & 0 \end{bmatrix} + \varepsilon \begin{bmatrix} -P_X(z_{i+1}|z_i) & P(\bar{z}_{i+1}|z_i) \\ -P_X(z_{i+1}|\bar{z}_i) & P(\bar{z}_{i+1}|\bar{z}_i) \end{bmatrix} \\
&\stackrel{\text{def}}{=} \mathbf{M}_i^{(0)} + \varepsilon\mathbf{M}_i^{(1)},
\end{aligned} \tag{7}$$

and

$$\mathbf{p}_0 = [P_X(z_1),\ 0] + \varepsilon[-P_X(z_1),\ P_X(\bar{z}_1)] \stackrel{\text{def}}{=} \mathbf{M}_0^{(0)} + \varepsilon\mathbf{M}_0^{(1)}. \tag{8}$$

Here, $P_X(b|a) = \pi_{ab}$ are transition probabilities of the Markov process X, which is assumed stationary, and $P_X(a)$ are its marginal (stationary) probabilities, $a, b \in$

[2]When no base is specified, logarithms are to base 2; $\ln x$ will denote the natural logarithm of x.

$\{0,1\}$. Notice that for the sake of a uniform notation, the matrices $\mathbf{M}_0^{(0)}$ and $\mathbf{M}_0^{(1)}$ are of dimensions 1×2, while $\mathbf{M}_i^{(0)}$ and $\mathbf{M}_i^{(1)}$ are 2×2 for $i > 0$. Using the above definitions and (5), we obtain

$$\begin{aligned} P_Z(z_1^n) &= P(Z_1^n = z_1^n) = \mathbf{p}_0\mathbf{M}_1\mathbf{M}_2\cdots\mathbf{M}_{n-1}\mathbf{1}^t \\ &= (\mathbf{M}_0^{(0)}+\varepsilon\mathbf{M}_0^{(1)})(\mathbf{M}_1^{(0)}+\varepsilon\mathbf{M}_1^{(1)})(\mathbf{M}_2^{(0)}+\varepsilon\mathbf{M}_2^{(1)})\cdots(\mathbf{M}_{n-1}^{(0)}+\varepsilon\mathbf{M}_{n-1}^{(1)})\mathbf{1}^t. \quad (9) \end{aligned}$$

In order to compute the Shannon entropy (and, later, Rényi entropies of any order), we resort to the following formal definition (which was also used in entropy computations in [23]):

$$R(s,\varepsilon) = \sum_{z_1^n} P_Z^s(z_1^n), \quad (10)$$

where s is a complex variable, and the summation is over all binary n-tuples. Define also the matrix

$$\mathbf{P}(s) = \begin{bmatrix} \pi_{00}^s & \pi_{01}^s \\ \pi_{10}^s & \pi_{11}^s \end{bmatrix}, \quad (11)$$

and the vector $\boldsymbol{\pi}(s) = [P_X^s(0), P_X^s(1)]$. The following lemma summarizes some key properties of the function $R(s,\varepsilon)$.

Lemma 1 *(i) The (unnormalized) Shannon entropy of Z_1^n is given by*

$$H(Z_1^n) = \mathbf{E}\left[-\log P(Z_1^n)\right] = -(\ln 2)\left.\frac{\partial}{\partial s}R(s,\varepsilon)\right|_{s=1}. \quad (12)$$

The entropy of the underlying Markov sequence is $H(X_1^n) = (-\ln 2)\frac{\partial}{\partial s}R(s,0)\big|_{s=1}$.
(ii) We have

$$R(s,0) = \sum_{z^n} P_X^s(z_1^n) = \boldsymbol{\pi}(s)\mathbf{P}^{n-1}(s)\mathbf{1}^t. \quad (13)$$

Using a formal Taylor expansion near $\varepsilon = 0$, we write $R(s,\varepsilon) = R(s,0) + \varepsilon\frac{\partial}{\partial\varepsilon}R(s,\varepsilon)|_{\varepsilon=0} + O(R_{\varepsilon,\varepsilon}(s,\varepsilon')\varepsilon^2)$, where $R_{\varepsilon,\varepsilon}(s,\varepsilon')$ is the second derivative with respect to ε computed at some ε'. It can be shown (details will be provided in the journal version of this paper) that $R_{\varepsilon,\varepsilon,s}(1,\varepsilon') = O(n)$ for $0 < \varepsilon < 0.5$, where $R_{\varepsilon,\varepsilon,s}(1,\varepsilon')$ is the first derivative with respect to s at $s = 1$ of $R_{\varepsilon,\varepsilon}(s,\varepsilon')$. Now, by Lemma 1(i) and the above estimate of $R_{\varepsilon,\varepsilon,s}(1,\varepsilon')$, we obtain

