Taylor series expansions for the entropy rate of Hidden Markov Processes

Or Zuk Eytan Domany
Dept. of Physics of Complex Systems
Weizmann Inst. of Science
Rehovot, 76100, Israel
Email: {or.zuk}/{eytan.domany}@weizmann.ac.il

Ido Kanter
Faculty of Physics
Bar-Ilan Univ.
Ramat-Gan, 52900, Israel
Email: kanter@mail.biu.ac.il

Michael Aizenman
Deptartment of Physics
Princeton Univ.
Princeton, NJ 08544-0708
Email: aizenman@princeton.edu

***Abstract*— Finding the entropy rate of Hidden Markov Processes is an active research topic, of both theoretical and practical importance. A recently used approach is studying the asymptotic behavior of the entropy rate in various regimes. In this paper we generalize and prove a previous conjecture relating the entropy rate to entropies of finite systems. Building on our new theorems, we establish series expansions for the entropy rate in two different regimes. We also study the radius of convergence of the two series expansions.**

I. Introduction

Let $\{X_N\}$ be a finite state stationary markov process over the alphabet $\Sigma = \{1, \dots, s\}$. Let $\{Y_N\}$ be its noisy observation (on the same alphabet). Let $M = M_{s\times s} = \{m_{ij}\}$ be the Markov transition matrix and $R = R_{s\times s}$ be the emission matrix, i.e. $P(X_{N+1} = j|X_N = i) = m_{ij}$ and $P(Y_N = j|X_N = i) = r_{ij}$. We assume that the Markov matrix M is strictly positive ($m_{ij} > 0$), and denote its stationary distribution by the (column) vector π, satisfying $\pi^t M = \pi^t$.

The process Y can be viewed as a noisy observation of X, through a noisy channel. It is known as a *Hidden Markov Process (HMP)*, and is determined by the parameters M and R. *HMPs* have a rich and developed theory, and enourmous applications in various fields (see [1], [2]).

An important quantity of the process Y is its entropy rate. The Shannon entropy rate of a stochastic process ([3]) measures the amount of 'uncertainty per-symbol'. More formally, for $i \le j$, let $[Y]_i^j$ denote the vector $(Y_i, \dots, Y_j)$. Then the entropy rate $\bar{H}(Y)$ is defined as:

$$\bar{H}(Y) = \lim_{N\to\infty} \frac{H([Y]_1^N)}{N} \tag{1}$$

Where $H(Y) = -\sum_Y P(Y)\log P(Y)$. Here and throughout the paper we use natural logarithms, so the entropy is measured in *NATS*, and also adopt the convention $0\log 0 \equiv 0$. We sometimes omit the realization y of the variable Y, so $P(Y)$ should be understood as $P(Y = y)$. The entropy rate can also be computed via the conditional entropy as: $\bar{H}(Y) = \lim_{N\to\infty} H(Y_N|[Y]_1^{N-1})$, since for a stationary process the two limits exist and coincide ([4]). The conditional entropy $H(Y|X)$ (where X, Y are sets of r.v.s.) represents the average uncertainty of Y, assuming that we know X, that is $H(Y|X) = \sum_x P(X = x)H(Y|X = x)$. By the chain rule for entropy, it can also be viewed as a difference of entropies, $H(Y|X) = H(X, Y) - H(X)$, which will be used later.

There is at present no explicit expression for the entropy rate of a *HMP* ([1], [5]). Few recent works ([5], [6], [7]) have dealt with finding the asymptotic behavior of $\bar{H}$ in several parameter regimes. However, they concentrated only on binary alphabet, and proved rigorously only bounds or at most second ([7]) order behavior.

Here we generalize and prove a conjecture posed in [7], which justifies (under some mild assumptions) the computation of $\bar{H}$ as a series expansion in the High Signal-to-Noise-Ratio ('High-SNR') regime. The expansion coefficients were given in [7], for the symmetric binary case. In this case, the matrices M and R are given by:

$$M = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix} \ , \ R = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix} \tag{2}$$

and the process is characterized by the two parameters p, ϵ. The High-SNR expansion in this case is an expansion in ϵ around zero.

In section II, we present and prove our two main theorems; Thm. 1 is a generalization of a conjecture raised in [7] which connects the coefficients of entropies using finite histories to the entropy rate. Proving it justifies the High-SNR expansion of [7]. We also give Thm. 2, which is the analogue of Thm. 1

in a different regime, termed 'Almost-Memoryless' ('A-M'). In section III we use our two new theorems to compute the first coefficients in the series expansions for the two regimes. We give the first-order asymptotics for a general alphabet, as well as higher order coefficients for the symmetric binary case. In section IV we estimate the radius of convergence of our expansions using a finite number of terms, and compare our results for the two regimes. We end with conclusions and future directions.

II. From Finite system entropy to entropy rate

In this section we prove our main results, namely Thms. 1 and 2, which relate the coefficients of the finite bounds C_N to those of the entropy rate $\bar{H}$ in two different regimes.

