The Relative Entropy Rate For Two Hidden Markov Processes

Or Zuk

Dept. of Physics of Complex Systems, Weizmann Inst. of Science, Rehovot, 76100, Israel, or.zuk@weizmann.ac.il

Abstract

The relative entropy rate is a natural and useful measure of distance between two stochastic processes. In this paper we study the relative entropy rate between two Hidden Markov Processes (HMPs), which is of both theoretical and practical importance. We give new results showing analyticity, representation using Lyapunov exponents, and Taylor expansion for the relative entropy rate of two discrete-time finite-alphabet HMPs.

1 Introduction

Let $\{X_N\}$ be a finite state stationary Markov process over the alphabet $\Sigma = \{1,..,s\}$. Let $\{Y_N\}$ be its noisy observation (on the same alphabet). We consider two probability laws for the processes $\{X_N, Y_N\}$, denoted λ, μ. For the law η ($\eta = \lambda, \mu$) we let $M_\eta = \{m_\eta(i,j)\}$ denote the Markov transition matrix and $R_\eta = \{r_\eta(i,j)\}$ denote the emission matrix, i.e. M_η and R_η are $s \times s$ real nonnegative stochastic matrices, $P_\eta(X_{N+1} = j|X_N = i) = m_\eta(i,j)$ and $P_\eta(Y_N = j|X_N = i) = r_\eta(i,j)$. We further assume that the Markov matrices M_η are strictly positive ($m_\eta(i,j) > 0$), and denote their stationary distributions by π_η. Note that the law η is entirely determined by M_η and R_η and we may write $\eta = \{M_\eta, R_\eta\}$.

The process Y can be viewed as a noisy version of X, observed through a noisy channel. It is known as a *Hidden Markov Process (HMP)*. *HMPs* have a rich and developed theory, and enormous applications in various fields (see [1], [2]). λ and μ can be viewed as the laws generating the process, and are also frequently termed *Hidden Markov Models (HMMs)*.

For the model η ($\eta = \lambda, \mu$), we associate a family of probability measures $\{P_\eta^{(N)}\}, P_\eta^{(N)} : \Sigma^N \to [0,1]$, defined by

$$P_\eta^{(N)}([Y]_1^N) = \sum_{[X]_1^N} \left\{ \pi_\eta(X_1) \prod_{i=1}^{N-1} m_\eta(X_i, X_{i+1}) \prod_{i=1}^{N} r_\eta(X_i, Y_i) \right\} \quad (1)$$

Where here and throughout the paper, $[X]_i^j$ denotes the vector $(X_i,..,X_j)$, uppercase denote random variables while lower case denote their realizations, and the latter are often omitted, i.e. $P_\eta(X)$ stands for $P_\eta(X = x)$. When no confusion may occur, we shall also omit the (N) superscript and simply write P_η.

In many applications, such as classification, monitoring the training process, or clustering, the need for a dissimilarity measure between two *HMP*s arises. A natural and appealing choice is the relative entropy rate (RE-rate), first introduced in [3] for *HMP*s. Let $D(P_\lambda^{(N)}||P_\mu^{(N)})$ be the relative entropy ([4]) distance (also known as Kullback-Leibler distance, or cross-entropy) between the two probability distributions[1]:

$$D(P_\lambda^{(N)}||P_\mu^{(N)}) = \sum_{[Y]_1^N} P_\lambda^{(N)}([Y]_1^N) \log \frac{P_\lambda^{(N)}([Y]_1^N)}{P_\mu^{(N)}([Y]_1^N)} \quad (2)$$

Then the RE-rate (also known as the Kullback-Leibler divergence rate), is defined by:

$$D(\lambda||\mu) = \lim_{N\to\infty} \frac{1}{N} D(P_\lambda^{(N)}||P_\mu^{(N)}) \quad (3)$$

$D(\lambda||\mu)$ can also be computed via the conditional relative-entropy ([5]), $D(\lambda||\mu) = \lim_{N\to\infty} D_N(\lambda||\mu)$, where D_N is defined as:

$$D_N(\lambda||\mu) = \sum_{[Y]_1^N} P_\lambda([Y]_1^N) \log \frac{P_\lambda(Y_N|[Y]_1^{N-1})}{P_\mu(Y_N|[Y]_1^{N-1})} \quad (4)$$

For two stationary ergodic *HMM*s, it is known that the above two limits exist and coincide ([3], [6]). Using the chain-rule for relative-entropies ([5]), D_N is also given as $D_N(\lambda||\mu) = D(P_\lambda^{(N)}||P_\mu^{(N)}) - D(P_\lambda^{(N-1)}||P_\mu^{(N-1)})$, which will be used later.