$$\begin{aligned} H(Z_1^n) &= H(X_1^n) - (\ln 2)\varepsilon\left.\frac{\partial^2}{\partial s\partial\varepsilon}R(s,\varepsilon)\right|_{\varepsilon=0,s=1} + O(n\varepsilon^2) \quad (14) \\ &= H(X_1^n) - (\ln 2)\varepsilon\frac{\partial}{\partial s}\frac{\partial}{\partial\varepsilon}\left.\sum_{z_1^n} P_Z^s(z_1^n)\right|_{\varepsilon=0,s=1} + O(n\varepsilon^2). \end{aligned}$$

To compute the derivative of $P_Z^s(z_1^n)$ at $\varepsilon = 0$, we first differentiate both sides of (9), obtaining

$$\left.\frac{\partial}{\partial\varepsilon}P_Z(z_1^n)\right|_{\varepsilon=0} = \sum_{i=0}^{n-1}\mathbf{M}_0^{(0)}\mathbf{M}_1^{(0)}\cdots\mathbf{M}_{i-1}^{(0)}\mathbf{M}_i^{(1)}\mathbf{M}_{i+1}^{(0)}\cdots\mathbf{M}_{n-1}^{(0)}\,\mathbf{1}. \quad (15)$$

Using the definitions of the matrices $\mathbf{M}_i^{(b)}$, $b = 0, 1$, from (7) and (8), it follows, after some algebraic and probability manipulations, that

$$\left.\frac{\partial}{\partial\varepsilon} P_Z(z_1^n)\right|_{\varepsilon=0} = P_X(z_1^n) \sum_{i=0}^{n-1} (g_i(z_1^n) - 1), \tag{16}$$

where

$$\begin{aligned} g_i(z_1^n) &= \frac{P_X(\bar{z}_{i+1}|z_i) P_X(z_{i+2}|\bar{z}_{i+1})}{P_X(z_{i+1}|z_i) P_X(z_{i+2}|z_{i+1})} = \frac{P_X(z_i \bar{z}_{i+1} z_{i+2})}{P_X(z_i z_{i+1} z_{i+2})}, \ 0 < i < n-1, \\ g_0(z_1^n) &= \frac{P_X(\bar{z}_1 z_2)}{P_X(z_1 z_2)} - 1, \quad g_{n-1}(z_1^n) = \frac{P_X(z_{n-2} \bar{z}_{n-1})}{P_X(z_{n-2} z_{n-1})} - 1. \end{aligned} \tag{17}$$

Thus,

$$\left.\frac{\partial}{\partial\varepsilon} P_Z^s(z_1^n)\right|_{\varepsilon=0} = \left[s P_Z^{s-1}(z_1^n) P_X(z_1^n) \sum_{i=0}^{n-1} (g_i(z_1^n) - 1) \right]_{\varepsilon=0}. \tag{18}$$

Observing that $P_Z(z_1^n)|_{\varepsilon=0} = P_X(z_1^n)$, taking derivatives on both sides of (10), and using (18), we obtain

$$\left.\frac{\partial}{\partial\varepsilon} R(s, \varepsilon)\right|_{\varepsilon=0} = s \sum_{z_1^n} P_X^s(z_1^n) \sum_{i=0}^{n-1} (g_i(z_1^n) - 1). \tag{19}$$

By an argument similar to that leading to the result of Lemma 1(ii), we obtain the following.

Lemma 2 *We have*

$$\left.\frac{\partial}{\partial\varepsilon} R(s, \varepsilon)\right|_{\varepsilon=0} = s\boldsymbol{\pi}(s) \sum_{i=1}^{n-1} \mathbf{P}^{i-1}(s) \left(\mathbf{Q}_1(s) \mathbf{Q}_2(s) - \mathbf{P}^2(s) \right) \mathbf{P}^{n-i-2}(s) \mathbf{1}^t \tag{20}$$

where

$$\mathbf{Q}_1(s) = \begin{bmatrix} \pi_{00}\pi_{01}^{s-1} & \pi_{01}\pi_{00}^{s-1} \\ \pi_{10}\pi_{11}^{s-1} & \pi_{11}\pi_{10}^{s-1} \end{bmatrix}, \qquad \mathbf{Q}_2(s) = \begin{bmatrix} \pi_{00}\pi_{10}^{s-1} & \pi_{01}\pi_{11}^{s-1} \\ \pi_{10}\pi_{00}^{s-1} & \pi_{11}\pi_{01}^{s-1} \end{bmatrix}.$$

