On the Entropy of a Hidden Markov Process*

Philippe Jacquet
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France
Philippe.Jacquet@inria.fr

Gadiel Seroussi†
Hewlett-Packard Laboratories
Palo Alto, California
U.S.A.
gseroussi@ieee.org

Wojciech Szpankowski‡
Dept. of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

We study the entropy rate of a 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 equal to a top Lyapunov exponent of the product. This offers an explanation for the elusiveness of explicit expressions for the HMP entropy rate, as Lyapunov exponents are notoriously difficult to compute. Consequently, we focus on asymptotic estimates, and apply the same product of random matrices to derive an explicit expression for a 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 higher order Markov processes and to Rényi entropies of any order.

Index terms: Hidden Markov process, Shannon entropy, Rényi entropy, product of random matrices, top Lyapunov exponent, spectral representation of matrices.

---

*A preliminary version of this work was presented at the *Data Compression Conference*, Snowbird, Utah, 2004.

*Work done in part at the Mathematical Sciences Research Institute, Berkeley, California.

†This research was supported in part by the NSF Grants CCF-0513636, and DMS-0503742, and the NIH Grant R01 GM068959-01.

1 Introduction

Let $X = \{X_k\}_{k\geq 1}$ be a first order stationary Markov process over a binary alphabet, with transition matrix $\mathbf{P} = \{p_{ab}\}$ such that $p_{ab} = P_X(X_k{=}b|X_{k-1}{=}a),\ a, b \in \{0, 1\}$. 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

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

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 p_{01}, p_{10}, and ε.

The process Z is one of the simplest examples of a *hidden Markov process* (HMP). More generally, a HMP is a process resulting from observing any discrete-time, finite state homogeneous Markov chain through a discrete-time memoryless channel [6, 7, 2]. 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 [32], speech recognition [17, 10, 31], statistics [23], communications and information theory [1], DNA sequencing [4, 24], etc., with just a small sample of references given for each application. A comprehensive survey of HMP research and applications can be found in [8], including an extensive bibliography. HMPs are also referred to in the literature as *hidden Markov models* (HMM) (cf. [10, 31]).

The simplicity of the definition of HMPs is misleading, and despite the extensive research on their properties and applications, some questions on fundamental properties of the processes remain open, even for the "simple" case defined in (1). Some of these questions concern the performance of filtering [21, 29], denoising [38, 29], and compression [8] 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 Shannon *entropy rate* $\mathbf{H}(Z)$ of the process Z in terms of the parameters of the process [8]. This is the problem addressed in this paper, where we also consider the case of the more general Rényi entropies [33].

The Shannon entropy is obviously relevant in data compression, but both Shannon and Rényi entropies arise also in other contexts, such as searching, sorting, and pattern matching [36]. For example, consider a sequence $Z_1, Z_2, \ldots, Z_n, \ldots$ generated by a strong mixing source. Define L_n^s as the length of the longest substring of $Z_1 \ldots Z_n$ that has s copies inside $Z_1, \ldots, Z_n$, $s \geq 1$. It is known (cf. [35]) that almost surely[1]

$$\lim_{n\to\infty} \frac{L_n^s}{\log n} = \frac{s}{(s-1)\mathbf{H}_s(Z)}$$

[1] All logarithms are natural, and entropies are measured in nats. Unnormalized entropies will be denoted by $H(\cdot)$, and normalized (per symbol) entropy rates by $\mathbf{H}(\cdot)$.

where $\mathbf{H}_s(Z)$ is the s-th order Rényi entropy rate of Z, as defined in (35). As $s \to 1$ the Rényi entropy approaches the Shannon entropy, and the above holds once L_n^1 is interpreted as the length of the longest prefix of $Z_{n+1}, Z_{n+2} \ldots$ that occurs at least once inside the sequence $Z_1, Z_2, \ldots, Z_n$ [35]. Furthermore, L_n^s can be viewed as the height of the so-called s-suffix trie in which an s-trie is built from suffixes of the sequence $Z_1, \ldots, Z_n$, as explained in [36]. In many applications, data modeled by Markov processes is affected by noise, and thus, the above mentioned asymptotic behaviors are governed by the corresponding HMP entropies.

The question of computing the Shannon entropy (or, simply, *entropy*) of a HMP was studied as early as [3], where the analysis suggests the intrinsic complexity of expressing the HMP entropy as a function of the process parameters. 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 [21], and explicit asymptotic formulas for the regime where $p_{ab} \to 0,\ a \neq b$ were obtained. Furthermore, there has been a flurry of activity on the subject of asymptotic estimates of the HMP entropy rate since (and partly stimulated by) the preliminary publication of the results of this paper in [19]. In particular, Ordentlich and Weissman [28, 29] present a different methodology for analyzing the HMP entropy rate, from which the first derivative can be obtained in various regimes of the parameters p_{01}, p_{10}, and ε. Zuk et al. [39] present formulas for higher order coefficients of the Taylor expansion in the symmetric case ($p_{01}{=}p_{10}$), obtained using mathematical tools from statistical physics. Han and Marcus [14, 15] characterize the analyticity of the HMP entropy rate, and obtain a broad generalization of the results of [39]. As this paper is going to press, insights gained from the study of the HMP entropy rate are being applied to another long standing open problem, namely, the capacity of a noisy constrained channel [16, 20].

Our study will focus on the estimation of the HMP entropy rate in the regime where the channel parameter (noise) ε is small. The paper is organized as follows: In Section 2 we outline the analysis and our main results. In particular, we derive the probability of a HMP sequence Z_1^n as a product of random matrices. Relying on classical results on products of random matrices, in Theorem 1 we show that the entropy we seek is a top Lyapunov exponent of a well defined matrix process. Lyapunov exponents are notoriously difficult, or even infeasible to compute, as argued in [37]. This provides additional supporting evidence to the elusiveness of the HMP entropy. In Theorem 2 we present an explicit first-order Taylor expansion of $\mathbf{H}(Z)$ near $\varepsilon = 0$, as a function of the parameters p_{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 also give an estimate of the quadratic term of the expansion, and extend our analysis to higher order HMPs and to Rényi's entropies. The section also includes results of empirical simulations of HMPs, which validate the entropy rate approximation. Most proofs are deferred to Section 3 where we use spectral representation of positive matrices to derive our main results.

2 Main Results

We denote by $\bar{Y}$ the Boolean complement of a binary variable Y, and, for any sequence $\{Y_k\}_{k\geq 1}$, we denote by Y_i^j the (sub-)sequence $Y_i, Y_{i+1} \ldots Y_j$, $j \geq i$. Recalling the definition of the HMP in (1), we observe that $Z_i = X_i$ if $E_i = 0$ and $Z_i = \bar{X}_i$ if $E_i = 1$. We next derive an expression for the probability distribution[2] $P(Z_1^n)$ of a sequence Z_1^n emitted by the HMP, for some $n \geq 1$. First, using elementary properties of probabilities, and our assumptions on the processes X and E, we have

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

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

$$\mathbf{p}_n = [\, P(Z_1^n, E_n = 0),\ P(Z_1^n, E_n = 1)\,] \tag{3}$$

and

$$\mathbf{M}_\varepsilon(Z_{n-1}, Z_n) = \begin{bmatrix} (1-\varepsilon) P_X(Z_n | Z_{n-1}) & \varepsilon P_X(\bar{Z}_n | Z_{n-1}) \\ (1-\varepsilon) P_X(Z_n | \bar{Z}_{n-1}) & \varepsilon P_X(\bar{Z}_n | \bar{Z}_{n-1}) \end{bmatrix} \tag{4}$$

where the expressions $P_X(Z_i|Z_{i-1})$ are the Markov transition probabilities computed on the components of the HMP Z, and ε is the parameter of the Bernoulli noise process.

One concludes from (2) that

$$\mathbf{p}_n = \mathbf{p}_{n-1} \mathbf{M}_\varepsilon(Z_{n-1}, Z_n), \quad n > 1. \tag{5}$$

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 (5), we obtain

$$P(Z_1^n) = \mathbf{p}_1 \mathbf{M}_\varepsilon(Z_1, Z_2) \cdots \mathbf{M}_\varepsilon(Z_{n-1}, Z_n) \mathbf{1}^t. \tag{6}$$

The joint distribution $P(Z_1^n)$ of the HMP, as presented in (6), has the form of a product of random matrices, since the conditionals $P_X(Z_i|Z_{i-1})$ are random variables. Notice that the components of the process $\{\mathbf{M}_\varepsilon(Z_i, Z_{i+1})\}_{i\geq 1}$ take values from a set of four different matrices. Applying a subadditive ergodic theorem, it is possible to show that the expectation $\mathbf{E}[\log P(Z_1^n)]$ must converge to a constant known as the *top Lyapunov exponent* of the matrix process [11, 27, 37]). We will rely on the following result by Furstenberg and Kesten [11] (see also [27]), which formally established this fact.

[2] In general, the measures governing probability expressions will be clear from the context. In cases when confusion is possible, we will explicitly indicate the measure, e.g., $P_X(Z_n|Z_{n-1})$.