Although it is not a norm, the RE-rate has several natural interpretations. For example, it represent the discriminative power of one model over the other. Thus if data is generated by the model λ, then $D(\lambda||\mu)$

[1] For simplicity, we use natural logarithms throughout the paper

represents the average difference in the likelihood score per-symbol between λ and μ. $D(\lambda||\mu)$ is also the average loss-per-symbol when compressing the data, assuming (erroneously) it was generated by μ ([5]).

Lately, many new results were obtained for the Shannon entropy-rate ([13]) of a *HMP*. These include analyticity ([8]), representation as a Lyapunov exponent for a random matrix product ([9], [10]), and asymptotic evaluations in various regimes ([10], [11], [12]). The Shannon entropy rate $\bar{H}(\lambda)$ is related to a special case of the RE-rate where one of the models is uniform, using the identity $\bar{H}(\lambda) = \log s - D(\lambda||u)$, where u is a uniform model (say, with $r_u(i,j) = s^{-1}$). The main purpose of this paper is to apply the same methods used in the aforementioned papers, and derive similar, yet more general results, for the RE-rate.

In section 2 we represent the RE-rate as a difference of two Lyapunov exponents, with the same probability laws, albeit different matrix values. In section 3 we prove, under mild positivity assumption, that the RE-rate is analytic in the processes parameters. While the RE-rate for two Markov chains is known ([7], [14]), there is at present no explicit expression for the RE-rate of two *HMP*s, in terms of the parameters of the two underlying models. So far, only bounds ([15]) or approximation algorithms ([7], [16], [17]) were obtained. In section 4 we study the representation of the RE-rate as a Taylor series expansion in various parameters. We show a relation between the Taylor series coefficients of the relative-entropy conditioned on finite histories, to those of the RE-rate, in two different parameter regimes. We also demonstrate the applicability of our results, by giving the first order asymptotic behavior of the RE-rate in one of the regimes. We end with conclusions and future directions.

2 Relative Entropy Rate and Lyapunov Exponents

It was previously shown ([9], [10]) that the entropy rate of a *HMP* can be represented as a top Lyapunov exponent of a random matrix product. Here we extend this result, and represent the RE-rate as a difference of two Lyapunov Exponents. For $\eta = \lambda, \mu$ and $a = 1, .., s$, we define the matrices $G_\eta^{(a)} = \{g_\eta^{(a)}(i,j)\}$ as:

$$g_\eta^{(a)}(i,j) = P_\eta(X_{N+1} = i, Y_{N+1} = a | X_N = j) = m_\eta(j,i) r_\eta(i,a) \tag{5}$$

Let $\rho_\eta^{(N)}$ be the (random) vector defined by

$$\rho_\eta^{(N)}(i) = P_\eta([Y]_1^N, X_N = i) \tag{6}$$

Using the forward equations, one can write the recursion relation

$$\rho_\eta^{(N)} = G_\eta^{(Y_N)} \rho_\eta^{(N-1)} \tag{7}$$

With the initial vector $\rho_\eta^{(1)}$ given by $\rho_\eta^{(1)}(i) = \pi_\eta(i) r_\eta(i, Y_1)$. By induction, the joint probability function $P_\eta([Y]_1^N)$, is given by:

$$P_\eta([Y]_1^N) = \xi^t \prod_{i=N}^{2} G_\eta^{(Y_i)} \rho_\eta^{(1)} \tag{8}$$