A. *The High SNR Regime*

This regime was dealt in further details in [7], [8], albeit with no rigorous justification for the obtained series expansion. In the High-SNR regime the observations are likely to be equal to the states, or in other words, the emission matrix R is close to the identity matrix I. We therefore write $R = I + \epsilon T$, where $\epsilon > 0$ is a small constant and $T = \{t_{ij}\}$ is a matrix satisfying $t_{ii} < 0, t_{ij} \geq 0, \forall i \neq j$ and $\sum_{j=1}^{s} t_{ij} = 0$. The entropy rate in this regime can be given as an expansion in ϵ around zero. We state here our new theorem, connecting the entropy of finite systems to the entropy rate in this regime.

Theorem 1: Let $H_N \equiv H_N(M,T,\epsilon) = H([Y]_1^N)$ be the entropy of a system of length N, and let $C_N = H_N - H_{N-1}$. Let $B_\rho(0) \subset \mathbb{C}$ be some (complex) neighborhood of zero, in which the functions $\{C_N\}$ and $\bar{H}$ are analytic in ϵ, with Taylor expansions given by:

$$C_N(M,T,\epsilon) = \sum_{k=0}^{\infty} C_N^{(k)} \epsilon^k, \quad \bar{H}(M,T,\epsilon) = \sum_{k=0}^{\infty} C^{(k)} \epsilon^k \quad (3)$$

(The coefficients $C_N^{(k)}$ are functions of the parameters M and T. From now on we omit this dependence). Then:

$$N \geq \left\lceil \frac{k+3}{2} \right\rceil \Rightarrow C_N^{(k)} = C^{(k)} \quad (4)$$

The analyticity of $\{C_N\}$ and $\bar{H}$ around $\epsilon = 0$ was recently shown in [9]. One can also use [10], which showed that the law of the process Y is Gibbsian, together with the complete analyticity results for Gibbsian measures in [11] to deduce analyticity of $\bar{H}$.

C_N is actually an upperbound ([4]) for $\bar{H}$. The behavior stated in Thm. 1 was discovered previously using symbolic computations, but was proven only for $k \leq 2$, and only for the symmetric binary case (see [7]).

Although it may appear technically involved, the proof of Thm. 1 is based on the following two simple ideas. First, we distinguish between the noise parameters at different sites. This is done by considering a more general process $\{Z_N\}$, where Z_i's emission matrix is $R_i = I + \epsilon_i T$. The joint distribution of $[Z]_1^N$ is thus determined by M,T and $[\epsilon]_1^N$. We define the following functions:

$$F_N(M,T,[\epsilon]_1^N) = H([Z]_1^N) - H([Z]_1^{N-1}) \quad (5)$$

Setting all the ϵ_i's equal reduces us back to the Y process, and in particular $F_N(M,T,(\epsilon,\dots,\epsilon)) = C_N(\epsilon)$.

Second, we observe that if a particular ϵ_i is set to zero, the corresponding observation Z_i must equal the state X_i. Thus, conditioning back to the past is 'blocked'. This can be used to prove the following:

Lemma 1: Assume $\epsilon_j = 0$ for some $1 < j < N$. Then:

$$F_N([\epsilon]_1^N) = F_{N-j+1}([\epsilon]_{j+1}^N)$$

Proof:

F can be written as a sum of conditional entropies:

$$F_N = -\sum_{[Z]_1^N} P([Z]_1^{N-1}) P(Z_N|[Z]_1^{N-1}) \log P(Z_N|[Z]_1^{N-1}) \quad (6)$$

Where the dependence on $[\epsilon]_1^N$ and M,T comes through the probabilities $P(..)$. Since $\epsilon_j = 0$, we must have $X_j = Z_j$, and therefore (since the X_i's form a Markov chain), conditioning further to the past is 'blocked', that is:

$$\epsilon_j = 0 \Rightarrow P(Z_N|[Z]_1^{N-1}) = P(Z_N|[Z]_j^{N-1}) \quad (7)$$

(Note that eq. (7) is true for $j < N$, but not for $j = N$). Substituting in eq. (6) gives:

$$\begin{aligned} F_N = &-\sum_{[Z]_1^N} P([Z]_1^{N-1}) P(Z_N|[Z]_j^{N-1}) \log P(Z_N|[Z]_j^{N-1}) = \\ &-\sum_{Z_j^N} P([Z]_j^{N-1}) P(Z_N|[Z]_j^{N-1}) \log P(Z_N|[Z]_j^{N-1}) \\ &= F_{N-j+1} \quad (8) \end{aligned}$$

$\blacksquare$

Let $\vec{k} = [k]_1^N$ be a vector with $k_i \in \{\mathbb{N} \cup 0\}$. Define its 'weight' as $\omega(\vec{k}) = \sum_{i=1}^{N} k_i$. Define also:

$$F_N^{\vec{k}} \equiv \left. \frac{\partial^{\omega(\vec{k})} F_N}{\partial \epsilon_1^{k_1}, \dots, \partial \epsilon_N^{k_N}} \right|_{\vec{\epsilon}=0} \quad (9)$$

With the above definition, $C_N^{(k)}$ is obtained by summing $F_N^{\vec{k}}$ on all $\vec{k}$'s with weight k, and dividing by $k!$:

$$C_N^{(k)} = \frac{1}{k!} \sum_{\vec{k}, \omega(\vec{k})=k} F_N^{\vec{k}} \tag{10}$$

As is shown next, one does not need to sum on all such $\vec{k}$'s, since many of them give zero contribution:

Lemma 2: Let $\vec{k} = [k]_1^N$. If $\exists i, j,\ 1 \leq i < j < N$, with $k_j \leq 1 \leq k_i$, then $F_N^{\vec{k}} = 0$.