Proposition 1 (Furstenberg and Kesten [11]) *Let $\{\mathbf{M}_i\}_{i\geq 1}$, be a stationary ergodic process, where the $\mathbf{M}_i$ are square matrices such that $\mathbf{E}[\log^+ ||\mathbf{M}_i||] < \infty$ for some matrix norm $||\cdot||$ (where $\log^+ x = \max\{0, \log x\}$). Then there exists a real constant μ such that*

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{E}[\log ||\mathbf{M}_1 \cdots \mathbf{M}_n||] = \lim_{n\to\infty} \frac{1}{n} \log ||\mathbf{M}_1 \cdots \mathbf{M}_n|| = \mu \qquad \text{a.s..} \tag{7}$$

The constant μ of Proposition 1 is known as the top Lyapunov exponent of the process $\{\mathbf{M}_\varepsilon(Z_i, Z_{i+1})\}_{i\geq 1}$. An immediate consequence of this result and (6) is that the entropy of the HMP is equal to a top Lyapunov exponent.

Theorem 1 *Consider the HMP Z defined in (1). The entropy rate*

$$\begin{aligned} \boldsymbol{H}(Z) &= \lim_{n\to\infty} \mathbf{E}[\,-\frac{1}{n}\log P(Z_1^n)\,] \\ &= \lim_{n\to\infty} \frac{1}{n}\mathbf{E}[-\log\left(\mathbf{p}_1 \mathbf{M}_\varepsilon(Z_1, Z_2)\mathbf{M}_\varepsilon(Z_2, Z_3)\cdots \mathbf{M}_\varepsilon(Z_{n-1}, Z_n)\mathbf{1}^t\right)] \end{aligned}$$

is the top Lyapunov exponent of the process $\{\mathbf{M}_\varepsilon(Z_i, Z_{i+1})\}_{i\geq 1}$.

Proof. Consider a 2×2 nonnegative matrix A. It is readily verified that $\mathbf{p}_1 A \mathbf{1}^t$ is a norm of A (cf. [26]). Clearly, in the case of the nonnegative matrices $\mathbf{M}_\varepsilon(Z_i, Z_{i+1})$, the norm is upper bounded. The theorem follows by direct application of Proposition 1, using this norm. ■

We note that the connection between discrete-time, finite state Markov chains and Lyapunov exponents, based also on the Furstenberg and Kesten theorem, is studied in [13], where the dual problem of a memoryless signal going through a Markov channel is considered. The results in [13] also link other interesting parameters like mutual information and channel capacity to Lyapunov exponents.

Although some upper and lower bounds for top Lyapunov exponents are available (cf. [12]), it is, in general, notoriously difficult (if not computationally infeasible) to compute them precisely [37]. Therefore, it is of interest to study asymptotic approximations of the HMP entropy rate. Next, we derive an explicit asymptotic expansion of the Shannon entropy rate $\mathbf{H}(Z)$, which does not depend on direct computation of Lyapunov exponents.

2.1 Shannon entropy

From now on we deal with the entropy rate $\mathbf{H}(Z)$ for the HMP Z of (1) as a function of ε, when ε is small. The following formal definition will be useful in computing the Shannon entropy (and later Rényi entropies of any order) of Z

$$R_n(s,\varepsilon) = \sum_{z_1^n} P_Z^s(z_1^n), \tag{8}$$

where the exponent s of P_Z is a complex variable, and the summation is over all binary n-tuples. (The function $R_n(s,\varepsilon)$ was also used in entropy calculations in [18].) It is readily verified, using the chain rule for derivatives, that

$$H(Z_1^n) = \mathbf{E}\left[-\log P(Z_1^n)\right] = -\left.\frac{\partial}{\partial s} R_n(s,\varepsilon)\right|_{s=1}. \tag{9}$$

The entropy of the underlying Markov sequence is

$$H(X_1^n) = -\left.\frac{\partial}{\partial s} R_n(s,0)\right|_{s=1}. \tag{10}$$

Let $\boldsymbol{\pi}(1) = [\,P_X(0),\ P_X(1)\,]$ denote the stationary distribution of the binary Markov process X. Define the matrix

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

and the vector $\boldsymbol{\pi}(s) = [\,P_X^s(0),\ P_X^s(1)\,]$. It is readily verified, again by direct computation, that the entries of the mth power, $\mathbf{P}^m(s)$, of $\mathbf{P}(s)$, $m \geq 2$, are given by

$$(\mathbf{P}^m(s))_{a,b} = \sum_{z_1^{m-1}} P_X^s(z_1|a) P_X^s(z_2|z_1) \dots P_X^s(b|z_{m-1}), \quad a,b \in \{0,1\}. \tag{12}$$

the case $m = 1$ being trivially covered by the definition $p_{ab} = P_X(b|a)$. It follows that

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

Using a formal Taylor expansion near $\varepsilon = 0$, we write

$$R_n(s,\varepsilon) = R_n(s,0) + \varepsilon \frac{\partial}{\partial \varepsilon} R_n(s,\varepsilon)|_{\varepsilon=0} + O(R_{\varepsilon,\varepsilon}(s,\varepsilon')\varepsilon^2), \tag{14}$$

where $R_{\varepsilon,\varepsilon}(s,\varepsilon')$ is the second derivative with respect to ε computed at some ε', provided these derivatives exist. In fact, it is easy to verify that one needs additional assumptions for this to happen. For example, for a HMP with an underlying Markov process defined by the following transition matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix},$$

the derivatives of the entropy rate with respect to ε do not exist, and the above Taylor expansion does not hold. In fact, it was shown in [29] that the main error term (corresponding to $R_n(s,\varepsilon) - R_n(s,0)$ in our notation) is $O(\varepsilon \log(1/\varepsilon))$ in that case. However, we will prove the two lemmas below, where we write $\mathbf{P} > 0$ to denote a transition matrix all of whose entries are *positive.*

Lemma 1 *Assume $\mathbf{P} > 0$, and define $\bar{\boldsymbol{\pi}}(s) = [\,P_X^s(1),\ P_X^s(0)\,]$ (the reverse of $\boldsymbol{\pi}$),*

$$\mathbf{Q}_1(s) = \begin{bmatrix} p_{00}p_{01}^{s-1} & p_{01}p_{00}^{s-1} \\ p_{10}p_{11}^{s-1} & p_{11}p_{10}^{s-1} \end{bmatrix}, \quad \textit{and} \quad \mathbf{Q}_2(s) = \begin{bmatrix} p_{00}p_{10}^{s-1} & p_{01}p_{11}^{s-1} \\ p_{10}p_{00}^{s-1} & p_{11}p_{01}^{s-1} \end{bmatrix}. \tag{15}$$

Then,

$$\begin{aligned} \left.\frac{\partial}{\partial \varepsilon} R_n(s,\varepsilon)\right|_{\varepsilon=0} &= s\Big(\bar{\boldsymbol{\pi}}(s)\mathbf{Q}_2(s) - \boldsymbol{\pi}(s)\mathbf{P}(s)\Big)\mathbf{P}^{n-2}(s)\mathbf{1}^t \\ &+ s\boldsymbol{\pi}(s)\sum_{i=1}^{n-3}\mathbf{P}^{i-1}(s)\Big(\mathbf{Q}_1(s)\mathbf{Q}_2(s) - \mathbf{P}^2(s)\Big)\mathbf{P}^{n-i-2}(s)\mathbf{1}^t \\ &+ s\boldsymbol{\pi}(s)\mathbf{P}^{n-2}(s)\Big(\mathbf{Q}_1(s) - \mathbf{P}(s)\Big)\mathbf{1}^t. \end{aligned} \tag{16}$$

The first and third terms in (16) reflect border conditions at the beginning and end of the sequence, respectively, and their contribution will be asymptotically negligible. The second, and main, term will determine the asymptotic behavior of the entropy. Lemma 1 will be useful to derive the linear term in the asymptotic expansion of the Shannon, and later the Rényi, entropy rate. This requires, however, bounding the growth of the quadratic term in the expansion, which is accomplished in the next lemma, proved in Section 3.

Lemma 2 *Let* $\mathbf{P} > 0$. *Then for all* $\varepsilon \in [0, \frac{1}{2})$ *we have,*

$$\frac{\partial}{\partial s} \frac{\partial^2}{\partial \varepsilon^2} R_n(s, \varepsilon)|_{\varepsilon=0, s=1} = O(n) \, . \tag{17}$$

With the lemmas established, it follows from (9), (10), (14), and (17) that

$$H(Z_1^n) = H(X_1^n) - \varepsilon \frac{\partial^2}{\partial s \partial \varepsilon} R_n(s, \varepsilon) \bigg|_{\varepsilon=0, s=1} + O(n\varepsilon^2) \, . \tag{18}$$

To find the linear term in the Taylor expansion (18), we need to differentiate (16) with respect to s, and evaluate at $s = 1$. To facilitate this computation, we use the spectral representation [22, 34, 36] of the matrix $\mathbf{P}(s)$. Let $\lambda(s)$ be the main eigenvalue of $\mathbf{P}(s)$ with $\mathbf{r}_1^t(s)$ and $\boldsymbol{\ell}_1(s)$ being the corresponding right and left main eigenvectors, respectively, normalized so that $\boldsymbol{\ell}_1(s)\mathbf{r}_1^t(s) = 1$. Let also $\mu(s)$ be the second eigenvalue, with $\mathbf{r}_2^t(s)$ and $\boldsymbol{\ell}_2(s)$ the respective right and left eigenvectors. Since $\mathbf{P} > 0$, the the Perron-Frobenius theorem [22, 36] applies, and we know that $\lambda(s) > |\mu(s)|$ and $\mathbf{r}_1$ and $\boldsymbol{\ell}_1$ are real-valued nonnegative vectors. The matrix spectral representation yields

$$\begin{aligned} \mathbf{P}^k(s) &= \lambda^k(s) \, \left(\mathbf{r}_1^t(s) \, \boldsymbol{\ell}_1(s)\right) + \mu^k(s) \, \left(\mathbf{r}_2^t(s) \, \boldsymbol{\ell}_2(s)\right) \\ &= \lambda^k(s) \, \left(\mathbf{r}_1^t(s) \, \boldsymbol{\ell}_1(s)\right) (1 + O(\rho^k)) \, , \end{aligned} \tag{19}$$

where $\rho = |\mu(s)|/\lambda(s) < 1$. Note that $\left(\mathbf{r}_i^t(s) \, \boldsymbol{\ell}_i(s)\right)$ is an outer product resulting in a 2×2 matrix of rank one.