Where here and throughout the paper, ξ denotes the (column) vector of s ones. Therefore, the RE-rate can be written as:

$$D(\lambda||\mu) = \lim_{N\to\infty} \frac{1}{N} E_\lambda \Big[\log(\xi^t \prod_{i=N}^{2} G_\lambda^{(Y_i)} \rho_\lambda^{(i)}) - \log(\xi^t \prod_{i=N}^{2} G_\mu^{(Y_i)} \rho_\mu^{(1)}) \Big] \tag{9}$$

The mappings $A \to \xi^t A \rho_\eta^{(1)}, \quad \eta = \lambda, \mu$ are easily shown to satisfy the requirements for being matrix norms. Moreover, the above limit exists also when taking only the first term in the sum (and is equal to the minus of the entropy rate). This immediately gives:

Proposition 1: $D(\lambda||\mu)$ is the difference of the two top Lyapunov exponents:

$$D(\lambda||\mu) = \lim_{N\to\infty} \frac{1}{N} E_\lambda \Big\| \log(\prod_{i=N}^{2} G_\lambda^{(Y_i)}) \Big\| - \lim_{N\to\infty} \frac{1}{N} E_\lambda \Big\| \log(\prod_{i=N}^{2} G_\mu^{(Y_i)}) \Big\| \tag{10}$$

Note that different matrices appear in the above two Lyapunov exponents, but the probability law of selecting the matrices is the same. The matrices are chosen according to the Markovian law, where the probability of picking $G_\eta^{(b)}$ after $G_\eta^{(a)}$ is

$$P_\lambda(Y_{N+1} = b | Y_N = a) = \sum_{X_N, X_{N+1}} P_\lambda(X_N, X_{N+1}, Y_{N+1} = b | Y_N = a) = \frac{\sum_{i,j=1}^{s} \pi_\lambda(i) r_\lambda(i,a) m_\lambda(i,j) r_\lambda(j,b)}{\sum_{j=1}^{s} \pi_\lambda(j) r_\lambda(j,a)} \tag{11}$$

One should not be concerned by the fact that a different norm was used for each exponent, as the Lyapunov exponent is independent of the norm chosen ([18]).

3 Analyticity

It was recently shown ([8]) that under mild positivity assumptions, the entropy rate of a *HMP* is analytic in the underlying process parameters. We show here that a similar result holds also for the RE-rate:

Theorem 1: Let $M(s,\mathbb{C})$ be the ring of complex square $s\times s$ matrices. Let $\Gamma\subset M(s,\mathbb{C})^4$ be the hyperplane defined by:

$$\Gamma=\Big\{(M_\lambda,R_\lambda,M_\mu,R_\mu)\Big|\forall\eta=\lambda,\mu,\ \forall i=1,..,s,\ \sum_{j=1}^{s}m_\eta(i,j)=\sum_{j=1}^{s}r_\eta(i,j)=1\Big\} \tag{12}$$

Then there is some relatively open domain $\Omega\subset\Gamma$, such that:

1) Ω contains the open domain of all real positive stochastic matrices:

$$\Lambda=\Big\{(M_\lambda,R_\lambda,M_\mu,R_\mu)\in\Gamma\Big|\ m_\eta(i,j)r_\eta(i,j)>0,\ \forall\eta=\lambda,\mu,\ \forall i,j=1,..,s\Big\} \tag{13}$$

2) $D(\lambda||\mu)$ is an analytic function of the parameters $(M_\lambda,R_\lambda,M_\mu,R_\mu)$ in Ω.

Proof: We represent the RE-rate as the sum:

$$D(\lambda||\mu)=-\bar{H}_\lambda-\lim_{N\to\infty}\frac{1}{N}E_\lambda\Big\|\log(\prod_{i=2}^{N}G_\mu^{(Y_i)})\Big\| \tag{14}$$

The first term $-\bar{H}_\lambda$, is simply the (minus) entropy rate of the process λ, which was recently shown to be analytic in some Ω, $\Omega\supset\Lambda$ ([8]). As for the second term, one can repeat the proof from [8] (with the appropriate modifications) and show analyticity by showing uniform convergence in some Ω of the finite conditional functions $\sum_{[Y]_1^N}P_\lambda([Y]_1^N)\log P_\lambda(Y_N|[Y]_1^{N-1})$. Alternatively, we observe that the parameters M_λ,R_λ influence only the probabilities in the Lyapunov exponent representation (prop. 1), whereas the parameters M_μ,R_μ influence only the matrix entries. Thus no parameter influence both, and we can rely directly on results from Lyapunov expnents theory, which guarantee the analyticity in both the matrices values themselves ([19], [20]), and their probabilities ([21]). ■

4 Taylor Expansions Using Finite-System Relative Entropies

In this section we show a relation between the Taylor series coefficients of finite conditional relative entropies, and those of the RE-rate. Our results apply in two specific parameters regimes. We then demonstrate an application of our results, by computing the first term in the Taylor series expansion for one of the regimes, termed 'High-SNR'.