Proof: Assume first $k_j = 0$. Using lemma 1 we get

$$F_N^{\vec{k}} \equiv \left. \frac{\partial^{\omega(\vec{k})} F_N([\epsilon]_1^N)}{\partial\epsilon_1^{k_1}, \ldots, \partial\epsilon_N^{k_N}} \right|_{\vec{\epsilon}=0} = \left. \frac{\partial^{\omega(\vec{k})} F_{N-j+1}([\epsilon]_j^N)}{\partial\epsilon_1^{k_1}, \ldots, \partial\epsilon_N^{k_N}} \right|_{\vec{\epsilon}=0} =$$
$$\left. \frac{\partial^{\omega(\vec{k})-1}}{\partial\epsilon_1^{k_1}, \ldots, \partial\epsilon_i^{k_i-1}, \ldots, \partial\epsilon_N^{k_N}} \left[\frac{\partial F_{N-j+1}([\epsilon]_j^N)}{\partial\epsilon_i} \right] \right|_{\vec{\epsilon}=0} = 0 \tag{11}$$

The case $k_j = 1$ is more difficult, but follows the same principles. Write the probability of Z:

$$P([Z]_1^N) = \sum_{[X]_1^N} P([X]_1^N) P([Z]_1^N | [X]_1^N) =$$
$$\sum_{[X]_1^N} P([X]_1^N) \prod_{i=1}^{N} (\delta_{X_i Z_i} + \epsilon_i t_{X_i Z_i}) \tag{12}$$

where δ_{ij} is Kronecker delta. Write the partial derivative with respect to ϵ_j:

$$\left. \frac{\partial P([Z]_1^N)}{\partial\epsilon_j} \right|_{\epsilon_j=0} =$$
$$\left. \sum_{[X]_1^N} \left[P([X]_1^N) t_{X_j Z_j} \prod_{i \neq j} (\delta_{X_i Z_i} + \epsilon_i t_{X_i Z_i}) \right] \right|_{\epsilon_j=0} =$$
$$\left. \left\{ \sum_{a=1}^{s} t_{a Z_j} P([Z]_1^{N\,(j \to a)}) \right\} \right|_{\epsilon_j=0} \tag{13}$$

Where $[Z]_1^{N\,(j \to a)}$ denotes the vector which is equal to $[Z]_1^N$ in all coordinates except on coordinate j, where $Z_j = a$. Using Bayes' rule $P(Z_N | [Z]_1^{N-1}) = \frac{P([Z]_1^N)}{P([Z]_1^{N-1})}$, we get:

$$\left. \frac{\partial P(Z_N | [Z]_1^{N-1})}{\partial\epsilon_j} \right|_{\epsilon_j=0} =$$
$$\frac{1}{P([Z]_1^{N-1})} \sum_{a=1}^{s} t_{a Z_j} \Big[P([Z]_1^{N\,(j \to a)}) -$$
$$\left. P(Z_N | [Z]_1^{N-1}) P([Z]_1^{N-1\,(j \to a)}) \Big] \right|_{\epsilon_j=0} \tag{14}$$

This gives:

$$\left. \frac{\partial [P([Z]_1^N) \log P(Z_N | [Z]_1^{N-1})]}{\partial\epsilon_j} \right|_{\epsilon_j=0} =$$
$$\sum_{a=1}^{s} t_{a Z_j} \Big\{ P([Z]_1^{N\,(j \to a)}) \log P(Z_N | [Z]_1^{N-1}) +$$
$$\left. P([Z]_1^{N\,(j \to a)}) - P(Z_N | [Z]_1^{N-1}) P([Z]_1^{N-1\,(j \to a)}) \Big\} \right|_{\epsilon_j=0} \tag{15}$$

And therefore:

$$\left. \frac{\partial F_N}{\partial\epsilon_j} \right|_{\epsilon_j=0} =$$
$$-\sum_{a=1}^{s} t_{a Z_j} \Bigg\{ \sum_{[Z]_1^N} \Big[P([Z]_1^{N\,(j \to a)}) \log P(Z_N | [Z]_1^{N-1}) -$$
$$\left. P(Z_N | [Z]_1^{N-1}) P([Z]_1^{N-1\,(j \to a)}) \Big] \Bigg\} \right|_{\epsilon_j=0} =$$
$$\Bigg\{ -\sum_{a=1}^{s} t_{a Z_j} \sum_{[Z]_j^N} \Big[P([Z]_j^{N\,(1 \to a)}) \log P(Z_N | [Z]_j^{N-1}) -$$
$$\left. P(Z_N | [Z]_j^{N-1}) P([Z]_j^{N-1\,(1 \to a)}) \Big] \Bigg\} \right|_{\epsilon_1=0} \tag{16}$$