Define $\mathbf{Q}(s) = \mathbf{Q}_1(s)\mathbf{Q}_2(s)$. It follows immediately from (11) and (15) that $\mathbf{Q}(1) = \mathbf{P}^2(1)$. Therefore, when differentiating (16) with respect to s and evaluating at $s = 1$, the only terms that do not vanish in the derivative of the middle term of (16) are those involving the derivative of $\mathbf{Q}(s) - \mathbf{P}^2(s)$. Also, since $\mathbf{P}(1)$ is a positive stochastic matrix, we have $\lambda(1) = 1$. Now, substituting the spectral representation for powers of $\mathbf{P}(s)$ from (19) in (16), differentiating with respect to s and evaluating at $s = 1$, and simplifying power sums, we obtain

$$\frac{\partial^2}{\partial \varepsilon \partial s} R_n(s, \varepsilon) \bigg|_{\substack{\varepsilon=0, \\ s=1}} = n \, \boldsymbol{\pi}(1) \left(\mathbf{r}_1^t(1) \, \boldsymbol{\ell}_1(1)\right) \frac{\partial}{\partial s} \left(\mathbf{Q}(s) - \mathbf{P}^2(s)\right) \bigg|_{s=1} \left(\mathbf{r}_1^t(1) \, \boldsymbol{\ell}_1(1)\right) \mathbf{1}^t + o(n). \tag{20}$$

The $o(n)$ term in (20) is contributed by the summation of powers of ρ originating in the approximation (19), the first and last terms in (16), and the adjustment needed to make n (rather than $n-2$) the multiplier in (20).

For the transition probability matrix $\mathbf{P} = \mathbf{P}(1)$, we have

$$\boldsymbol{\ell}_1(1) = \boldsymbol{\pi}(1) = \left[\frac{p_{10}}{p_{10}+p_{01}}, \frac{p_{01}}{p_{10}+p_{01}} \right],$$

and $\mathbf{r}_1(1) = [1,\, 1]$. Thus, $\boldsymbol{\pi}(1)\mathbf{r}_1^t(1) = \boldsymbol{\ell}_1(1)\mathbf{1}^t = 1$, and (20) simplifies to

$$\frac{\partial^2}{\partial\varepsilon\partial s} R_n(s,\varepsilon)\Big|_{\substack{\varepsilon=0,\\ s=1}} = n\,\boldsymbol{\pi}(1)\,\frac{\partial}{\partial s}\left(\mathbf{Q}(s) - \mathbf{P}^2(s)\right)\Big|_{s=1} \mathbf{1}^t + o(n). \tag{21}$$

The derivative in (21) is obtained through a rather straightforward symbolic manipulation. We show the derivative of each term, evaluated at $s = 1$, to give some insight into how the final result takes its form. We have

$$\frac{\partial \mathbf{P}^2(s)}{\partial s}\Big|_{s=1} = \begin{bmatrix} 2p_{00}^2 \log(p_{00}) + p_{01}p_{10}\log(p_{01}p_{p10}) & p_{00}p_{01}\log(p_{00}p_{01}) + p_{01}p_{11}\log(p_{01}p_{11}) \\ p_{10}p_{00}\log(p_{10}p_{00}) + p_{11}p_{10}\log(p_{11}p_{10}) & p_{10}p_{01}\log(p_{10}p_{01}) + 2p_{11}^2\log(p_{11}) \end{bmatrix}, \tag{22}$$

and

$$\frac{\partial \mathbf{Q}(s)}{\partial s}\Big|_{s=1} = \begin{bmatrix} p_{00}^2 \log(p_{01}p_{10}) + +2p_{01}p_{10}\log(p_{00}) & p_{00}p_{01}\log(p_{01}p_{11}) + p_{01}p_{11}\log(p_{00}p_{01}) \\ p_{10}p_{00}\log(p_{11}p_{10}) + p_{11}p_{10}\log(p_{10}p_{00}) & 2p_{10}p_{01}\log(p_{11}) + p_{11}^2\log p_{10}p_{01} \end{bmatrix}. \tag{23}$$

Putting it all together, from (18), (21), (22) and (23), after some symbolic manipulation and rearrangement of terms, we obtain the following result for the Shannon entropy.

Theorem 2 *Let* $\mathbf{P} > 0$*. The first order term in the entropy rate of the process* Z*,*

$$\boldsymbol{H}(Z) = \lim_{n\to\infty} \frac{1}{n} H_n(Z^n) = \boldsymbol{H}(X) + f_1(p_{01}, p_{10})\varepsilon + O(\varepsilon^2), \tag{24}$$

is given by

$$f_1(p_{01}, p_{10}) = \mathbb{D}\left(P_X(z_1z_2z_3)||P_X(z_1\bar{z}_2z_3)\right) = \sum_{z_1z_2z_3} P_X(z_1z_2z_3)\log\frac{P_X(z_1z_2z_3)}{P_X(z_1\bar{z}_2z_3)}, \tag{25}$$

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

Remark. The expansion of Theorem 2 does not hold, in general, when the condition $\mathbf{P} > 0$ is not satisfied, i.e., some transition probabilities are zero. For example, consider the Markov chain discussed above Lemma 1, with the following general transition probabilities

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ 1 & 0 \end{bmatrix} \tag{26}$$

where $0 \leq p \leq 1$. Clearly, some sequences x_1^n are not reachable by the Markov process in this case, i.e., the set of sequences of nonzero probability under the process is *constrained*. This set of sequences, however, may have nonzero probability with respect to the channel output process Z. The probability of the set in this case is polynomial in ε, which will generally contribute a term $O(\varepsilon \log \varepsilon)$ to the entropy rate $H(Z)$ when ε is small. This was observed for the transition matrix $\mathbf{P}$ of (26) in [29], where it was shown that

$$H(Z) = H(X) - \frac{p(2-p)}{1+p}\varepsilon \log \varepsilon + O(\varepsilon) \tag{27}$$

as $\varepsilon \to 0$. Recently, Han and Marcus [15] generalized this result to HMPs with underlying Markov processes of arbitrary order, showing that, in general,

$$H(Z) = H(X) - f_0(\mathbf{P})\varepsilon \log \varepsilon + O(\varepsilon) \tag{28}$$

when at least one of the transition probabilities in the Markov chain is zero. This setting is closely related to the long-standing *noisy constrained capacity* problem [9], originally posed by Shannon. The link between the two problems is also studied in [20], where the coefficient $f_0(\mathbf{P})$ in (28) is determined explicitly for the important family of (d, k) constraints [25].

The methodology leading to the proof of Theorem 2 allows for the computation of further terms of the Taylor expansion, provided one bounds the error term as done in Lemma 2. For example, after verifying that the third derivative of $\mathbf{H}(Z_1^n)$ with respect to ε is $O(n)$, we obtain an explicit expression for the second derivative of $\mathbf{H}(Z_1^n)$ at $\varepsilon = 0$, denoted $f_2(p_{01}, p_{10})$ (this coefficient multiplies $\varepsilon^2/2$ in the Taylor expansion). Let

$$\mathbf{Q}_3(s) = \begin{bmatrix} p_{00}p_{11}^{s-1} & p_{01}p_{10}^{s-1} \\ p_{10}p_{01}^{s-1} & p_{11}p_{00}^{s-1} \end{bmatrix}. \tag{29}$$

Then,

$$\begin{aligned} f_2(p_{01}, p_{10}) &= -f_1(p_{0,1}, p_{10}) - \boldsymbol{\pi}(1)[\mathbf{Q}_1(0)\mathbf{Q}_2(0) - \mathbf{P}^2(0)]\mathbf{r}_1(1) \\ &- \boldsymbol{\ell}_1(1)\frac{\partial}{\partial s}\left(\mathbf{P}^3(s) - \mathbf{P}^2(s)\mathbf{Q}_2(s) - \mathbf{Q}_3(s)\mathbf{Q}_1(s)\mathbf{P}(s) + \mathbf{Q}_3^2(s)\mathbf{P}(s)\right)\Big|_{s=1} \mathbf{1}^t, \end{aligned} \tag{30}$$

As above, $\boldsymbol{\ell}_1(1)$ and $\mathbf{r}_1(1)$ are left and right eigenvectors of $\mathbf{P}(1)$. At the end of Section 3.2 we present an outline of the derivation of (30), using tools similar to those leading to Theorem 2.

We illustrate these results in the following example.