To find the linear term in the Taylor expansion (14), we need to differentiate (20) with respect to s, and evaluate at $s = 1$. To facilitate this computation, we use the spectral representation [16] of the matrix $\mathbf{P}(s)$. Let $\lambda(s)$ be the main eigenvalue of $\mathbf{P}(s)$ with $\mathbf{r}_1^t(s)$ and $\mathbf{l}_1(s)$ being the corresponding right and left main eigenvectors, respectively, normalized so that $\mathbf{l}_1(s)\mathbf{r}_1^t(s) = 1$. Let also $\mu(s)$ be the second eigenvalue, with $\mathbf{r}_2^t(s)$ and $\mathbf{l}_2(s)$ the respective right and left eigenvectors. The matrix spectral representation yields (cf. [16, 24])

$$\mathbf{P}^k(s) = \lambda^k(s)\, \mathbf{r}_1^t(s) \mathbf{l}_1(s) + \mu^k(s)\, \mathbf{r}_2^t(s) \mathbf{l}_2(s). \tag{21}$$

Notice that each term on the right hand side of (21) is rank-one matrix obtained as the outer product (column $\times$ row) of the respective eigenvectors.

Operating with the spectral representations, we obtain

$$\frac{\partial^2}{\partial\varepsilon\partial s}R(s,\varepsilon)\bigg|_{\substack{\varepsilon=0,\\ s=1}} = n\,\boldsymbol{\pi}(1)\mathbf{r}_1^t(1)\;\mathbf{l}_1(1)\mathbf{1}^t\;\mathbf{l}_1(1)\;\frac{\partial}{\partial s}\left(\mathbf{Q}(s)-\mathbf{P}^2(s)\right)\bigg|_{s=1}\mathbf{r}_1^t(1), \tag{22}$$

where $\mathbf{Q}(s)=\mathbf{Q}_1(s)\mathbf{Q}_2(s)$. Observe that

$$\frac{\partial}{\partial s}\left(\mathbf{Q}(s)-\mathbf{P}^2(s)\right)\bigg|_{s=1} = \mathbf{Q}_1'(1)\mathbf{Q}_2(1)-\mathbf{P}'(1)\mathbf{P}(1)+\mathbf{Q}_1(1)\mathbf{Q}_2'(1)-\mathbf{P}(1)\mathbf{P}'(1),$$

using an obvious shorthand for differentiation with respect to s. Also, for the transition probability matrix $\mathbf{P}=\mathbf{P}(1)$, we have

$$\mathbf{l}_1(1)=\boldsymbol{\pi}(1)=[\frac{\pi_{10}}{\pi_{10}+\pi_{01}},\;\frac{\pi_{01}}{\pi_{10}+\pi_{01}}],$$

the vector of stationary probabilities of the Markov process X, and $\mathbf{r}_1(1)=[1,\,1]$. Thus, $\boldsymbol{\pi}(1)\mathbf{r}_1^t(1)=\mathbf{l}_1(1)\mathbf{1}^t=1$. The main result of this section now follows from (22) and from (14) by carrying out the computations and arranging terms.

Theorem 2 *The entropy rate of the process Z is*

$$h(Z)=\lim_{n\to\infty}\frac{1}{n}H_n(Z^n)=h(X)+f_1(\pi_{01},\pi_{10})\varepsilon+O(\varepsilon^2),$$

with

$$f_1(\pi_{01},\pi_{10})=\mathbb{D}\left(P_X(z_1z_2z_3)||P_X(z_1\bar{z}_2z_3)\right)=\sum_{z_1z_2z_3}P_X(z_1z_2z_3)\log\frac{P_X(z_1z_2z_3)}{P_X(z_1\bar{z}_2z_3)}, \tag{23}$$

where $h(X)$ is the entropy rate of the Markov process X, $\mathbb{D}$ denotes the Kullback-Liebler divergence, and the summation is over all binary triplets.