Where the latter equality comes from using eq. (7), which 'blocks' the dependence backwards. Eq. (16) shows that $\left. \frac{\partial F_N}{\partial\epsilon_j} \right|_{\epsilon_j=0}$ does not depend on ϵ_i for $i < j$, therefore $\frac{\partial^{k_i+1} F_N}{\partial\epsilon_i^{k_i} \partial\epsilon_j} = 0$ and $F_N^{\vec{k}} = 0$. ■

Before proving Thm. 1, we need one more lemma, which already shows a 'settling' behavior. More precisely, we prove here that adding zeros to the left of $\vec{k}$ leaves $F_N^{\vec{k}}$ unchanged:

Lemma 3: Let $\vec{k} = [k]_1^N$ with $k_1 \leq 1$. Denote $\vec{k}^{(r)}$ the concatenation of $\vec{k}$ with r zeros on the left: $\vec{k}^{(r)} = (\underbrace{0, \ldots, 0}_{r}, k_1, \ldots, k_N)$. Then:

$$F_N^{\vec{k}} = F_{r+N}^{\vec{k}^{(r)}} \quad, \forall r \in \mathbb{N}$$

Proof: Assume first $k_1 = 0$. Using lemma 1, we get:

$$F_{r+N}^{\vec{k}^{(r)}}([\epsilon]_1^{r+N}) = \left. \frac{\partial^{\omega(\vec{k}^{(r)})} F_{r+N}([\epsilon]_1^{r+N})}{\partial\epsilon_{r+2}^{k_2}, \ldots, \partial\epsilon_{r+N}^{k_N}} \right|_{\vec{\epsilon}=0} =$$
$$\left. \frac{\partial^{\omega(\vec{k})} F_N([\epsilon]_{r+1}^{r+N})}{\partial\epsilon_{r+2}^{k_2}, \ldots, \partial\epsilon_{r+N}^{k_N}} \right|_{\vec{\epsilon}=0} = F_N^{\vec{k}}([\epsilon]_{r+1}^{r+N}) \tag{17}$$

The case $k_1 = 1$ is reduced back to the case $k_1 = 0$ by taking the derivative. We next prove the claim for $r = 1$ and for greater values it follows by induction. Using eqs. (16,17), we get:

$$
\begin{gathered}
F_{N+1}^{\vec{k}^{(1)}}([\epsilon]_1^{N+1}) = \frac{\partial^{\omega(\vec{k})-1}}{\partial\epsilon_3^{k_2}\ldots\partial\epsilon_{N+1}^{k_N}}\left[\left.\frac{\partial F_{N+1}}{\partial\epsilon_2}\right|_{\epsilon_2=0}\right]\Bigg|_{\vec{\epsilon}=0} = \\
\frac{\partial^{\omega(\vec{k})-1}}{\partial\epsilon_3^{k_2}\ldots\partial\epsilon_{N+1}^{k_N}}\Bigg\{-\sum_{a=1}^{s} t_{aZ_2}\sum_{[Z]_1^{N+1}} \\
\Big[P([Z]_1^{N+1\,(2\to a)})\log P(Z_{N+1}|[Z]_1^N)- \\
P(Z_{N+1}|[Z]_1^N)P([Z]_1^{N\,(2\to a)})\Big]\Big|_{\epsilon_2=0}\Bigg\}\Bigg|_{[\epsilon]_1^{N+1}=0} = \\
\frac{\partial^{\omega(\vec{k})-1}}{\partial\epsilon_2^{k_2}\ldots\partial\epsilon_{N}^{k_N}}\Bigg\{-\sum_{a=1}^{s} t_{aZ_2}\sum_{[Z]_1^{N}} \\
\Big[P([Z]_1^{N\,(1\to a)})\log P(Z_{N}|[Z]_1^{N-1})- \\
P(Z_{N}|[Z]_1^{N-1})P([Z]_1^{N\,(1\to a)})\Big]\Big|_{\epsilon_1=0}\Bigg\}\Bigg|_{[\epsilon]_1^{N}=0} = F_N^{\vec{k}}([\epsilon]_1^N)
\end{gathered}
\tag{18}
$$

■

We are now ready to prove Thm. 1, which follows directly from lemmas 2 and 3:

Proof:
Let $\vec{k} = [k]_1^N$ with $\omega(\vec{k}) = k$. Define its 'length' (from right, considering only entries *strictly* larger than one) as $l(\vec{k}) = N + 1 - \min_{k_i>1}\{i\}$. It easily follows from lemma 2 that if $F_N^{\vec{k}} \neq 0$, then all the entries of $\vec{k}$ except some of its leftmost entries are at least '2', and thus we must have $l(\vec{k}) \leq \lceil\frac{k+3}{2}\rceil - 1$. Therefore, according to lemma 3 we have:

$$
F_N^{\vec{k}} = F_{\lceil\frac{k+3}{2}\rceil}^{(k_{N-\lceil\frac{k+3}{2}\rceil+1},\ldots,k_N)}
\tag{19}
$$

for all $\vec{k}$'s in the sum. From eq. 10, by summing over all $F_N^{\vec{k}}$ with the same 'weight', we get $C_N^{(k)} = C_{\lceil\frac{k+3}{2}\rceil}^{(k)}, \quad \forall N > \lceil\frac{k+3}{2}\rceil$. From the analyticity of C_N and $\bar{H}$ around $\epsilon = 0$, one can show by induction on k that $\lim_{N\to\infty} C_N^{(k)} = C^{(k)}$, therefore we must have $C_N^{(k)} = C^{(k)}, \quad \forall N \geq \lceil\frac{k+3}{2}\rceil$. ■