Example. *Symmetric Markov Process*

Consider a Markov process with symmetric transition probabilities $p_{01} = p_{10} = p$, $p_{00} = p_{11} = 1-p$. 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 $P_X(000) = P_X(111) = \frac{1}{2}(1-p)^2$, $P_X(001) = P_X(011) = P_X(100) = P_X(110) = \frac{1}{2}p(1-p)$, $P_X(010) = P_X(101) = p^2$. Substituting these values into (25), we obtain

$$f_1(p, p) = 2(1-2p)\log\frac{1-p}{p}, \tag{31}$$

Parameters			Calculated				Empirical
ε	p	n	$\mathbf{H}(X)$	$f_1(p,p)$	$f_2(p,p)$	$\mathbf{H}(X)+$ $f_1\varepsilon+f_2\frac{\varepsilon^2}{2}$	$\frac{-\log P_Z(z_1^n)}{n}$
0.001	0.005	$4\cdot10^9$	0.031	10.481	-19810.0	0.032	0.039
0.001	0.010	$4\cdot10^9$	0.056	9.006	-4908.5	0.063	0.063
0.001	0.025	$1\cdot10^9$	0.117	6.961	-766.5	0.123	0.123
0.010	0.050	$1\cdot10^8$	0.199	5.300	-184.8	0.242	0.242
0.010	0.100	$1\cdot10^8$	0.325	3.516	-43.0	0.358	0.357
0.010	0.300	$1\cdot10^8$	0.611	0.678	-2.5	0.618	0.617

Table 1: Second order Taylor approximation of $\mathbf{H}(Z)$ vs. empirical estimation.

and, for the second order term, from (30),

$$f_2(p,p) = -f_1(p,p) - \frac{1}{2}\left(\frac{2p-1}{p(1-p)}\right)^2 . \tag{32}$$

This expression coincides with the one given in [39] for the second derivative of the entropy rate of a symmetric HMP.

HMPs for various values of the parameters ε and $p_{01} = p_{10} = p$ were simulated, generating pseudo-random HMP sequences of lengths between $n = 10^8$ and $n = 4\cdot10^9$. 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 of the entropy rate. This is justified by the fact that a sequence emitted by the HMP is "typical" and, thus, satisfies $|\frac{1}{n}\log P_Z(z_1^n) + \mathbf{H}_n(Z)| < \delta$ for any $\delta > 0$, with probability that approaches one exponentially fast as $n\to\infty$ [5]. A sample of results for some values of ε and p are given in Table 1, where values of $\mathbf{H}(Z)$ estimated using a second-order Taylor approximation according to (31) and (32) are compared with simulation estimates. The slope $\partial\mathbf{H}_n(Z)/\partial\varepsilon|_{\varepsilon=0}$, as a function of p, is plotted in Figure 1. The empirical slope was estimated using first differences of the estimated values of $\mathbf{H}_n(Z)$ near $\varepsilon=0$, and the result compared with the analytical value produced by (31).

The methods leading to the results of this section can also be applied to HMPs with underlying Markov processes of higher order. Consider a binary stationary Markov process, X, of order r. The process is defined by the set of conditional probabilities $P(X_t{=}a_{r+1}\,|\,X_{t-r}^{t-1}{=}a_1^r)$, $a_1^{r+1}\in\{0,1\}^{r+1}$ (we assume X_{-r+1}^0 is defined and distributed according to the stationary distribution of the process). The HMP is still defined by Equation (1), with a Bernoulli process $\{E_i\}$. A generalization of Theorem 1 holds also in this case, where the process $\{\mathbf{M}_\varepsilon(Z_i,Z_{i+1})\}_{i\geq1}$ is supported by a set of 2^{r+1} binary matrices of dimensions $2^r\times2^r$. The generalization of Theorem 2, in turn, takes the following form.

Theorem 3 *Let* $\mathbf{P}$ *be a stationary binary Markov process such that* $P(X_t = a_{r+1}|X_{t-r}^{t-1} = a_1^r) > 0$ *for all* $a_1^{r+1}\in\{0,1\}^{r+1}$. *For a binary sequence* z_1^{2r+1}, *let* $\check{z}_1^{2r+1}$ *denote a sequence that is identical to* z_1^{2r+1}, *except for the* $(r{+}1)$*st coordinate, where* $\check{z}_{r+1} = \bar{z}_{r+1}$. *The first*

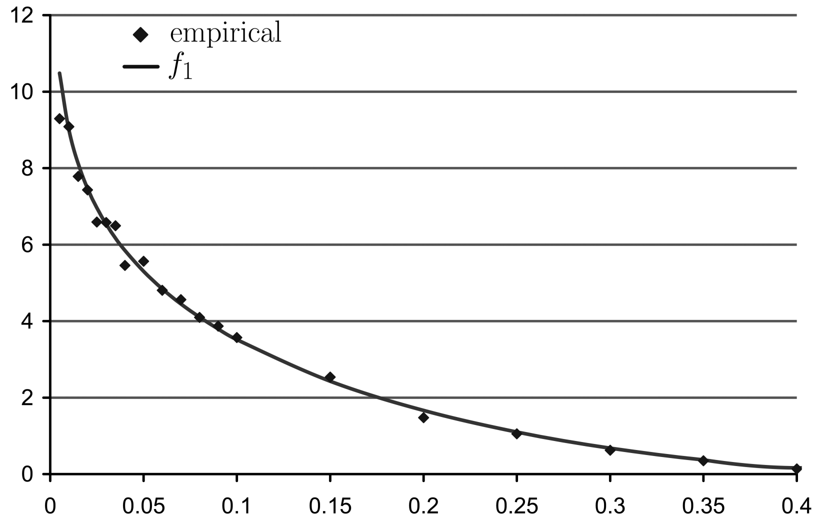

Figure 1: Values of f_1 and empirical estimation of $\partial h/\partial\varepsilon|_{\varepsilon=0}$ as a function of p (entropy measured in nats).

order term in the entropy rate of the process Z*,*

$$\boldsymbol{H}(Z) = \lim_{n\to\infty} \frac{1}{n} H_n(Z^n) = \boldsymbol{H}(X) + f_1(P)\varepsilon + O(\varepsilon^2), \tag{33}$$

is given by

$$f_1(P) = \mathbb{D}\left(P_X(z_1^{2r+1})||P_X(\check{z}_1^{2r+1})\right) = \sum_{z_1^{2r+1}} P_X(\check{z}_1^{2r+1}) \log \frac{P_X(z_1^{2r+1})}{P_X(\check{z}_1^{2r+1})}, \tag{34}$$

where the summation is over all binary strings of length $2r+1$*.*

The proof of Theorem 3 follows along the same lines as that of Theorem 2 (and, of course, generalizes it), but with more cumbersome notation and symbolic expressions. Therefore, for clarity, in the next section we provide details for the proof of Theorem 2, but just a sketch of the proof of Theorem 3, pointing out the parallelism between the two.

2.2 Rényi entropy

We next deal with the Rényi [33] entropy, H_s, of order s, defined as

$$H_s(Z_1^n) = \frac{\log R_n(s,\varepsilon)}{1-s}. \tag{35}$$

As in the case of the Shannon entropy, we focus our attention on deriving an asymptotic approximation of $\mathbf{H}_s(Z)$, and, in particular, on the first order error term of the Taylor expansion around $\varepsilon = 0$. Observe first, however, that the Rényi entropy of the underlying Markov process can be expressed as (cf. [36])

$$\mathbf{H}_s(X) = \frac{1}{1-s} \log \lambda(s),$$

where $\lambda(s)$ is the main eigenvalue of the matrix $\mathbf{P}$ of (11). Using arguments similar to those of Section 2.1, we derive the following result, proved in Section 3.

Theorem 4 *If* $\mathbf{P} > 0$, *then for any* s *and small* ε

$$\boldsymbol{H}_s(Z) = \boldsymbol{H}_s(X) + \varepsilon \frac{s}{(1-s)\lambda^2(s)} \boldsymbol{\ell}_1(s) \left(\mathbf{Q}_1(s)\mathbf{Q}_2(s) - \mathbf{P}^2(s) \right) \mathbf{r}_1(s) + O(\varepsilon^2), \tag{36}$$

where $\boldsymbol{\ell}_1(s)$ *and* $\mathbf{r}_1(s)$ *are, respectively, the left and the right main eigenvectors of* $\mathbf{P}(s)$, *and* $\mathbf{Q}_1(s)$ *and* $\mathbf{Q}_2(s)$ *are defined in (15).*

3 Analysis and Proofs

In this section we first prove Lemmas 1 and 2 leading directly to the proof of Theorem 2. We also present an outline of the derivation of the second order term in the expansion of the Shannon entropy rate, given in (30). We then proceed to the proof of Theorem 4. Throughout this section we assume that $\mathbf{P} > 0$.

