From finite-system entropy to entropy rate for a Hidden Markov Process

Or Zuk*, Eytan Domany*, Ido Kanter†, and Michael Aizenman‡

January 23, 2006

Abstract

A recent result presented the expansion for the entropy rate of a Hidden Markov Process (HMP) as a power series in the noise variable ϵ. The coefficients of the expansion around the noiseless ($\epsilon = 0$) limit were calculated up to 11^{th} order, using a conjecture that relates the entropy rate of a HMP to the entropy of a process of finite length (which is calculated analytically). In this paper we generalize and prove the conjecture, and discuss its theoretical and practical consequences.

1 Introduction

Let $\{X_N\}$ be a finite state stationary Markov process over the alphabet $\Sigma = \{1, .., s\}$, and let $\{Y_N\}$ be its noisy observation (on the same alphabet). The process Y is generated by the Markov transition matrix $M = M_{s\times s} = \{m_{ij}\}$ and the emission matrix $I + \epsilon T$, where I is the $s \times s$ identity matrix, the matrix $T = T_{s\times s} = \{t_{ij}\}$ satisfies $t_{ii} < 0$, $t_{ij} \geq 0$, $\forall i \neq j$ and $\sum_{j=1}^{s} t_{ij} = 0$, and $\epsilon > 0$ is some constant. (There is no loss of generality here, as any stochastic matrix can be represented as $I+\epsilon T$.) This yields the probabilities $P(X_{N+1} = j|X_N = i) = m_{ij}$ and $P(Y_N = j|X_N = i) = \delta_{ij} + \epsilon t_{ij}$, where δ is Kronecker's delta. We consider the case of high signal to noise ratio ('High-SNR'), characterized by small values of ϵ, and assume strictly positive M ($m_{ij} > 0$) with a unique stationary distribution.

The process Y can be viewed as an observation of X through a noisy channel. It is a *Hidden Markov Process (HMP)*, governed by the parameters M, T and ϵ. *HMPs* have a rich

*E.D. and O.Z. are in the Dept. of Physics of Complex Systems, Weizmann Inst. of Science, Rehovot, 76100, Israel, Tel: 972-8-9344984, Email: {or.zuk/eytan.domany}@weizmann.ac.il

†I.K. is in the Dept. of Physics, Bar-Ilan Univ., Ramat-Gan, 52900, Israel

‡M.A. is in the Depts. of Physics and Mathematics, Princeton Univ., Jadwin Hall Princeton, NJ 08544-0708

theory, with applications in various fields, such as speech recognition ([1]), information theory ([2]) and signal processing ([3]). While we concentrate on a finite-state first-order *HMP*, our results can be easily generalized to more cases (e.g. continuous observations).

An important quantity for a stochastic process is the Shannon entropy rate, which measures its 'uncertainty per-symbol' ([4]). More formally, for $i \leq j$ let $[Y]_i^j$ denote the vector $(Y_i, .., Y_j)$. The entropy rate of Y is defined as:

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

Where $H(Y) = -\sum_Y P(Y) \log P(Y)$; Sometimes we omit the realization y of the variable Y, so $P(Y)$ should be understood as $P(Y = y)$. For a finite-entropy stationary process the limit (1) exists and $\bar{H}$ can also be computed via the conditional entropy ([5]) as: $\bar{H}(Y) = \lim_{N \to \infty} H(Y_N | [Y]_1^{N-1})$. Here $H(U|V)$ represents the conditional entropy, which for random variables U and V is the average uncertainty of the conditional distribution of U given V, that is $H(U|V) = \sum_v P(V = v) H(U|V = v)$. By the entropy chain rule, it is also given as a difference of entropies, $H(U|V) = H(U, V) - H(V)$. This relation will be used below.