B. The Almost Memoryless Regime

In the A-M regime, the Markov transition matrix is close to uniform. Thus, throughout this section, we assume that M is given by $M = U + \delta T$, such that U is a constant (uniform) matrix, $u_{ij} = s^{-1}$, $\delta > 0$ is a small constant and T satisfies $\sum_{j=1}^{s} t_{ij} = 0$. Thus the process is entirely characterized by the set of parameters (R, T, δ), where R again denotes the emission matrix.

Interestingly, similarly to the High-SNR regime, the conditional entropy given a finite history gives the correct entropy rate up to a certain order which depends on the finite history taken. In the A-M regime we can also prove analyticity of $\{C_N\}$ and $\bar{H}$ in δ near $\delta = 0$. This is stated as:

Theorem 2: Let $H_N \equiv H_N(R,T,\delta) = H([Y]_1^N)$ be the entropy of a finite system of length N, and let $C_N = H_N - H_{N-1}$. Let $B_\rho(0) \subset \mathbb{C}$ be some (complex) neighborhood of $\delta = 0$, in which the (one-variable) functions $\{C_N\}, \bar{H}$ are analytic in δ, with Taylor expansions given by:

$$
C_N(M,T,\delta) = \sum_{k=0}^{\infty} C_N^{(k)}\delta^k, \quad \bar{H}(M,T,\epsilon) = \sum_{k=0}^{\infty} C^{(k)}\delta^k
\tag{20}
$$

(The coefficients $C_N^{(k)}$ are functions of the parameters M and T.) Then:

$$
N \geq \left\lceil\frac{k+3}{2}\right\rceil \Rightarrow C_N^{(k)} = C^{(k)}
\tag{21}
$$

Proof: The proof is very similar to that of Thm. 1. Distinguishing between the sites by setting $M_i = U + \delta_i T$ in site i, we notice that if one sets $\delta_i = 0$ for some i, then M_i becomes uniform, and thus knowing Z_i 'blocks' the dependence of Z_N on previous Z_j's ($\forall j < i$). The rest of the proof continues in an analogous way to the proof of Thm. 1 (including the three lemmas therein), and its details are thus omitted here. ■

III. Computation of the series coefficients

An immediate application of Thms. 1 and 2 is the computation of the first terms in the series expansion for $\bar{H}$ (assuming its existence), by simply computing these terms for C_N for N large enough. In this section we compute, for both regimes, the first order for the general alphabet case, and also give few higher order terms for the simple symmetric binary case. Our method for computing $C^{(k)}$ is straightforward. We compute $C_N^{(k)}$ for $N = \lceil\frac{k+3}{2}\rceil$ by simply enumerating all sequences $[Y]_1^N$, computing the k-th coefficient in $P([Y]_1^N)\log P([Y]_1^N)$ for each one, and summing their contribution. This computation is, however, exponential in k, and thus raises the challenge of designing more efficient algorithms, in order to compute further orders and for larger alphabets.

Before giving the calculated coefficients, we need some new notations. For a vector α, $diag(\alpha)$ denotes the square matrix with α's elements on the diagonal. We use Matlab-like notation to denote element-by-element operations on matrices. Thus, for matrices A and B, $logA$ is a matrix whose elements are $\{\log a_{ij}\}$, and $[A. * B]$ is a matrix whose elements are $\{a_{ij}b_{ij}\}$. ξ denotes the (column) vector of N ones.

A. The High-SNR expansion

According to Thm. 1, computing C_2 enables us to extract $\bar{H}^{(k)}$. This is used to show the following:

Proposition 1: Let $R = I + \epsilon T$. Assume that the entropy rate $\bar{H}$ is analytic in some neighborhood of $\epsilon = 0$. Then $\bar{H}$ satisfies:

$$\bar{H} = -\pi^t [M. * \log M] \xi + \xi^t \Big\{ diag(\log(\pi)) T^t diag(\pi) M -$$

$$[diag(\pi) M T + T^t diag(\pi) M]. * [\log(diag(\pi) M)] \Big\} \xi \epsilon + O(\epsilon^2) \tag{22}$$

Proof: Noting that according to Thm. 1, $\bar{H} = C_2 + O(\epsilon^2)$, we first compute (exactly) C_2, and then expand it by substituting $R = I + \epsilon T$. Write C_2 as:

$$C_2 = H(Y_N | Y_{N-1}) =$$

$$-\sum_{i,j} P(Y_N = j, Y_{N-1} = i) \log \frac{P(Y_N = j, Y_{N-1} = i)}{P(Y_{N-1} = i)} \tag{23}$$