4.1 The High SNR regime

Loosely speaking, the term high SNR regime represents a regime in the parameters domain in which the observations Y_N are likely to equal the hidden states X_N. In other words, the emission matrices $R_\eta\quad(\eta=\lambda,\mu)$ are close to the identity matrix I. We may therefore write $R_\eta=I+\epsilon T_\eta$, where $\epsilon>0$ is a small constant and $T_\eta=\{t_\eta(i,j)\}$ are matrices satisfying $t_\eta(i,i)<0$, $t_\eta(i,j)\geq 0$, $\forall i\neq j$ and $\sum_{j=1}^{s}t_\eta(i,j)=0$. The RE-rate in this regime can be given as an expansion in ϵ around zero. We state here our new theorem, connecting the relative entropy of finite systems to the RE-rate:

Theorem 2: Let $D_N\equiv D_N(\lambda||\mu,\epsilon)$ be the conditional relative entropy between the probability laws of λ and μ on a finite system of length N, where $R_\eta=I+\epsilon T_\eta$, $\eta=\lambda,\mu$. Assume[2] that there is some (complex) neighborhood of $\epsilon=0$ in which the (one-variable) functions $\{D_N\},D$ are analytic in ϵ, with a Taylor expansion given by

$$D_N(\lambda||\mu,\epsilon)=\sum_{k=0}^{\infty}D_N^{(k)}\epsilon^k,\ \ D(\lambda||\mu,\epsilon)=\sum_{k=0}^{\infty}D^{(k)}\epsilon^k \tag{15}$$

(Here the coefficients $D_N^{(k)},D^{(k)}$ are functions of the parameters M_λ,M_μ and T_λ,T_μ. From now on we omit this dependence). Then we have:

$$N\geq k+2\Rightarrow D_N^{(k)}=D^{(k)} \tag{16}$$

The behavior stated in Thm. 2 was discovered using symbolic computations, but was not proven before. A stronger statement (of settling of the coefficients for $N\geq\lceil\frac{k+3}{2}\rceil$) was proven (using similar methods) for the special case of the entropy rate of a *HMP* ([22]).

The proof of Thm. 2 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 according to the model η is $R_{\eta,i}=I+\epsilon_i T_\eta$.

[2]Analyticity around $\epsilon=0$ was shown in [8], albeit only for the entropy rate. The functions D_N are easily shown to be differentiable to all orders in ϵ, at $\epsilon=0$. The unproven assumption here is that they are also analytic with a radius of analyticity uniform in N, and are uniformly bounded within some common neighborhood of $\epsilon=0$

The joint distribution of $[Z]_1^N$ is thus determined by M_η,T_η and $[\epsilon]_1^N$. We define the following functions:

$$F_N(\lambda,\mu,[\epsilon]_1^N)=\sum_{[Z]_1^N}P_\lambda([Z]_1^N)\log\frac{P_\lambda(Z_N|[Z]_1^{N-1})}{P_\mu(Z_N|[Z]_1^{N-1})} \tag{17}$$

Setting all the ϵ_i's equal, reduces us back to the Y process, so in particular $F_N(\lambda,\mu,(\epsilon,..,\epsilon))=D_N(\epsilon)$. 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 :

$$F_N([\epsilon]_1^N)=F_{N-j+1}([\epsilon]_j^N) \tag{18}$$

Proof: F can be written as:

$$F_N=\sum_{[Z]_1^N}P_\lambda([Z]_1^N)\log\frac{P_\lambda(Z_N|[Z]_1^{N-1})}{P_\mu(Z_N|[Z]_1^{N-1})} \tag{19}$$