Example. Consider a Markov process with symmetric transition probabilities $\pi_{01}=\pi_{10}=\pi$, $\pi_{00}=\pi_{11}=1-\pi$. This process has stationary probabilities $P_X(0)=P_X(1)=\frac{1}{2}$. The probabilities $P_X(z_1^3)$ of binary triplets are readily computed as $P_X(000)=P_X(111)=\frac{1}{2}(1-\pi)^2$, $P_X(001)=P_X(011)=P_X(100)=P_X(110)=\frac{1}{2}\pi(1-\pi)$, $P_X(010)=P_X(101)=\pi^2$. Substituting these values into (23), we obtain

$$f_1(\pi,\pi)=2(1-2\pi)\log\frac{1-\pi}{\pi}. \tag{24}$$

Remarks. We note that similar results can be obtained for Rényi entropies H_s of any order s and the respective entropy rates, by noting that

$$H_s(Z_1^n)=\frac{\log R(s,\varepsilon)}{1-s}. \tag{25}$$

A derivation similar to that leading to Theorem 2 yields

$$h_s(Z)=h_s(X)+\frac{\varepsilon}{(1-s)\lambda(s)}\mathbf{l}_1(s)\left(\mathbf{Q}(s)-\mathbf{P}^2(s)\right)\mathbf{r}_1(s)+O(\varepsilon^2), \tag{26}$$

where the Markov Renyi's entropy is $h_s(X)=\frac{1}{1-s}\log\lambda(s)$ (cf. [16]).

Also, higher order terms for the Shannon (or Rényi) entropy rate expansion can be obtained by considering terms as in (15), but with, say, k factors of type $\mathbf{M}_i^{(1)}$ for the k-th order term.

Parameters			Calculated			Empirical
ε	π	n	$h(X)$	$f_1(\pi,\pi)$	$h(X)+$ $f_1(\pi,\pi)\varepsilon$	$-\frac{1}{n}\log P_Z(z_1^n)$
0.001	0.005	$4\cdot 10^9$	0.045	15.121	0.061	0.056
0.001	0.010	$4\cdot 10^9$	0.080	12.994	0.094	0.091
0.001	0.025	$1\cdot 10^9$	0.168	10.042	0.179	0.177
0.01	0.050	$1\cdot 10^8$	0.286	7.646	0.363	0.349
0.01	0.100	$1\cdot 10^8$	0.469	5.072	0.520	0.514
0.01	0.300	$1\cdot 10^8$	0.881	0.978	0.891	0.891

Table 1: First order approximation of $h(Z)$ according to Theorem 2 and empirical estimation.

5 Experimental Results and Validation

HMPs for various values of the parameters ε and $\pi_{01} = \pi_{10} = \pi$ were simulated, generating pseudo-random HMP sequences of lengths between $n = 10^8$ and $n = 4\cdot 10^9$. For each generated sequence z_1^n, the probability $P_Z(z_1^n)$ assigned by the hidden Markov model of the given parameters was computed, and $-\frac{1}{n}\log P_Z(z_1^n)$ was taken as an estimate for the entropy rate. This is justified by the fact that a sequence emitted by the HMP is "typical" and, thus, satisfies $|\frac{1}{n}\log P_Z(z_1^n) + h_n(Z)| < \delta$ for any $\delta > 0$, with probability approaching one as $n \to \infty$ [25]. A sample of results for some values of ε and π are given in Table 1, where values of $h(Z)$ estimated using the result of Theorem 2 are compared with simulation estimates. The slope $\partial h_n(Z)/\partial\varepsilon|_{\varepsilon=0}$, as a function of π, is plotted in Figure 1. The empirical slope was estimated using first differences of the estimated values of $h_n(Z)$ near $\varepsilon = 0$, and the result compared with the analytical value produced by (24).

References

[1] A. W. Drake, *Observation of a Markov Process Through a Noisy Channel*, PhD thesis, Massachusetts Institute of Technology, 1962.

[2] A.W. Drake, "Observation of a Markov source through a noisy channel," in *Proc. IEEE Symp. Signal Transmission and Processing*, Columbia Univ., New York, 1965, pp. 12–18.

[3] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 37, pp. 1554–1563, 1966.

[4] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inform. Theory*, vol. IT-48, pp. 1518–1569, June 2002.

[5] J. Raviv, "Decision making in Markov chains applied to the problem of pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-3, pp. 536–551, Oct. 1967.

[6] F. Jelinek, L. R. Bahl, and R. L. Mercer, "Design of a linguistic statistical decoder for recognition of continuous speech," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 250–256, May 1975.