We can express the above probabilities as:

$$P(Y_{N-1} = i) = [\pi^t R]_i$$

$$P(Y_N = j, Y_{N-1} = i) = [R^t diag(\pi) M R]_{ij} \equiv F_{ij} \tag{24}$$

Substituting eq. (24) in eq. (23), and writing in matrix form, we get:

$$C_2 = \Big\{ [log(\pi^t R)] F - \xi^t [F. * log F] \Big\} \xi \tag{25}$$

Substituting $R = I + \epsilon T$ gives:

$$F = diag(\pi) M + [diag(\pi) M T + T^t diag(\pi) M] \epsilon + O(\epsilon^2),$$

$$F. * \log F = [diag(\pi) M]. * \log(diag(\pi) M) +$$

$$\Big\{ [diag(\pi) M T + T^t diag(\pi) M]. * [I + \log(diag(\pi) M)] \Big\} \epsilon +$$

$$O(\epsilon^2) \tag{26}$$

Substituting these in eq. (25) gives, after simplification, the result (22). ■

We note that prop. 1 above is a generalization of the result obtained by [5] for a binary alphabet.

Turning now into the symmetric binary case, the first eleven orders of the series expansion were given in [7], but only the first two were proved to be correct. Thm. 1 proves the correctness of the entire expansion from [7], which is not repeated here.

B. The almost memoryless expansion

By Thm. 2, one can expand the entropy rate around $M = U$ by simply computing the coefficients $C_N^{(k)}$ for N large enough. For example, by computing C_2 we have established, in analogy to prop. 1, the first order:

Proposition 2: Let $M = U + \delta T$. Then $\bar{H}$ satisfies:

$$\bar{H} = \log s - s^{-1} \xi^t R [\log(R^t \xi)] -$$

$$\xi^t \Big[(s^{-1} R^t T R). * log(s^{-1} R^t U R) \Big] \xi \delta + O(\delta^2) \tag{27}$$

Proof: Since $\bar{H} = C_2 + O(\delta^2)$, we expand C_2 (as given in eq. (25)) in δ. M is simply replaced by $U + \delta T$. Dealing with π is more problematic. Note that the stationary distribution of U is $s^{-1} \xi$. We write $\pi = s^{-1} \xi + \delta \psi + O(\delta^2)$, and solve:

$$(s^{-1} \xi^t + \delta \psi^t)(U + \delta T) = (s^{-1} \xi^t + \delta \psi^t) + O(\delta^2) \tag{28}$$

It follows that ψ should satisfy $\psi^t (I - U) = \xi^t T$, where I is the identity matrix. We cannot invert $I - U$ since it is of rank $s - 1$. The extra equation needed for determining ψ uniquely comes from the requirement $\sum_{i=1}^{s} \psi_i = 0$. Substituting $M = U + \delta T$ and $\pi = s^{-1} \xi + \delta \psi + O(\delta^2)$ in eq. (25), one gets:

$$C_2 = \Big\{ log(s^{-1} \xi^t R) s^{-1} R^t U R -$$

$$\xi^t [(s^{-1} R^t U R). * log(s^{-1} R^t U R)] \Big\} \xi +$$

$$\Big\{ \log(s^{-1} \xi^t R) R^t [s^{-1} diag(\xi) T + diag(\psi) U] R -$$

$$\xi^t \Big[\Big(R^t (s^{-1} diag(\xi) T + diag(\psi) U) R \Big). *$$

$$\Big(sU + \log(s^{-1} R^t U R) \Big) \Big] \Big\} \xi \delta + O(\delta^2) \tag{29}$$

After further simplification, most terms in eq. (29) cancel out, and we are left with the result (27). ■

In [12] it was shown that the first order term vanishes for the symmetric binary case, which is consistent with eq. (27). Our result holds for general alphabets and process parameters. Looking at the symmetric binary case might be misleading here, since by doing so one fails to see the linear behavior in δ for the general case.

We have computed higher orders for the symmetric binary case by expanding C_N for $N = 8$, which gives us $C^{(k)}$ for $k \leq 13$. In this case the expansion is in the parameter $\delta = \frac{1}{2} - p$, and gives (for better readability the dependency on ϵ is represented here via $\mu = 1 - 2\epsilon$):

$$\bar{H} = \log(2) - \mu^4 \Big[2\delta^2 + \frac{4}{3}(7\mu^4 - 12\mu^2 + 6)\delta^4 +$$

$$\frac{32}{15}(46\mu^8 - 120\mu^6 + 120\mu^4 - 60\mu^2 + 15)\delta^6 +$$

$$\frac{32}{21}(1137\mu^{12}-4088\mu^{10}+5964\mu^{8}-4536\mu^{6}+1946\mu^{4}-504\mu^{2}+84)\delta^{8}+\frac{512}{45}(3346\mu^{16}-15120\mu^{14}+28800\mu^{12}-30120\mu^{10}+18990\mu^{8}-7560\mu^{6}+1980\mu^{4}-360\mu^{2}+45)\delta^{10}+\frac{1024}{165}(159230\mu^{20}-874632\mu^{18}+2091100\mu^{16}-2857360\mu^{14}+2465100\mu^{12}-1400960\mu^{10}+532312\mu^{8}-135960\mu^{6}+24145\mu^{4}-3300\mu^{2}+330)\delta^{12}\Bigg]+O(\delta^{14}); \qquad (30)$$

The above expansion generalizes a result from [12], who proved $\bar{H}=\log(2)-2\mu^4\delta^2+o(\delta^2)$. Note that for the first few coefficients, all odd powers of δ vanish, and the coefficients are all polynomials of μ^2, which makes this series simpler than the one obtained in the High-SNR regime ([7]).