There is at present no explicit expression for the entropy rate of a *HMP* ([2, 6]). Few recent works ([6, 7, 8]) have studied the asymptotic behavior of $\bar{H}$ in several regimes, albeit giving rigorously only bounds or at most second ([8]) order behavior. Here we generalize and prove a relationship, first posed in [8] as a conjecture, thereby turning the computation presented there, of $\bar{H}$ as a series expansion up to 11th order in ϵ, into a rigorous statement.

2 Theorem Statement and Proof

We first state our main result, which will be proven at the end of the section.

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:*

$$C_N(M, T, \epsilon) = \sum_{k=0}^{\infty} C_N^{(k)} \epsilon^k, \quad \bar{H}(M, T, \epsilon) = \sum_{k=0}^{\infty} C^{(k)} \epsilon^k \tag{2}$$

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

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

Analyticity of $\{C_N\}$ and $\bar{H}$ around $\epsilon = 0$ was recently shown in [9]. One may also use [10], which showed that the law of the process Y is Gibbsian, together with the complete analyticity results for Gibbsian measures of [12], to deduce analyticity of $\bar{H}$. C_N is in fact an upper-bound ([5]) for $\bar{H}$. The behavior stated in Thm. 1 was discovered using symbolic computations, but was proven only for $k \leq 2$, in the binary symmetric case ([8]). Although it may appear technically involved, our proof is based on two simple ideas.

First, we distinguish between the noise parameters at different sites. We thus consider a more general process $\{Z_N\}$, where Z_i's emission matrix is $I + \epsilon_i T$. The process $\{Z_N\}$ is determined by M,T and $[\epsilon]_1^N$. We define the following functions:

$$F_N(M, T, [\epsilon]_1^N) = H([Z]_1^N) - H([Z]_1^{N-1}) \tag{4}$$

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

Second, we observe that if a particular ϵ_i is set to zero, we must have $Z_i = X_i$. Thus, conditioning back to the past is 'blocked'. This is used to prove:

Lemma 1 *If $\epsilon_j = 0$ for some $1 < j < N$, then:*

$$F_N([\epsilon]_1^N) = F_{N-j+1}([\epsilon]_j^N) \tag{5}$$

Proof *F can be written as the sum:*

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

The dependence on $[\epsilon]_1^N$ and M, T is hidden in the probabilities $P(..)$. Since $\epsilon_j = 0$, we have

$X_j = Z_j$, and conditioning further to the past is 'blocked':

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

Substituting in eq. 6 gives:

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

$$-\sum_{[Z]_j^N} P([Z]_j^N) \log P(Z_N|[Z]_j^{N-1}) = F_{N-j+1} \tag{8}$$

■

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:

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

$C_N^{(k)}$ is obtained by summing the contributions $F_N^{\vec{k}}$ of all the vectors $\vec{k}$'s with weight k:

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

The next lemma shows that many such $\vec{k}$'s give zero contribution to the sum:

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

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

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

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

Assume now $k_j = 1$. Write the probability of Z:

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

Let $[Z]_1^{N(j\to a)}$ *denote the vector we get from* $[Z]_1^N$ *by changing* Z_j *to* a *(while keeping other coordinates). Differentiating with respect to* ϵ_j *gives (see [11] for more details):*

$$\frac{\partial P([Z]_1^N)}{\partial \epsilon_j}\bigg|_{\epsilon_j=0} = \sum_{[X]_1^N} \left[P([X]_1^N) t_{X_j Z_j} \prod_{i\neq j} (\delta_{X_i Z_i} + \epsilon_i t_{X_i Z_i}) \right]\Bigg|_{\epsilon_j=0} =$$

$$\left\{ \sum_{a=1}^{s} t_{aZ_j} P([Z]_1^{N(j\to a)}) \right\}\Bigg|_{\epsilon_j=0} \tag{13}$$