3.1 Proof of Lemma 1

Our goal is to prove Equation (16) of Lemma 1. Recall the definition of the matrices $\mathbf{M}_\varepsilon(Z_i, Z_{i+1})$ in (4). We construct these matrices for a given realization z_1^n of Z_1^n. Also, to reduce clutter, we will use the abbreviated notation $\mathbf{M}_i = \mathbf{M}_\varepsilon(z_i, z_{i+1})$. Hence, we have

$$\begin{aligned}
\mathbf{M}_i &= \begin{bmatrix} (1-\varepsilon)P_X(z_{i+1}|z_i) & \varepsilon P_X(\bar{z}_{i+1}|z_i) \\ (1-\varepsilon)P_X(z_{i+1}|\bar{z}_i) & \varepsilon P_X(\bar{z}_{i+1}|\bar{z}_i) \end{bmatrix} \\
&= \begin{bmatrix} P_X(z_{i+1}|z_i) & 0 \\ P_X(z_{i+1}|\bar{z}_i) & 0 \end{bmatrix} + \varepsilon \begin{bmatrix} -P_X(z_{i+1}|z_i) & P_X(\bar{z}_{i+1}|z_i) \\ -P_X(z_{i+1}|\bar{z}_i) & P_X(\bar{z}_{i+1}|\bar{z}_i) \end{bmatrix} \\
&\stackrel{\text{def}}{=} \mathbf{M}_i^{(0)} + \varepsilon \mathbf{M}_i^{(1)}, \qquad 1 \le i \le n-1\,.
\end{aligned} \tag{37}$$

Similarly, for the vector $\mathbf{p}_1$ defined in (3), we have

$$\mathbf{p}_1 = [\,P_X(z_1),\ 0\,] + \varepsilon[\,-P_X(z_1),\ P_X(\bar{z}_1)\,] \stackrel{\text{def}}{=} \mathbf{p}_1^{(0)} + \varepsilon \mathbf{p}_1^{(1)}\,. \tag{38}$$

Now, we recall the expression (6) for the probability, which we recast as

$$\begin{aligned}
P_Z(z_1^n) &= \mathbf{p}_1 \mathbf{M}_1 \mathbf{M}_2 \cdots \mathbf{M}_{n-1} \mathbf{1}^t \\
&= (\mathbf{p}_1^{(0)} + \varepsilon \mathbf{p}_1^{(1)})(\mathbf{M}_1^{(0)} + \varepsilon \mathbf{M}_1^{(1)})(\mathbf{M}_2^{(0)} + \varepsilon \mathbf{M}_2^{(1)}) \cdots (\mathbf{M}_{n-1}^{(0)} + \varepsilon \mathbf{M}_{n-1}^{(1)}) \mathbf{1}^t\,.
\end{aligned} \tag{39}$$

To compute the derivative of $P_Z(z_1^n)$ at $\varepsilon = 0$, we differentiate both sides of (39), obtaining

$$\begin{aligned}
\frac{\partial}{\partial \varepsilon} P_Z(z_1^n)\Big|_{\varepsilon=0} &= \mathbf{p}_1^{(1)} \prod_{j=1}^{n-1} \mathbf{M}_j^{(0)} \mathbf{1}^t + \\
&\quad \mathbf{p}_1^{(0)} \sum_{i=1}^{n-2} \left[\prod_{j=1}^{i-1} \mathbf{M}_j^{(0)} \cdot \mathbf{M}_i^{(1)} \cdot \prod_{j=i+1}^{n-1} \mathbf{M}_j^{(0)} \right] \mathbf{1}^t + \\
&\quad \mathbf{p}_1^{(0)} \prod_{j=1}^{n-2} \mathbf{M}_j^{(0)} \cdot \mathbf{M}_{n-1}^{(1)} \mathbf{1}^t\,.
\end{aligned} \tag{40}$$

Focusing on a typical term in the summation on the second line of (40), it follows from the definitions in (37)–(38) that

$$\mathbf{p}_1^{(0)} \prod_{j=1}^{i-1} \mathbf{M}_j^{(0)} = \left[\, P_X(z_1^i),\ 0 \,\right] .$$

Multiplication by $\mathbf{M}_i^{(1)}$ yields the vector $[\,-P_X(z_1^{i+1}),\ P_X(z_1^i \bar{z}_{i+1})\,]$, and further multiplication by $\prod_{j=i+1}^{n-1} \mathbf{M}_j^{(0)} \mathbf{1}^t$ results in

$$\mathbf{p}_1^{(0)} \prod_{j=1}^{i-1} \mathbf{M}_j^{(0)} \cdot \mathbf{M}_i^{(1)} \cdot \prod_{j=i+1}^{n-1} \mathbf{M}_j^{(0)} \mathbf{1}^t \;=\; P_X(z_1^i \bar{z}_{i+1} z_{i+2}^n) - P_X(z_1^n)$$

The first and third lines of (40) deal with the edge cases ($i = 0$ and $i = n-1$), but are otherwise similar. Let $\mathbf{e}_i$ denote a binary unit vector of length n, with a one in the ith coordinate. It follows from the foregoing discussion that

$$\frac{\partial}{\partial \varepsilon} P_Z(z_1^n)\bigg|_{\varepsilon=0} = -P_X(z_1^n) + \sum_{i=1}^{n} P_X(z_1^n \oplus \mathbf{e}_i) \tag{41}$$

$$\begin{aligned} = \;\; & -P_X(z_1^n) + \Bigg(P_X(\bar{z}_1) P_X(z_2|\bar{z}_1) \prod_{j=2}^{n-1} P_X(z_{j+1}|z_j) + \\ & \sum_{i=2}^{n-2} P_X(z_1) \prod_{j=1}^{i-2} P_X(z_{j+1}|z_j) \cdot P_X(\bar{z}_i|z_{i-1}) P_X(z_{i+1}|\bar{z}_i) \cdot \prod_{j=i+1}^{n-1} P_X(z_{j+1}|z_j) + \\ & P_X(z_1) \prod_{j=1}^{n-2} P_X(z_{j+1}|z_j) \cdot P_X(\bar{z}_n|z_{n-1}) \Bigg) \mathbf{1}^t . \end{aligned} \tag{42}$$

By the definition of $R(s,\varepsilon)$ in (8), we have

$$\frac{\partial}{\partial \varepsilon} R(s,\varepsilon)\bigg|_{\varepsilon=0} \;=\; \sum_{z_1^n} s P_Z^{s-1}(z^n)\, \frac{\partial}{\partial \varepsilon} P_Z(z_1^n)\bigg|_{\varepsilon=0} \tag{43}$$

Applying the definitions of $\mathbf{P}(s), \mathbf{Q}_1(s)$, and $\mathbf{Q}_2(s)$ in (11) and (15), and recalling the characterization of powers of $\mathbf{P}(s)$ in (12), we verify

$$\bar{\boldsymbol{\pi}}(s)\mathbf{Q}_2(s)\mathbf{P}^{n-2}(s)\mathbf{1}^t \;=\; \sum_{z_1^n} P_X^{s-1}(z_1^n) P_X(\bar{z}_1 z_2^n), \tag{44}$$

$$\boldsymbol{\pi}(s) P_X^{i-1}(s)\mathbf{Q}_1(s)\mathbf{Q}_2(s)\mathbf{P}^{n-i-2}(s)\mathbf{1}^t \;=\; \sum_{z_1^n} P_X^{s-1}(z_1^n) P_X(z_1^{i-1}\bar{z}_i z_{i+1}^n), \;\; 1 \le i \le n-1, \tag{45}$$

$$\boldsymbol{\pi}(s)\mathbf{P}^{n-2}(s)\mathbf{Q}_1(s)\mathbf{1}^t \;=\; \sum_{z_1^n} P_X^{s-1}(z_1^n) P_X(z_1^{n-1}\bar{z}_n). \tag{46}$$

Equation (16) now follows from (42)–(46). ■

3.2 Proof of Lemma 2

Here we are to prove that the second derivative of $H(Z_1^n)$ with respect to ε is $O(n)$. We have

$$\frac{\partial^2}{\partial\varepsilon^2}H(Z_1^n) = -\sum_{z_1^n}\frac{\partial^2 P(z_1^n)}{\partial\varepsilon^2}\log P(z_1^n) - \sum_{z_1^n}\left(\frac{\partial P(z_1^n)}{\partial\varepsilon}\right)^2\frac{1}{P(z_1^n)}.$$

It follows from (37) that

$$\frac{\partial}{\partial\varepsilon}\mathbf{M}_\varepsilon(a,b) = \frac{\mathbf{M}_{1-\varepsilon}(a,b) - \mathbf{M}_\varepsilon(a,b)}{1-2\varepsilon}. \tag{47}$$

Now, from (39), applying (47), we obtain

$$\frac{\partial}{\partial\varepsilon}P(z_1^n) = \frac{1}{1-2\varepsilon}\sum_{i=1}^{n}\left(P_X(z_1^n\oplus\mathbf{e}_i) - P_X(z_1^n)\right). \tag{48}$$

Using (48), we write

$$\frac{\partial^2 P(z_1^n)}{\partial\varepsilon^2} = \\ \frac{2}{1-2\varepsilon}\frac{\partial P(z_1^n)}{\partial\varepsilon} + \frac{1}{(1-2\varepsilon)^2}\sum_{1\le j,k\le n}\Big(P_X(z_1^n\oplus\mathbf{e}_j\oplus\mathbf{e}_k) - P_X(z_1^n\oplus\mathbf{e}_j) - P_X(z_1^n\oplus\mathbf{e}_k) + P_X(z_1^n)\Big).$$

Thus,

$$\begin{aligned}\frac{\partial^2 H(Z_1^n)}{\partial\varepsilon^2} &= -\frac{2}{1-2\varepsilon}\frac{\partial H(Z_1^n)}{\partial\varepsilon} - \sum_{z_1^n}\left(\frac{\partial P(z_1^n)}{\partial\varepsilon}\right)^2\frac{1}{P(z_1^n)} - \\ &\frac{1}{(1-2\varepsilon)^2}\sum_{1\le j,k\le n}\sum_{z_1^n}\Big(P_X(z_1^n\oplus\mathbf{e}_j\oplus\mathbf{e}_k) - P_X(z_1^n\oplus\mathbf{e}_j) - P_X(z_1^n\oplus\mathbf{e}_k) + P_X(z_1^n)\Big)\log P_X(z_1^n).\end{aligned} \tag{49}$$

Observe that we already established, in (21), that the first term of (49) is $O(n)$. Hence we only deal with the other two terms that we denote by D_2 and D_1, respectively.