IV. Radius of Convergence

The usefulness of a series expansion such as the ones derived in eq. (30) and in [7] for practical purposes, highly depends on the radius of convergence. Determining the radius is a difficult problem, as it relates to the domain of analyticity of $\bar{H}$. In Thm. 2 we proved that the radius for the A-M expansion is positive.

For the High-SNR case, we gave a numerical estimation of the radius of convergence $\rho(p)$ as a function of p ([8]), based on the first few known terms. When one applies the same procedure to the coefficients of the A-M expansion, the numerical values of the estimated radius are much higher. The difference is demonstrated in fig. 1. In this figure, the (finite) series expansions with up to twelfth order is compared to two known bounds on $\bar{H}$ from [4]. The upper bound is simply $C_N = H(Y_N|[Y]_1^{N-1})$ and the lower bound is $c_N \equiv H(Y_N|X_1,[Y]_1^{N-1})$, for $N=2$. As can be seen from the figure, for the High-SNR case at $p=0.2$, the finite-order expansions are not within the bounds for large values of ϵ. For the A-M case, for $\epsilon=0.2$, the finite-order expansions remain within the bounds for any $0<p<\frac{1}{2}$.

The estimated radius $\rho(p)$ for the High-SNR expansion, is plotted as a function of p in fig. 2.a. In our context, the result of [9] proves that $\bar{H}(p,\epsilon)$ is real analytic in the domain $\Omega\subset\mathbb{R}^2$, $\Omega=\{(p,\epsilon):0<p,\epsilon<1\}$ (it is not known whether Ω is maximal with that respect). This domain is shown in fig. 2.b. For any $0<\epsilon<1$, the A-M expansion is near the point $(\epsilon,\frac{1}{2})$ which is an interior point of Ω. The High-SNR expansion is near some point $(p,0)$, which lies on the boundary of Ω.

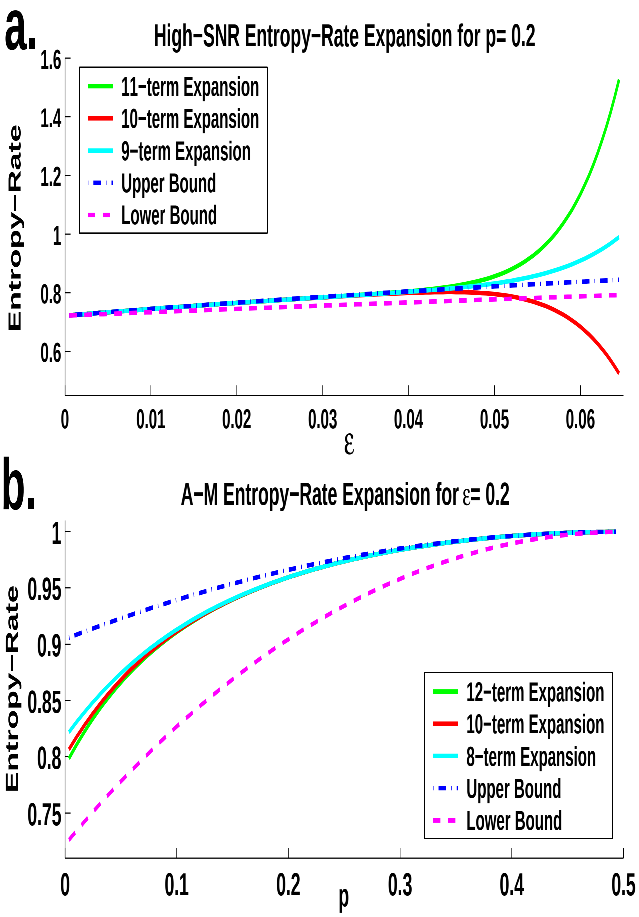

Fig. 1. Approximations for $\bar{H}$ using first few terms in its series expansion. a. The High-SNR expansions using $9, 10$ and 11 terms for $p=0.2$ deviate from the bounds for large values of ϵ. The first few terms of the expansion have alternating signs, therefore the direction of the deviation is determined by the parity of the number of terms taken. b. The A-M expansions using $8, 10$ and 12 terms for $\epsilon=0.2$ remain within the bounds for any value of p.

V. Conclusion

We presented a generalization and proof of the conjecture introduced in [7], relating the expansion coefficients of finite system entropies to those of the entropy rate for *HMPs*. Our new theorems shed light on the connection between finite and infinite chains, as well as give a practical and straightforward way to compute the entropy rate as a series expansion up to an arbitrary power.