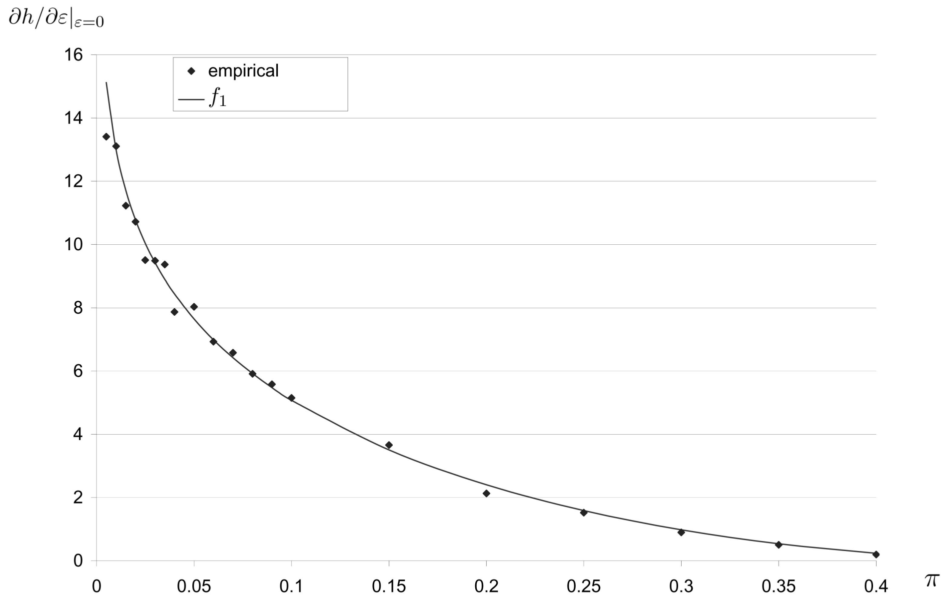

Figure 1: Values of f_1 and empirical estimation of $\partial h/\partial \varepsilon|_{\varepsilon=0}$ as a function of π

[7] J. D. Ferguson, ed., *Application of Hidden Markov Models to Text and Speech*, (Princeton, NJ), IDA-CRD, 1980.

[8] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[9] B. G. Leroux, "Maximum-likelihood estimation for hidden Markov models," *Stochastic Processes Their Applic.*, vol. 40, pp. 127–143, 1992.

[10] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284–287, mar 1974.

[11] G. A. Churchill, "Stochastic models for heterogeneous DNA sequences," *Bull. Math. Biology*, vol. 51, no. 1, pp. 79–94, 1989.

[12] A. V. Lukashin and M. Borodovsky, "GeneMark.hmm: New solutions for gene finding," *Nucleic Acids Res*, vol. 26, no. 4, pp. 1107–1115, 1998.

[13] R. Khasminskii and O. Zeitouni, "Asymptotic filtering for finite state Markov chains," *Stochastic Processes and their Applications*, vol. 63, pp. 1–10, 1996.

[14] E. Ordentlich and T. Weissman, "On the optimality of symbol by symbol filtering and denoising." Preprint, 2003.

[15] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," Technical Report HPL-2003-29, Hewlett-Packard Laboratories, Feb. 2003. Available at: `http://www.hpl.hp.com/techreports/2003/HPL-2003-29.pdf`.

[16] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*. New York: John Wiley & Sons, Inc., 2001.

[17] D. Blackwell, "The entropy of functions of finite-state Markov chains," in *Trans. First Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, (Prague, Czechoslovakia), pp. 13–20, Pub. House Chechoslovak Acad. Sci., 1957.

[18] T. Petrie, "Probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 40, no. 1, pp. 97–115, 1969.

[19] J. Tsitsiklis and V. Blondel, "The Lyapunov exponent and joint spectral radius of pairs of matrices are hard - when not impossible - to compute and to approximate," *Mathematics of Control, Signals, and Systems*, vol. 10, pp. 31–40, 1997.

[20] H. Furstenberg and H. Kesten, "Products of random matrices," *Ann. Math. Statist*, pp. 457–469, 1960.

[21] V. Oseledec, "A multiplicative ergodic theorem," *Trudy Moskov Mat. Obsc*, 1968.

[22] T. Holliday, A. Goldsmith, and P. Glynn, "On entropy and Lyapunov exponents for finite state channels." Preprint, `http://wsl.stanford.edu/Publications/THolliday/Lyapunov.pdf`, 2003.

[23] P. Jacquet and W. Szpankowski, "Entropy computations via analytic depoissonization," *IEEE Trans. Inform. Theory*, vol. IT-45, pp. 1072–1081, 1999.

[24] S. Karlin and H. Taylor, *A First Course in Stochastic Processes*. New York: Academic Press, 1975.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: John Wiley & Sons, Inc., 1991.