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, for $\eta=\lambda,\mu$:

$$\epsilon_j=0\Rightarrow P_\eta(Z_N|[Z]_1^{N-1})=P_\eta(Z_N|[Z]_j^{N-1}) \tag{20}$$

(Note that eq. (20) is true for $j<N$, but not for $j=N$). Substituting in eq. (19) gives:

$$\begin{aligned}F_N&=\sum_{[Z]_1^N}\left\{P_\lambda([Z]_1^{N-1})P_\lambda(Z_N|[Z]_j^{N-1})\right.\\&\left.\log\frac{P_\lambda(Z_N|[Z]_j^{N-1})}{P_\mu(Z_N|[Z]_j^{N-1})}\right\}=\sum_{Z_j^N}\left\{P_\lambda([Z]_j^{N-1})\right.\\&\left.P_\lambda(Z_N|[Z]_j^{N-1})\log\frac{P_\lambda(Z_N|[Z]_j^{N-1})}{P_\mu(Z_N|[Z]_j^{N-1})}\right\}=F_{N-j+1}\end{aligned} \tag{21}$$

■

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:

$$F_N^{\vec{k}}\equiv\left.\frac{\partial^{\omega(\vec{k})}F_N}{\partial\epsilon_1^{k_1},..,\partial\epsilon_N^{k_N}}\right|_{\vec{\epsilon}=0} \tag{22}$$

The next lemma shows that adding zeros to the left of $\vec{k}$ leaves $F_N^{\vec{k}}$ unchanged:

Lemma 2: Let $\vec{k}=[k]_1^N$ with $k_1=0$. Denote $\vec{k}^{(c)}$ the concatenation of $\vec{k}$ with c zeros to the left, $\vec{k}^{(c)}=(\underbrace{0,..,0}_{c},k_1=0,..,k_N)$. Then:

$$F_N^{\vec{k}}=F_{c+N}^{\vec{k}^{(c)}}\quad,\forall c\in\mathbb{N} \tag{23}$$

Proof: Using lemma 1, we get :

$$\begin{aligned}F_{c+N}^{\vec{k}^{(c)}}([\epsilon]_1^{c+N})&=\left.\frac{\partial^{\omega(\vec{k}^{(c)})}F_{c+N}([\epsilon]_1^{c+N})}{\partial\epsilon_{c+2}^{k_2},..,\partial\epsilon_{c+N}^{k_N}}\right|_{\vec{\epsilon}=0}=\\&\left.\frac{\partial^{\omega(\vec{k})}F_N([\epsilon]_{c+1}^{c+N})}{\partial\epsilon_{c+2}^{k_2},..,\partial\epsilon_{c+N}^{k_N}}\right|_{\vec{\epsilon}=0}=F_N^{\vec{k}}([\epsilon]_{c+1}^{c+N})\end{aligned} \tag{24}$$

■

Summing $F_N^{\vec{k}}$ over all $\vec{k}$'s with weight k gives $D_N^{(k)}$:

$$D_N^{(k)}=\frac{1}{k!}\sum_{\vec{k},\omega(\vec{k})=k}F_N^{\vec{k}} \tag{25}$$

We now show that one does not need to sum on all such $\vec{k}$'s, as many of them give zero contribution:

Lemma 3: Let $\vec{k}=[k]_1^N$. If $\exists i<j<N$, with $k_i>k_j=0$, then $F_N^{\vec{k}}=0$.

Proof: Using lemma 1 we get

$$\begin{aligned}F_N^{\vec{k}}&=\left.\frac{\partial^{\omega(\vec{k})}F_N([\epsilon]_1^N)}{\partial\epsilon_1^{k_1},..,\partial\epsilon_N^{k_N}}\right|_{\vec{\epsilon}=0}=\\&\left.\frac{\partial^{\omega(\vec{k})}F_{N-j+1}([\epsilon]_j^N)}{\partial\epsilon_1^{k_1},..,\partial\epsilon_N^{k_N}}\right|_{\vec{\epsilon}=0}=\\&\left.\frac{\partial^{\omega(\vec{k})-1}}{\partial\epsilon_1^{k_1},..,\partial\epsilon_i^{k_i-1},..,\partial\epsilon_N^{k_N}}\left[\frac{\partial F_{N-j+1}([\epsilon]_j^N)}{\partial\epsilon_i}\right]\right|_{\vec{\epsilon}=0}=0\end{aligned} \tag{26}$$