We first estimate D_1 defined as

$$D_1 = \sum_{1\le j,k\le n}\sum_{z_1^n}\Big(P_X(z_1^n\oplus\mathbf{e}_j\oplus\mathbf{e}_k) - P_X(z_1^n\oplus\mathbf{e}_j) - P_X(z_1^n\oplus\mathbf{e}_k) + P_X(z_1^n)\Big)\log P_X(z_1^n).$$

Observe that

$$\begin{aligned}D_1 &= \sum_{1\le j,k\le n}\sum_{z_1^n}P_X(z_1^n)\Big(\log P_X(z_1^n\oplus\mathbf{e}_j\oplus\mathbf{e}_k) - \log P_X(z_1^n\oplus\mathbf{e}_j) - \log P_X(z_1^n\oplus\mathbf{e}_k) + \log P_X(z_1^n)\Big)\\ &= \sum_{1\le j,k\le n}\sum_{z_1^n}P_X(z_1^n)\log\frac{P_X(z_1^n\oplus\mathbf{e}_j\oplus\mathbf{e}_k)P_X(z_1^n)}{P_X(z_1^n\oplus\mathbf{e}_j)P_X(z_1^n\oplus\mathbf{e}_k)}\end{aligned}$$

To complete our derivation we will use the following lemma, which we prove at the end of this section.

Lemma 3 *There exists $\rho < 1$ such that*

$$\frac{P_X(z_1^n \oplus \mathbf{e}_j \oplus \mathbf{e}_k)P_X(z_1^n)}{P_X(z_1^n \oplus \mathbf{e}_j)P_X(z_1^n \oplus \mathbf{e}_k)} = 1 + O(\rho^k) + O(\rho^j) + O(\rho^{|j-k|}) + O(\rho^{n-j}) + O(\rho^{n-k}) \qquad (50)$$

uniformly over all z_1^n.

Granted Lemma 3, we proceed as follows

$$\begin{aligned} D_1 &= \sum_{1\le j,k\le n} \sum_{z_1^n} P_X(z_1^n) \log \frac{P_X(z_1^n \oplus \mathbf{e}_j \oplus \mathbf{e}_k)P_X(z_1^n)}{P_X(z_1^n \oplus \mathbf{e}_j)P_X(z_1^n \oplus \mathbf{e}_k)} \\ &= \sum_{z_1^n} P_X(z_1^n) \sum_{1\le j,k\le n} \log\left(1 + O(\rho^k) + O(\rho^j) + O(\rho^{|j-k|}) + O(\rho^{n-j}) + O(\rho^{n-k})\right) \\ &= \sum_{z_1^n} P_X(z_1^n) O\left(\frac{n}{1-\rho}\right) = O(n) \end{aligned}$$

as needed.

Now we deal with D_2 defined as

$$D_2 = \sum_{z_1^n} \left(\frac{\partial P(z_1^n)}{\partial \varepsilon}\right)^2 \frac{1}{P(z_1^n)}.$$

Using (48) we find

$$\begin{aligned} D_2 &= \sum_{z_1^n} \sum_{1\le j,k\le n} \left(P_X(z_1^n \oplus \mathbf{e}_j) - P_X(z_1^n)\right)\left(P_X(z_1^n \oplus \mathbf{e}_k) - P_X(z_1^n)\right) \frac{1}{P(z_1^n)} \\ &= \sum_{z_1^n} \sum_{1\le j,k\le n} \Bigg(P_X(z_1^n) - P_X(z_1^n \oplus \mathbf{e}_j) - P_X(z_1^n \oplus \mathbf{e}_k) + P_X(z_1^n \oplus \mathbf{e}_j \oplus \mathbf{e}_k) \\ &\quad + \left(\frac{P_X(z_1^n \oplus \mathbf{e}_j)P_X(z_1^n \oplus \oplus \mathbf{e}_k)}{P_X(z_1^n \oplus \mathbf{e}_j \oplus \mathbf{e}_k)P_X(z_1^n)} - 1\right) P_X(z_1^n \oplus \mathbf{e}_j \oplus \mathbf{e}_k)\Bigg) \\ &= 0 + 0 + \sum_{z_1^n} P_X(z_1^n) O\left(\frac{n}{1-\rho}\right) \\ &= O(n) \end{aligned}$$

where the first two zeros are due to $0 = \sum_{z_1^n} \left(P_X(z_1^n) - P_X(z_1^n \oplus \mathbf{e}_j)\right)$, while the last estimate follows from Lemma 3.

To complete the proof of Lemma 2 it remains to establish Lemma 3.

Proof of Lemma 3. To facilitate the parsing of matrix formulas, we introduce a somewhat redundant notation for inner and outer products. For row vectors $\mathbf{l}$ and $\mathbf{r}$ we write $\langle \mathbf{l}, \mathbf{r} \rangle$ for the scalar product, while $\mathbf{r} \otimes \mathbf{l}$ is the outer (tensor) product. Furthermore, for a matrix $\mathbf{M}$ and row vectors $\mathbf{l}$ and $\mathbf{r}$ we write $\langle \mathbf{l}, \mathbf{M}, \mathbf{r} \rangle := \langle \mathbf{l}\mathbf{M}, \mathbf{r} \rangle = \langle \mathbf{l}, \mathbf{M}\mathbf{r}^t \rangle$.

Let now $\{\mathbf{B}_i\}_{i\geq 1}$ be a set of positive matrices. We write $\mathbf{B}_i^j = \mathbf{B}_i \cdots \mathbf{B}_j$. Seneta in Section 3.2 of [34] presented an interesting generalization of the Perron-Frobenius theorem for positive *set* of matrices $\{\mathbf{B}_i\}_{i\geq 1}$ that we briefly review. Let $\mathbf{l}(\mathbf{B}_i^j)$ and $\mathbf{r}(\mathbf{B}_i^j)$ be left and right main eigenvectors of $\mathbf{B}_i^j$ corresponding to the main eigenvalue $\lambda(\mathbf{B}_i^j)$. Corollary 2 of Section 3.2 in [34] asserts that there exists $0 < \rho < 1$ such that

$$\mathbf{B}_i^j = \mathbf{B}_i \cdots \mathbf{B}_j = \lambda(\mathbf{B}_i^j)\mathbf{r}(\mathbf{B}_i^j) \otimes \mathbf{l}(\mathbf{B}_i^j)\left(1 + O(\rho^{|j-i|})\right), \tag{51}$$

that is, positive matrices are well approximated by the rank one matrix $\mathbf{r}(\mathbf{B}_i^j) \otimes \mathbf{l}(\mathbf{B}_i^j)$.

We apply (51) to our problem, in particular to the subset $\mathbf{M}(z_i^j) = \mathbf{M}(z_i, z_{i+1}) \cdots \mathbf{M}(z_{j-1}, z_j)$ of matrices $\{\mathbf{M}(z_k, z_{k+1})\}_{k=0}^{n-1}$. With notations as above we find that

$$\mathbf{M}(z_i^j) = \lambda(z_i^j)\mathbf{r}(z_i^j) \otimes \mathbf{l}(z_i^j)\left(1 + O(\rho^{|j-i|})\right).$$

Applying above to the matrix product representation (6) of $P(z_1^n)$ with $\mathbf{p}_1 = \boldsymbol{\pi}$ we arrive at

$$\begin{aligned} P(z_1^n) &= \lambda(z_1^{j-1})\lambda(z_{j+1}^{k-1})\lambda(z_{k+1}^n)\langle\boldsymbol{\pi}, \mathbf{r}(z_1^{j-1})\rangle\langle\mathbf{l}(z_1^{j-1}), \mathbf{M}(z_{j-1}^{j+1}), \mathbf{r}(z_{j+1}^{k-1})\rangle \\ &\quad \langle\mathbf{l}(z_{j+1}^{k-1}), \mathbf{M}(z_{k-1}^{k+1})\mathbf{r}(z_{k+1}^n)\rangle\langle\mathbf{l}(z_{k+1}^n)\mathbf{1}\rangle \cdot \left(1 + O(\rho^j + \rho^k + \rho^{|k-j|} + \rho^{n-k})\right) \end{aligned}$$

Applying the above to $P(z_1^n \oplus \mathbf{e}_j)$, $P(z_1^n \oplus \mathbf{e}_k)$ and $P(z_1^n \oplus \mathbf{e}_j + \mathbf{e}_k k)$ we obtain the same formulas except that $\mathbf{M}(z_{j-1}^{j+1})$ is replaced by $\mathbf{M}(z_{j-1}, \bar{z}_j, z_{j+1})$ and $\mathbf{M}(z_{k-1}^{k+1})$ is replaced by $\mathbf{M}(z_{k-1}\bar{z}_k z_{k+1})$ which cancel out in the ratio (50) proving Lemma 3. The proof of Theorem 2 is now completed. ∎

Finally, let us briefly show how we derived the second term of the Taylor expansion presented in the remark after Theorem 2. Thus we want to establish (30). Observe first that

$$\frac{\partial^2 P^s(z_1^n)}{\partial\varepsilon^2} = \frac{1}{s}\frac{\partial P^s(z_1^n)}{\partial\varepsilon} + s(s-1)P^{s-2}(z_1^n) + sP^{s-1}(z_1^n)\frac{\partial^2 P(z_1^n)}{\partial\varepsilon^2}. \tag{52}$$

When computing the derivative with respect to s at $s = 1$, the first term above will lead to f_1 term of Theorem 2, the second term gives us the second term of (30), thus we are left with the third term that we compute now.