By Bayes' rule $P(Z_N|[Z]_1^{N-1}) = \frac{P([Z]_1^N)}{P([Z]_1^{N-1})}$*, we get:*

$$\frac{\partial P(Z_N|[Z]_1^{N-1})}{\partial \epsilon_j}\bigg|_{\epsilon_j=0} = \frac{1}{P([Z]_1^{N-1})} \sum_{a=1}^{s} t_{aZ_j} \Big[P([Z]_1^{N(j\to a)}) -$$

$$P(Z_N|[Z]_1^{N-1}) P([Z]_1^{N-1(j\to a)}) \Big]\Big|_{\epsilon_j=0} \tag{14}$$

This gives:

$$\frac{\partial [P([Z]_1^N) \log P(Z_N|[Z]_1^{N-1})]}{\partial \epsilon_j}\bigg|_{\epsilon_j=0} = \sum_{a=1}^{s} t_{aZ_j} \Big\{ P([Z]_1^{N(j\to a)}) \log P(Z_N|[Z]_1^{N-1}) +$$

$$P([Z]_1^{N(j\to a)}) - P(Z_N|[Z]_1^{N-1}) P([Z]_1^{N-1(j\to a)}) \Big\}\Big|_{\epsilon_j=0} \tag{15}$$

And therefore:

$$\frac{\partial F_N}{\partial \epsilon_j}\bigg|_{\epsilon_j=0} =$$

$$-\sum_{a=1}^{s} t_{aZ_j} \left\{ \sum_{[Z]_1^N} \Big[P([Z]_1^{N(j\to a)}) \log P(Z_N|[Z]_1^{N-1}) - P(Z_N|[Z]_1^{N-1}) P([Z]_1^{N-1(j\to a)}) \Big] \right\}\Bigg|_{\epsilon_j=0} =$$

$$\left\{ -\sum_{a=1}^{s} t_{aZ_j} \sum_{[Z]_j^N} \Big[P([Z]_j^{N(1\to a)}) \log P(Z_N|[Z]_j^{N-1}) - P(Z_N|[Z]_j^{N-1}) P([Z]_j^{N-1(1\to a)}) \Big] \right\}\Bigg|_{\epsilon_1=0} \tag{16}$$

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

Before proving Thm. 1, we show 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 $\vec{k}$ and r zeros to the left: $\vec{k}^{(r)} = (\underbrace{0,..,0}_{r}, k_1,..,k_N)$. Then:*

$$F_N^{\vec{k}} = F_{r+N}^{\vec{k}^{(r)}} \quad , \forall r \in \mathbb{N} \tag{17}$$

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

$$F_{r+N}^{\vec{k}^{(r)}}([\epsilon]_1^{r+N}) = \frac{\partial^{\omega(\vec{k}^{(r)})} F_{r+N}([\epsilon]_1^{r+N})}{\partial\epsilon_{r+2}^{k_2},..,\partial\epsilon_{r+N}^{k_N}}\Bigg|_{\vec{\epsilon}=0} = \frac{\partial^{\omega(\vec{k})} F_N([\epsilon]_{r+1}^{r+N})}{\partial\epsilon_{r+2}^{k_2},..,\partial\epsilon_{r+N}^{k_N}}\Bigg|_{\vec{\epsilon}=0} = F_N^{\vec{k}}([\epsilon]_{r+1}^{r+N}) \tag{18}$$

The case $k_1 = 1$ is reduced back to the case $k_1 = 0$ by taking the derivative. Using eqs. (16,18), we get:

$$F_{N+1}^{\vec{k}^{(1)}}([\epsilon]_1^{N+1}) = \frac{\partial^{\omega(\vec{k})-1}}{\partial\epsilon_3^{k_2}\ldots\partial\epsilon_{N+1}^{k_N}}\left[\frac{\partial F_{N+1}}{\partial\epsilon_2}\bigg|_{\epsilon_2=0}\right]\Bigg|_{\vec{\epsilon}=0} =$$

$$\frac{\partial^{\omega(\vec{k})-1}}{\partial\epsilon_3^{k_2}\ldots\partial\epsilon_{N+1}^{k_N}}\Bigg\{-\sum_{a=1}^{s} t_{aZ_2}\sum_{[Z]_1^{N+1}}\Big[P([Z]_1^{N+1(2\to a)})\log P(Z_{N+1}|[Z]_1^N) -$$