■

We are now ready to prove Thm. 2, 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 non-zero entries) as $l(\vec{k})=N+1-\min_{k_i>0}\{i\}$. It easily follows from lemma 3 that if $F_N^{\vec{k}}\neq 0$, we must have $l(\vec{k})\leq k+1$. Therefore, according to lemma 2:

$$F_N^{\vec{k}}=F_{k+2}^{(k_{N-k-1},..,k_N)} \tag{27}$$

for all $\vec{k}$'s in the sum. Summing on all $F_N^{\vec{k}}$ with the same 'weight', we get $D_N^{(k)}=D_{k+2}^{(k)},\quad\forall N>k+2$. From the analyticity of D_N and D around $\epsilon=0$, one can show by induction that $\lim_{N\to\infty}D_N^{(k)}=D^{(k)}$, therefore we must have $D_N^{(k)}=D^{(k)},\quad\forall N\geq k+2$.

■

4.2 The Almost Memoryless Regime

In the almost memoryless (A-M) regime, we assume that the Markov transition matrices are close to a 'memoryless' matrix. A matrix Q is called memoryless, if all its rows are identical, i.e. $q(i,j)=q(j)$. Thus, a Markov process with a memoryless transition matrix

is in fact an i.i.d. process. Throughout this section we assume that M_η is given by $M_\eta = Q_\eta + \delta T_\eta$, such that Q_η are memoryless matrices, $\delta > 0$ is a small constant and $\sum_{j=1}^{s} t_\eta(i,j) = 0$. Interestingly, in similar to the high-SNR regime, the conditional relative entropy given a finite history gives the correct RE-rate up to a certain order. This is stated in:

Theorem 3: Let $D_N \equiv D_N(\lambda||\mu,\delta)$ be the conditional relative entropy between the probability laws λ and μ on a finite system of length N, where $M_\eta = Q_\eta + \delta T_\eta$, $\eta = \lambda,\mu$ and the Q_η's are memoryless matrices. Let the Taylor expansions of D_N and D around $\delta = 0$ be given by:

$$D_N(\lambda||\mu,\delta) = \sum_{k=0}^{\infty} D_N^{(k)}\delta^k, \;\; D(\lambda||\mu,\delta) = \sum_{k=0}^{\infty} D^{(k)}\delta^k \tag{28}$$

Then we have:

$$N \geq k+2 \Rightarrow D_N^{(k)} = D^{(k)} \tag{29}$$

Proof: The proof of Thm. 3 is very similar to that of Thm. 2. Distinguishing between the sites by setting $M_{\eta,i} = Q_\eta + \delta_i T_\eta$ in site i, we note that setting $\delta_i = 0$ for some i makes the transition matricx $M_{\eta,i}$ memoryless, and thus knowing Y_i 'blocks' the dependence of Y_N on previous Y_j's ($\forall j < i$). The rest of the proof continues in an analogous way to that of Thm. 2 (including the three lemmas therein), and its details are thus omitted here. ■

4.3 Computing the series-coefficients

An immediate application of Thms. 2 and 3 is the computation of the first terms in the series expansion for D, by simply computing these terms for D_N for N large enough. In this section we demonstrate, for the High-SNR regime, the computation of the first order coefficient. If one wishes to compute higher orders, a straightforward way is to compute $D_N^{(k)}$ for $N = k+2$. This can be done by simply enumerating all sequences $[Y]_1^N$, computing the k-th coefficient in $P_\lambda([Y]_1^N)\log\frac{P_\lambda([Y]_1^N)}{P_\mu([Y]_1^N)}$ and summing their contributions. This computation is, however, exponential in k and raises the challenge of designing faster algorithms. For the High-SNR regime we have:

Proposition 2: Let $\eta = \{M_\eta, R_\eta\}$, with $R_\eta = I + \epsilon T_\eta$, $\eta = \lambda,\mu$. Then the RE-rate $D(\lambda||\mu)$ satisfies:

$$D(\lambda||\mu) = \sum_{i,j=1}^{s} \pi_\lambda(i) m_\lambda(i,j)\log\left(\frac{m_\lambda(i,j)}{m_\mu(i,j)}\right) +$$

$$\epsilon \sum_{i,j,k,l=1}^{s} \left\{ \pi_\lambda(i) m_\lambda(i,l) m_\lambda(l,k) \Big[t_\lambda(l,j)\log\left(\frac{\pi_\lambda(i)m_\lambda(i,j)m_\lambda(j,k)}{\pi_\mu(i)m_\mu(i,j)m_\mu(j,k)}\right) - \frac{m_\mu(i,j)m_\mu(j,k)t_\mu(j,l)}{m_\mu(i,l)m_\mu(l,k)}\Big]\right\} + O(\epsilon^2) \tag{30}$$

Proof: According to Thm. 2, $D = D_3 + O(\epsilon^2)$. We thus expand D_3 around $\epsilon = 0$, by substituting $R_\eta = I + \epsilon T_\eta$, $\eta = \lambda,\mu$:

$$D_3(\lambda||\mu) = \sum_{i,j,k}\left\{ P_\lambda([Y]_1^3 = (i,j,k)^t) \log\frac{P_\lambda([Y]_1^3=(i,j,k)^t)P_\mu([Y]_1^2=(i,j)^t)}{P_\mu([Y]_1^3=(i,j,k)^t)P_\lambda([Y]_1^2=(i,j)^t)}\right\} \tag{31}$$

The above probabilities are of the form $P_\eta([Y]_1^N)$, and are given in eq. (1). One can, however, sum in eq. (1) only on vectors $[X]_1^N$ which differ from $[Y]_1^N$ in at most one site, and still get the correct probability up to an $O(\epsilon^2)$ correction. This gives:

$$P_\eta([Y]_1^2 = (i,j)^t) = \pi_\eta(i)m_\eta(i,j) + \epsilon\sum_{k=1}^{s}\Big[\pi_\eta(k)m_\eta(k,j)t_\eta(k,i) + \pi_\eta(i)m_\eta(i,k)t_\eta(k,j)\Big] + O(\epsilon^2) \tag{32}$$

And

$$P_\eta([Y]_1^3 = (i,j,k)^t) = \pi_\eta(i)m_\eta(i,j)m_\eta(j,k) + \epsilon\sum_{l=1}^{s}\Big[\pi_\eta(l)m_\eta(l,j)m_\eta(j,k)t_\eta(l,i) + \pi_\eta(i)m_\eta(i,j)m_\eta(l,k)t_\eta(l,j) + \pi_\eta(i)m_\eta(i,j)m_\eta(j,l)t_\eta(l,k)\Big] + O(\epsilon^2) \tag{33}$$

Substituting eqs. (32, 33) in eq. (31) , and using the Taylor expansion of the logarithm function $\log(a+x) = \log a + \frac{x}{a} + O(x^2)$ gives, after simplification, the result (30). ■

The rigorous result (30) was compared to simulation-based computations of the RE-rate, and good agreement was found for small values of ϵ (results are omitted here, due to lack of space). One can perform a similar expansion of D_3 in δ, to obtain the first order coefficient in the A-M regime. Note that for memoryless matrices Q_η we have $q_\eta(i,j) = \pi_\eta(j)$, thus the RE-rate equals $D(P_\lambda^{(1)}||P_\mu^{(1)})$, and is given by:

$$D(\lambda||\mu) =$$

$$\sum_{i=1}^{s}\left\{\Big[\sum_{j=1}^{s}\pi_\lambda(j)r_\lambda(j,i)\Big]\log\left(\frac{\sum_{j=1}^{s}\pi_\lambda(j)r_\lambda(j,i)}{\sum_{j=1}^{s}\pi_\mu(j)r_\mu(j,i)}\right)\right\} \tag{34}$$

When expanding near a memoryless matrix, one needs to take into account both the perturbations in the Markov matrices M_η, and in the stationary vectors π_η. We note that if one is given two models λ, μ, one can choose to expand around *any* memoryless matrices, and different matrices will naturally give different Taylor coefficients. Two such possibilities are, for example, taking uniform matrices, $q_\eta(i,j) = s^{-1}$, or taking matrices which preserve the singleton distributions of X_N, i.e. $q_\eta(i,j) = \pi_\eta(j)$. Unlike the High-SNR regime, when one wishes to compute of the first order of, say D_3, one needs to sum over all vectors $[X]_1^3$, as they all contribute to the sum.