With the notation as in Section 3.1, we find

$$\begin{aligned} \frac{\partial^2 P(z_1^n)}{\partial\varepsilon^2} &= \sum_{k=1}^{n-1} \mathbf{M}_0^{(0)}\mathbf{M}_1^{(0)} \cdots \mathbf{M}_{k-1}^{(1)}\mathbf{M}_k^{(1)}\mathbf{M}_{i+1}^{(0)} \cdots \mathbf{M}_{n-1}^{(0)}\,\mathbf{1}^t \\ &+ \sum_{|i-j|>1} \mathbf{M}_0^{(0)} \cdots \mathbf{M}_i^{(1)} \cdots \mathbf{M}_j^{(1)} \cdots \mathbf{M}_{n-1}^{(0)}\,\mathbf{1}^t \end{aligned}$$

But the second term of the above will give us zero when differentiating with respect to s at $s = 1$, so we only consider the first term which we can write as

$$\begin{aligned} &\sum_{k=1}^{n-1} P_X(z_1^{k-1})\Big(P_X(z_{k+2}z_{k+1}z_k|z_{k-1}) - P_X(z_{k+2}|\bar{z}_{k+1})P_X(z_{k+1}\bar{z}_k|z_{k-1}) \\ &-P_X(z_{k+2}|z_{k+1})P_X(\bar{z}_{k+1}|z_k)P_X(z_k|z_{k-1}) + P_X(z_{k+2}|\bar{z}_{k+1})P_X(\bar{z}_{k+1}\bar{z}_k|z_{k-1})\Big)P_X(z_{k+3}^n|z_{k+2}). \end{aligned}$$

Thus the third term of (52) can be written in matrix form as follows

$$s\boldsymbol{\pi}(s)\sum_{k=1}^{n-1}\mathbf{P}^{k-1}(s)\Big(\mathbf{P}^3(s)-\mathbf{P}^2(s)\mathbf{Q}_2(s)-\mathbf{Q}_1^2(s)\mathbf{Q}_2(s)+\mathbf{Q}_1(s)\mathbf{Q}_2(s)\mathbf{Q}_3(s)\Big)\mathbf{P}^{n-k-3}\mathbf{1}^t$$
$$+s\boldsymbol{\pi}(s)\sum_{|i-j|>1}\mathbf{P}^{i-1}(s)\Big(\mathbf{Q}_1(s)\mathbf{Q}_2(s)-\mathbf{P}^2(s)\Big)\mathbf{P}^{|j-i-2|}(s)[\mathbf{Q}_1(s)\mathbf{Q}_2(s)-\mathbf{P}^2(s)]\mathbf{P}^{n-j-2}(s)\mathbf{1}^t$$

where $\mathbf{Q}_3(s)$ is defined in (29). From the above it should be clear why the derivative of second term is zero at $s=1$. Thus the only contribution comes from the first term (which obviously is $O(n)$, as desired). This completes a brief derivation of (30).

3.3 Sketch of Proof of Theorem 3

To derive an analogue of (5), we consider the Markov chain of *states* $s_t = X_{t-r}^{t-1}$, $t>0$ (we assume X_{-r+1}^{0} is defined and distributed according to the stationary distribution of the process) of the rth order Markov process X. Thus, we will focus on r-symbol sliding windows of the binary processes of interest. In what follows, vectors are of dimension 2^r, and matrices are of dimensions $2^r\times 2^r$ (e.g., $\mathbf{1}$ is now a row vector of 2^r ones). Entries in vectors and matrices are indexed by vectors in $\{0,1\}^r$, according to some fixed order, so that $\{0,1\}^r=\{\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_{2^r}\}$. Reasoning as in the derivation of (5), we obtain

$$P(Z_1^n)=\mathbf{p}_1\mathbf{M}_\varepsilon(Z_2^{r+1},Z_1^r)\cdots\mathbf{M}_\varepsilon(Z_{n-r+1}^n,Z_{n-r}^{n-1})\mathbf{1}^t, \tag{53}$$

where

$$\mathbf{p}_1=\Big[P(Z_1^r,\,E_1^r{=}\mathbf{a}_1),\,P(Z_1^r,E_1^r{=}\mathbf{a}_2),\,\ldots\,,P(Z_1^r,E_1^r{=}\mathbf{a}_{2^r})\Big],$$

and $\mathbf{M}_\varepsilon(Z_i^{i+r-1},Z_{i-1}^{i+r-2})$ is a $2^r\times 2^r$ matrix defined as follows: for each binary $(r+1)$-tuple a_1^{r+1}, we have

$$\big(\mathbf{M}_\varepsilon(Z_i^{i+r-1},Z_{i-1}^{i+r-2})\big)_{a_1^r,a_2^{r+1}}=P_X(Z_i^{i+r-1}\oplus a_2^{r+1}\,|\,Z_{i-1}^{i+r-2}\oplus a_1^r)P(E_i^{i+r-1}{=}a_2^{r+1}),\;\; i>1. \tag{54}$$

All other entries of the matrix are zero. Clearly, $\mathbf{M}_\varepsilon(Z_i^{i+r-1},Z_{i-1}^{i+r-2})$ is a random matrix, drawn from a set of 2^{r+1} possible realizations. We now proceed as in the case $r=1$, and, for a realization z_1^n of Z_1^n, we write

$$\mathbf{M}_i=\mathbf{M}_\varepsilon(z_i^{i+r-1},z_{i-1}^{i+r-2})=\mathbf{M}_i^{(0)}+\varepsilon\mathbf{M}_i^{(1)},$$

and $\mathbf{p}_1=\mathbf{p}_1^{(0)}+\varepsilon\mathbf{p}_1^{(0)}$. For example, for $r=2$, we have

$$\mathbf{M}_{n-1}^{(0)}=\left[\begin{array}{cccc} P_X(z_n,z_{n+1}|z_{n-1},z_n) & 0 & 0 & 0\\ 0 & 0 & P_X(\bar z_n,z_{n+1}|z_{n-1},\bar z_n) & 0\\ P_X(z_n,z_{n+1}|\bar z_{n-1},z_n) & 0 & 0 & 0\\ 0 & 0 & P_X(\bar z_n,z_{n+1}|\bar z_{n-1},\bar z_n) &)\end{array}\right]$$

and

$$
\mathbf{M}_{n-1}^{(1)} = \begin{bmatrix} -P_X(z_n, z_{n+1}|z_{n-1}, z_n) & P_X(z_n, \bar{z}_{n+1}|z_{n-1}, z_n) & 0 & 0 \\ 0 & 0 & -P_X(\bar{z}_n, z_{n+1}|z_{n-1}, \bar{z}_n) & P_X(\bar{z}_n, \bar{z}_{n+1}|z_{n-1}, \bar{z}_n) \\ -P_X(z_n, z_{n+1}|\bar{z}_{n-1}, z_n) & P_X(z_n, \bar{z}_{n+1}|\bar{z}_{n-1}, z_n) & 0 & 0 \\ 0 & 0 & -P_X(\bar{z}_n, z_{n+1}|\bar{z}_{n-1}, \bar{z}_n) & P_X(\bar{z}_n, \bar{z}_{n+1}|\bar{z}_{n-1}, \bar{z}_n) \end{bmatrix}.
$$

Using the above definitions and (53), we arrive at

$$
\frac{\partial}{\partial \varepsilon} P_Z(z_1^n)\bigg|_{\varepsilon=0} = \sum_{i=1}^{n-3} \mathbf{p}_1^{(0)} \mathbf{M}_1^{(0)} \cdots \mathbf{M}_{i-1}^{(0)} \mathbf{M}_i^{(1)} \mathbf{M}_{i+1}^{(0)} \cdots \mathbf{M}_{n-1}^{(0)} \mathbf{1}^t + O(1),
$$

where the $O(1)$ term contains the boundary cases. This is an analogue of (40). After some further manipulations, we obtain, in analogy to (42),

$$
\begin{aligned}
\frac{\partial}{\partial \varepsilon} P_Z(z_1^n)\bigg|_{\varepsilon=0} &= -P_X(z_1^n) + \\
& \sum_{i=2}^{n-r} P_X(z_1^{i-1}) \cdot P_X(z_{i-r+1} \bar{z}_i z_{i+1}^{i+r-1} | z_{i-r}^{i-1}) \cdot P_X(z_{i+1}^n | \bar{z}_i z_{i+1}^{i+r-1}) \; + \; O(1).
\end{aligned}
$$