$$P(Z_{N+1}|[Z]_1^N)P([Z]_1^{N(2\to a)})\Big]\Big|_{\epsilon_2=0}\Bigg\}\Bigg|_{[\epsilon]_1^{N+1}=0} =$$

$$\frac{\partial^{\omega(\vec{k})-1}}{\partial\epsilon_2^{k_2}\ldots\partial\epsilon_N^{k_N}}\Bigg\{-\sum_{a=1}^{s} t_{aZ_2}\sum_{[Z]_1^N}\Big[P([Z]_1^{N(1\to a)})\log P(Z_N|[Z]_1^{N-1}) -$$

$$P(Z_N|[Z]_1^{N-1})P([Z]_1^{N(1\to a)})\Big]\Big|_{\epsilon_1=0}\Bigg\}\Bigg|_{[\epsilon]_1^N=0} = F_N^{\vec{k}}([\epsilon]_1^N) \tag{19}$$

This proved the claim for $r = 1$. The claim for larger r's follows by induction. ■

We are now ready to prove our main theorem, which follows directly from lemmas 2 and 3:

Proof (Thm. 1) Let $\vec{k} = [k]_1^N$ with $\omega(\vec{k}) = k$. Define its 'length' as $l(\vec{k}) = N+1-\min_{k_i>1}\{i\}$. It easily follows from lemma 2 that $F_N^{\vec{k}} \neq 0 \Rightarrow l(\vec{k}) \leq \lceil\frac{k+3}{2}\rceil - 1$. Thus, according to lemma 3, we have:

$$F_N^{\vec{k}} = F_{\lceil\frac{k+3}{2}\rceil}^{(k_{N-\lceil\frac{k+3}{2}\rceil+1},..,k_N)} \tag{20}$$

for all $\vec{k}$'s in the sum. Summing over all $F_N^{\vec{k}}$ with the same 'weight' gives $C_N^{(k)} = C_{\lceil \frac{k+3}{2} \rceil}^{(k)}$, $\forall N > \lceil \frac{k+3}{2} \rceil$. But from the analyticity of C_N and $\bar{H}$ near $\epsilon = 0$ it can be shown by induction that $\lim_{N \to \infty} C_N^{(k)} = C^{(k)}$, therefore $C_N^{(k)} = C^{(k)}$, $\forall N \geq \lceil \frac{k+3}{2} \rceil$. ∎

3 Conclusion

The theorem proven above sheds light on the connection between finite and infinite chains, and gives a practical and straightforward way to compute the entropy rate as a series expansion in ϵ 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 'settle' for $N \geq k + 2$. One can expand the entropy rate in several parameter regimes. As it turns out, exactly the same 'settling' as was proven in Thm. 1 happens in the 'almost memoryless' regime, where the transition matrix M is close to a matrix which makes the X_i's i.i.d, (i.e. a matrix whose rows are identical). This and other regimes, as well as the analytic behavior of the *HMP* ([9]), will be discussed elsewhere.

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

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

[3] M. Crouse, R. Nowak, R. Baraniuk *Wavelet-Based Statistical Signal Processing Using Hidden Markov Models*, IEEE Tran. Sig. Proc. 46(4), pp. 886-902, 1998.

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

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

[6] P. Jacquet, G. Seroussi and W. Szpankowski, *On the Entropy of a Hidden Markov Process*, Data Compression Conference, Snowbird 2004, pp. 362-371.

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

[8] O. Zuk, I. Kanter and E. Domany, *Asymptotics of the Entropy Rate for a Hidden Markov Process*, Data Compression Conference, Snowbird 2005, pp. 173-182.

[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] For detailed proofs see: O. Zuk, I. Kanter, E. Domany and M. Aizenman *Taylor Series Expansions for the Entropy Rate of Hidden Markov Processes*, International Conference on Communication, Istanbul 2006, accepted.

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