5 Conclusion

We have obtained new results about the RE-rate between two finite-alphabet *HMM*s. We have represented the relative entropy as a difference of two Lyapunov exponents from random matrix theory. We established the analyticity of the RE-rate, in the interior of the allowable parameters range. We have also shown a connection between the relative entropy of distributions on a finite chain and the RE-rate. This gives a straightforward way to compute the RE-rate as a (one parameter) Taylor series-expansion, in two different regimes. Other parameter regimes still need to be explored. For example, it is interesting to determine the behavior of the RE-rate when one of the models is a small perturbation of the other. In the context of model selection, it is also of interest to compare HMMs of different sizes, where here the RE-rate gives the (average) increase in the likelihood score when adding more parameters. Determining the (maximal) domain of analyticity for the entropy rate is also of considerable interest, as it relates to the radius of convergence of the Taylor series we have obtained.

Acknowledgment

I thank Libi Hertzberg for many helpful discussions and comments on the manuscript. This work was partially supported by the Minerva Foundation and 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 Trans. Inf. Th., 48, pp. 1518-1569, 2002.

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

[3] B.H. Juang and L.R. Rabiner, *A Probabilistic distance measure for HMMs*, AT&T Technical Journal, 64(2), pp. 391-408, 1985.

[4] S. Kullback and R. Leibler, *On information and sufficiency*, Ann. Math. Stat., 22, pp. 79-86, 1951.

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

[6] L. Xie, *Finite horizon robust state estimation for uncertain finite-alphabet hidden Markov models*, Phd thesis, Univ. of New South Wales - Australian Defence Force Academy, 2004

[7] X. Li, V.A. Ugrinovskii and I. R. Petersen, *Probabilistic distances between finite-state finite-alphabet hidden Markov models*, IEEE Trans. Auto. Control, 50(4), pp. 505-511, 2005.

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

[9] T. Holliday, P. Glynn and A. Goldsmith, *On Entropy and Lyapunov Exponents for Finite-State Channels*, Submitted to IEEE Trans. Inf. Th.

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

[11] 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.

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

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

[14] Z. Rached, F. Alajaji and L. L. Campbell, *The Kullback-Leibler Divergence Rate Between Markov Sources*, IEEE Trans. Inf. Th., 50(5), pp. 917-921, 2004.

[15] J. Silva and S. Narayanan, *An Upper Bound for the Kullback-Leibler Divergence for left-to-right Transient Hidden Markov Models*, submitted to IEEE Trans. Inf. Th.

[16] M.N. Do, *Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models*, IEEE Sig. Proc. Letters, 10(4) pp. 115- 118, 2003.

[17] M. Mohammad and W. H. Tranter, *A novel divergence measure for hidden Markov models*, Proc. IEEE Southeast Conference, pp. 240 - 243, 2005.

[18] I. Y. Goldsheid and G. A. Margulis, *Lyapunov indices of a product of random matrices*, Russian Mathematical Surveys, 44, pp. 11-71, 1989.

[19] D. Ruelle, *Analyticity properties of the characteristic exponents of random matrix products*, Adv. Math. 32, pp. 68-80, 1979.

[20] L. Arnold, M. Gundlach and L. Demetrius, *Evolutionary formalism for products of positive random matrices*, Ann. Appl. Prob. 4, pp. 859-901, 1994.

[21] Y. Peres, *Domains of analytic continuation for the top Lyapunov exponent*, Ann. Inst. H. Poincare Probab. Statist. 28(1) pp. 131-148, 1992.

[22] O. Zuk, E. Domany, I. Kanter and M. Aizenman, *Taylor series expansions for the entropy rate of Hidden Markov Processes*, to appear in ICC 2006.