The matrix $\mathbf{P}(s)$ for an rth order process is defined by

$$
(\mathbf{P}(s))_{(a_1,\ldots a_i \ldots a_r),(a_2,\ldots a_i,\ldots a_{r+1})} = p^s_{(a_1,\ldots \bar{a}_i \ldots a_r),(a_2,\ldots \bar{a}_i,\ldots a_{r+1})}, \quad a_1^{r+1} \in \{0,1\}^r,
$$

with zeroes in the remaining locations, where

$$
p_{(a_1,\ldots a_i \ldots a_r),(a_2,\ldots a_i,\ldots a_{r+1})} = P_X(a_2,\ldots a_i,\ldots a_{r+1}|a_1,\ldots a_i \ldots a_r)
$$

are the transition probabilities of the Markov chain. We also define

$$
\mathbf{Q}_i(s) = \mathbf{P}(1) \circ \widetilde{\mathbf{Q}}_i(s)
$$

where $\circ$ denotes the Schur (element-wise) product of matrices, and the $2^r \times 2^r$ matrix $\widetilde{\mathbf{Q}}_i(s)$ is defined by

$$
\left(\widetilde{\mathbf{Q}}_i\right)_{(a_1,\ldots a_i \ldots a_r),(a_2,\ldots a_i,\ldots a_{r+1})} = p^{s-1}_{(a_1,\ldots \bar{a}_i \ldots a_r),(a_2,\ldots \bar{a}_i,\ldots a_{r+1})}, \quad a_1^{r+1} \in \{0,1\}^r,
$$

with zeroes in the remaining locations (notice that this generalizes the definitions in (15)). With these definitions, using (55), we compute the derivative of $R(s,\varepsilon)$ at $\varepsilon = 0$. Thus,

$$
\frac{\partial}{\partial \varepsilon} R_n(s,\varepsilon)\bigg|_{\varepsilon=0} = s\boldsymbol{\pi}(s) \sum_{i=r}^{n-r-2} \mathbf{P}^{i-r}(s) \left[\mathbf{Q}_1(s) \cdots \mathbf{Q}_{r+1}(s) - \mathbf{P}^{r+1}(s)\right] \mathbf{P}^{n-i-r-1}(s) \mathbf{1}^t + O(1). \tag{55}
$$

To find the linear term in the Taylor expansion for the entropy rate, we need to differentiate (55) with respect to s, and evaluate it at $s = 1$. To facilitate this computation, we apply, as for the $r=1$ case, spectral matrix representations. Defining $\mathbf{Q}(s) = \mathbf{Q}_1(s)\cdots\mathbf{Q}_{r+1}(s)$, we obtain

$$\frac{\partial^2}{\partial\varepsilon\partial s}R_n(s,\varepsilon)\Big|_{\substack{\varepsilon=0,\\ s=1}} = n\,\boldsymbol{\pi}(1)\left(\mathbf{r}_1^t(1)\,\boldsymbol{\ell}_1(1)\right)\frac{\partial}{\partial s}\left(\mathbf{Q}(s)-\mathbf{P}^{r+1}(s)\right)\Big|_{s=1}\left(\mathbf{r}_1^t(1)\,\boldsymbol{\ell}_1(1)\right)\mathbf{1}^t\ + o(n),$$

which is derived using the relation $\mathbf{Q}(1) = \mathbf{P}^{r+1}(1)$. Theorem 3 now follows by observing that, as before, we have $\boldsymbol{\pi}(1) = \boldsymbol{\ell}_1(1)$, $\mathbf{r}_1(1) = \mathbf{1}$, and $\boldsymbol{\ell}_1(1)\mathbf{r}_1(1) = 1$, writing explicit expressions for the derivatives of $\mathbf{Q}(s)$ and $\mathbf{P}^{r+1}(s)$ at $s=1$ (in analogy to (22)–(23)), and carrying out the ensuing symbolic computations. ■

3.4 Proof of Theorem 4

In this section we derive the Taylor expansion for the Rényi entropy of order s, establishing Theorem 4. Taking the Taylor expansion of $\log R_n(s,\varepsilon)$ around $\varepsilon=0$ we arrive at

$$(1-s)H_s(Z_1^n) = \log R_n(s,\varepsilon) = \log R_n(s,0) + \varepsilon\frac{R'_\varepsilon(s,0)}{R_n(s,0)} + O(n\varepsilon^2),$$

where the error term follows from Lemma 2. From (13) and (19) (or Lemma 3 above) we conclude that

$$R_n(s,0) = \boldsymbol{\pi}(s)\mathbf{P}^{n-1}(s)\mathbf{1}^t = \lambda^{n-1}(s)\langle\boldsymbol{\pi}(s),\mathbf{r}_1(s)\rangle\langle(\boldsymbol{\ell}_1(s),\mathbf{1}\rangle(1+O(\rho^n)),$$

for $\rho<1$, where $\lambda(s)$ is the main eigenvalue of $\mathbf{P}(s)$ and $\boldsymbol{\ell}_1(s)$ and $\mathbf{r}_1(s)$ are the main left and right eigenvectors.

In a similar fashion we can express $R'_\varepsilon(s,0)$ given by (16) of Lemma 1. Indeed,

$$\begin{aligned}
R'_\varepsilon(s,0) &= s\boldsymbol{\pi}(s)\sum_{i=1}^{n-1}\mathbf{P}^{i-1}(s)\left(\mathbf{Q}_1(s)\mathbf{Q}_2(s)-\mathbf{P}^2(s)\right)\mathbf{P}^{n-i-2}(s)\mathbf{1}^t \\
&= s\sum_{i=1}^{n-1}\lambda^{n-3}\langle(\boldsymbol{\pi}(s),\mathbf{r}_1(s)\rangle\langle\boldsymbol{\ell}_1(s),\mathbf{Q}(s)-\mathbf{P}^2(s),\mathbf{r}_1(s)\rangle\langle\boldsymbol{\ell}_1(s),\mathbf{1}\rangle(1+O(\rho^n)) \\
&= s(n-1)\lambda^{n-3}\langle(\boldsymbol{\pi}(s),\mathbf{r}_1(s)\rangle\langle\boldsymbol{\ell}_1(s),\mathbf{Q}(s)-\mathbf{P}^2(s),\mathbf{r}_1(s)\rangle\langle\boldsymbol{\ell}_1(s),\mathbf{1}\rangle
\end{aligned}$$

Thus

$$\begin{aligned}
\log R_n(s,0) &= (n-1)\log\lambda(s) + O(1) \\
\frac{R'_\varepsilon(s,0)}{R_n(s,0)} &= \frac{s(n-1)\langle\boldsymbol{\ell}_1(s),\mathbf{Q}(s)-\mathbf{P}^2(s),\mathbf{r}_1(s)\rangle}{\lambda^2(s)}(1+O(\rho^n))
\end{aligned}$$

Putting everything together, we finally establish Theorem 4. ■

4 Conclusions

We studied the entropy rate of a hidden Markov process (HMP) defined as an output of a binary symmetric channel whose input is a binary Markov process. We first established that the entropy rate of the HMP coincides with the value of the top Lyapunov exponent of a well defined product of random matrices. These exponents are notoriously difficult to compute. Therefore, we turned our attention to asymptotic expansions, and derived a Taylor expansion of the HMP entropy rate when the probability of error is small. We observed that the linear term of the expansion is the Kullback-Liebler divergence between distributions of triplets of symbols, which are determined from marginals of the underlying Markov process. We also determined the second order term of the expansion explicitly, and validated the accuracy of the Taylor approximation with empirical simulation results. We showed extensions of our results to HMPs with underlying Markov processes of arbitrary order, and to the computation of HMP Rényi entropies of any order.

References

[1] 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.

[2] 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.

[3] 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.

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

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

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

[7] 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.

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

[9] J. Fan, T. L. Poo, and B. Marcus, "Constraint Gain," *IEEE Trans. Information Theory*, **50**, 1989–1999, 2001.

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

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

[12] R. Gharavi and V. Anantharam, An Upper Bound for the Largest Lyapunov Exponent of a Markovian Product of Nonnegative Matrices, preprint, 2004.

[13] T. Holliday, A. Goldsmith, and P. Glynn, "Capacity of Finite State Channels Based on Lyapunov Exponents of Random Matrices," *IEEE Trans. Information Theory*, **52**, 3509- 3532, 2006.

[14] G. Han and B. Marcus, "Analyticity of Entropy Rate of Hidden Markov Chains," *IEEE Trans. Information Theory*, **52**, 5251–5266, 2006.

[15] G. Han and B. Marcus, "Analyticity of Entropy Rate in Families of Hidden Markov Chains (II)," *Proc. ISIT 2006*, 103–107, Seattle, 2006.

[16] G. Han and B. Marcus, "Capacity of noisy constrained channels," preprint, Dec. 2006.

[17] 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.

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

[19] P. Jacquet, G. Seroussi, and W. Szpankowski, "On the Entropy of a Hidden Markov Process" (extended abstract), *Data Compression Conference*, 362–371, Snowbird, 2004.

[20] P. Jacquet, G. Seroussi, and W. Szpankowski, "Noisy constrained capacity", preprint, December 2006.

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

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

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

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

[25] B. Marcus, R. Roth and P. Siegel, "Constrained Systems and Coding for Recording Channels," Chap. 20 in *Handbook of Coding Theory* (eds. V. S. Pless and W. C. Huffman), Elsevier Science, 1998.

[26] B. Noble, and J. Daniel, *Applied Linear Algebra*, Prentice Hall, Englewood Cliffs, NJ, 1988.

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

[28] E. Ordentlich and T. Weissman, "New Bounds on th Entropy Rate of Hidden Markov Process," *Information Theory Workshop*, 117–122, San Antonio, 2004.

[29] E. Ordentlich and T. Weissman, "On the optimality of symbol by symbol filtering and denoising," *IEEE Trans. Information Theory*, **52**, 19-40, 2006.

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

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

[32] 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.

[33] A. Rényi, "On measures of entropy and information," Proceedings of 4th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, California, pp. 547–561, 1960.

[34] E. Seneta, *Non-Negative Matrices.* New York: John Wiley & Sons, 1973.

[35] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, 22, 1176-1198, 1993.

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

[37] 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.

[38] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Information Theory* , **52**, 5–28, 2005.

[39] O. Zuk, I. Kanter and E. Domany, "Asymptotics of the Entropy Rate for a Hidden Markov Process," *J. Stat. Phys.*, **121**, 343–360, 2005.