The surprising 'settling' of the expansion coefficients $C_N^{(k)}=C^{(k)}$ for $N\geq\lceil\frac{k+3}{2}\rceil$, holds for the entropy. For other functions involving only conditional probabilities (e.g. relative entropy between two *HMPs*) a weaker result holds: the coefficients

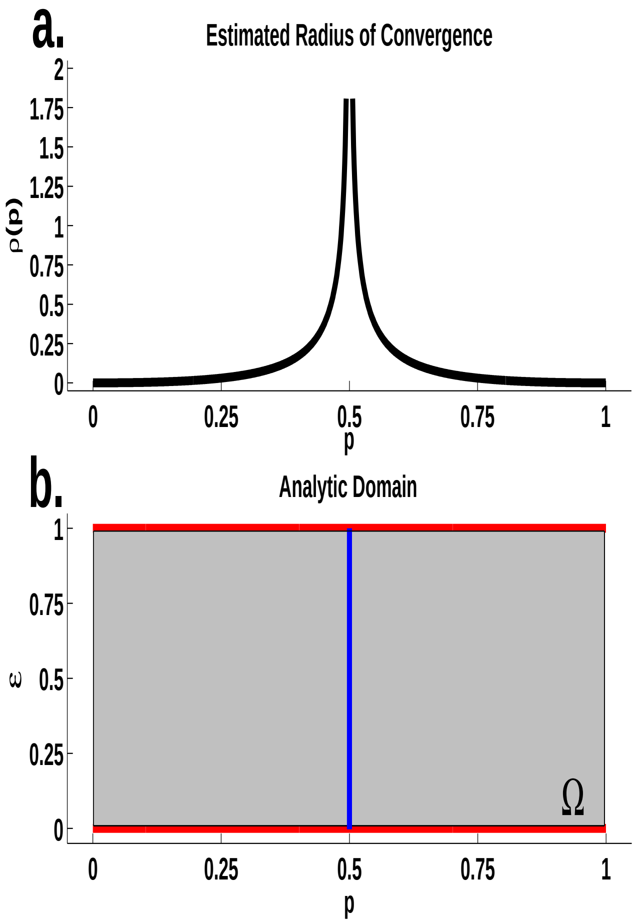

Fig. 2. a. The estimated radius of convergence $\rho(p)$ for the High-SNR expansion as a function of p. b. The domain Ω (shaded gray area) in the $\mathbb{R}^2$ plane for which it is known [9] that $\bar{H}$ is real analytic in (p, ϵ). The A-M expansion is near the vertical line $p = \frac{1}{2}$. The High-SNR expansion is near the horizonal boundaries at $\epsilon = 0$ and $\epsilon = 1$.

'settle' for $N \geq k + 2$. We note that this is still a highly non-trivial result, as it is known that for other regimes (e.g. 'rare-transitions' [13]), a finite chain of any length does not give the correct asymptotic behavior even to the first order. We also estimated the radius of convergence for the expansion in the two regimes, 'High-SNR' and 'A-M', and demonstrated their quantitatively different behavior. Further research in this direction, which closely relates to the domain of analyticity of the entropy rate, is still required.

Acknowledgment

M.A. is grateful for the hospitality shown him at the Weizmann Institute, where his work was supported by the Einstein Center for Theoretical Physics and the Minerva Center for Nonlinear Physics. The work of I.K. at the Weizmann Institute was supported by the Einstein Center for Theoretical Physics. E.D. and O.Z. were partially supported by the Minerva Foundation and by the European Community's Human Potential Programme under contract HPRN-CT-2002-00319, STIPCO.

References

[1] Y. Ephraim and N. Merhav, *Hidden Markov processes*, IEEE Tran. Inf. Th., 48(6), pp. 1518-1569, 2002.

[2] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proc. IEEE, 77(2), pp. 257-286, 1989.

[3] C. E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal, 27, pp. 379-423 and 623-656, 1948.

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

[5] P. Jacquet, G. Seroussi and W. Szpankowski, *On the Entropy of a Hidden Markov Process*, DCC 2004, pp. 362-371.

[6] E. Ordentlich and T. Weissman, *New Bounds on the Entropy Rate of Hidden Markov Processes*, San Antonio Inf. Th. Workshop, 2004.

[7] O. Zuk, I. Kanter and E. Domany, *Asymptotics of the Entropy Rate for a Hidden Markov Process*, DCC 2005, pp. 173-182.

[8] O. Zuk, I. Kanter and E. Domany, *The Entropy of a Binary Hidden Markov Process*, J. Stat. Phys. 121(3-4), pp. 343-360, 2005.

[9] G. Han and B. Marcus *Analyticity of Entropy Rate in Families of Hidden Markov Chains*, submitted to IEEE Tran. Inf. Th.

[10] J. Lorinczi, C. Maes and K. Vande Velde *Transformations of Gibbs measures*, Prob. Th. Related Fields, 112, pp. 121-147, 1998.

[11] R. L. Dobrushin and S. B. Shlosman, *Completely analytical interactions: Constructive description*, J. Stat. Phys. 46(5-6), pp. 983-1014, 1987.

[12] E. Ordentlich and T. Weissman, *On the optimality of symbol-by-symbol filtering and denoising*, IEEE Tran. Inf. Th. 52(1) pp. 19-40, 2006.

[13] C. Nair, E. Ordentlich and T. Weissman, *On asymptotic filtering and entropy rate for a hidden Markov process in the rare transitions regime*, ISIT 2005, pp. 1838-1842.