Entropy rate calculations of algebraic measures

Katy Marchand, Jaideep Mulherkar, and Bruno Nachtergaele

Abstract—Let $K = \{0, 1, ..., q - 1\}$. We use a special class of translation invariant measures on $K^\mathbb{Z}$ called algebraic measures to study the entropy rate of a hidden Markov process. Under some irreducibility assumptions of the Markov transition matrix we derive exact formulas for the entropy rate of a general $q$ state hidden Markov process derived from a Markov source corrupted by a specific noise model. We obtain upper bounds on the error when using an approximation to the formulas and numerically compute the entropy rates of two and three state hidden Markov models.

Index Terms—Entropy rate, hidden Markov process, algebraic measures

I. INTRODUCTION

In this paper we study the entropy rate of a hidden Markov process using a class of translation-invariant measures on the chain $K^\mathbb{Z}$, where $K = \{0, ..., q - 1\}$. These measures known as manifestly positive algebraic measures and their properties were first studied in [1]. A one to one correspondence was shown in [1] between manifestly positive algebraic measures and hidden Markov process. We use the properties of the algebraic measures to give formulas for the entropy rate of a hidden Markov process derived from a certain noise model that we will describe later.

In information theory one models an information source as a stochastic process $\{X_i\}_{i=1}^\infty$ with each $X_i$ a random variable taking values in the alphabet set $K = \{0, ..., q - 1\}$. The Shannon entropy of a random variable $X$ taking values in a set $K$ is defined as

$$S(X) = -\sum_{x \in K} p(x) \log_2 p(x)$$

(1)

The entropy of the source for the first $n$ transmitted symbols is given by the joint entropy of $X_1, ..., X_n$

$$S_n(X_1, X_2, ..., X_n) = -\sum_{x_1, ..., x_n} p(x_1, ..., x_n) \log_2 (x_1, ..., x_n)$$

(2)

The entropy rate of the source is defined by

$$H(\mu) = \lim_{n \to \infty} \frac{S_n(X_1, X_2, ..., X_n)}{n}$$

(3)

where $\mu$ is the measure associated with the sequence of random variables. This limit exits when $\{X_n\}$ forms a stationary stochastic process. Entropy rate is an important quantity in information theory as it is a measure of the average amount of information per symbol of a stochastic process. There is a well known formula for the entropy rate of a Markov source. Hidden Markov chains can be thought of as a noisy observation Markov source emitting a sequence of symbols. Hidden Markov models have been extensively studied and the statistical methods based on hidden Markov models have been successfully applied in diverse fields such as speech recognition, image analysis and restoration, DNA sequencing, communication and information theory. Even though extensive research has been carried out on hidden Markov models [2], [3], the problem of deriving explicit expressions for the entropy rate of the in terms of the parameters of the process is still an open issue.

The entropy rate of a hidden Markov process was first studied by Blackwell in 1957 [4] who showed that the entropy rate is given by the integral

$$H(\mu) = \int_{\mathcal{W}} S(w) \phi(dw)$$

(4)

where $S$ is the Shannon entropy and $w$ belongs to the simplex $\mathcal{W} = \{w = (w_1, w_2, ..., w_q) : w_i \geq 0, \sum w_i = 1\}$ is a $\phi(dw)$ is a particular measure called the Blackwell measure on the simplex $\mathcal{W}$. Recently there has been a renewed interest in problem of calculating the entropy rate of a hidden Markov chain. The papers [5], [6] showed a connection between the entropy rate and top Lyapunov exponent of a product of random matrices. The study of entropy rate in the context of filtering and denoising was done in [7], [8]. The study of the asymptotic behavior, the smoothness and analytic properties and obtaining new bound and improved bounds of the entropy rate in terms of the process parameters is carried out in [9], [10], [11], [12]. Calculation of entropy rate based on ideas from statistical mechanics is done in [13]. In this paper we follow the approach of [1], wherein a formula similar to (4) was derived. Moreover, in [1] the support of the Blackwell measure was explicitly characterized. We see that in the case of the particular noise model that we study in this paper the support of the measure $\phi(dw)$ simplifies and leads to an analytic solution of the entropy rate.

In section II-A we review some key results on algebraic measures from [1] that we will use in this paper. The description of the noise model and the support of the measure $\phi(dw)$ is computed in in section III. The main theorem about the formula for the entropy rate of the hidden Markov process is proved in IV. Lastly in section V we show numerical computations of the entropy rate using approximations to the formulas obtained in section IV.

II. SETUP

A. Manifestly positive algebraic measures

Let $q \in \mathbb{N}_0$ and $K = \{0, 1, ..., q - 1\}$ and consider the chain $\Omega = K^\mathbb{Z}$ consisting of configurations $\omega = (\omega_1, \omega_2, ...)$. 

Katy Marchand is currently unaffiliated.
Jaideep Mulherkar is with DA institute of information and communication technology, Gandhinagar, India. email: jaideep_mulherkar@daiict.ac.in.
Bruno Nachtergaele is with the Department of Mathematics University of California, Davis USA e-mail: bxn@math.ucdavis.edu.
with each \( w_i \in K \). Let \( F_K \) be the sigma algebra generated by all the cylinder sets. Let \( T : K^Z \to K^Z \) be the shift transformation given by
\[
T\omega = \delta \quad \text{where} \quad \delta_n = \omega_{n+1}
\]
A measure \( \mu \) on \( \Omega \) is called translation invariant if
\[
\mu(E) = \mu(T^{-1}E) \quad \forall E \in F_K
\]
In [1] a special class of translation-invariant measures called algebraic measures was constructed on the chain \( \Omega \) in terms of a triplet \((U, \rho, (E_\alpha)_{\alpha \in K})\) where
- \( U \) is a real ordered algebra (the order determined by a convex cone \( \mathbb{U}^+ \)),
- \( \rho \) is a positive linear functional on \( U \),
- \( E_\alpha U^+ \) satisfying \( \rho(AE) = \rho(AE) + \rho(A) \forall A \in U, \rho(E) = 1 \) where \( E = \sum_{\alpha \in K} E_\alpha \).

We state the following proposition from [1]

**Proposition II.1:** Given the triplet \((U, \rho, (E_\alpha)_{\alpha \in K})\) there exists a unique translation invariant probability measure \( \mu \) on \( K^Z \) such that
\[
\mu(\omega_{m,n}) = \rho(E_{w_m}, E_{w_{m+1}}, ..., E_{w_n})
\]
where \( \omega_{m,n} = (\omega_m, \omega_{m+1}, ..., \omega_n) \).

The measure \( \mu \) was called an algebraic measure and it was shown that if \( U \) is finite dimensional the triplet \((U, \rho, (E_\alpha)_{\alpha \in K})\) has a unique representation in a finite dimensional vector space. Let \( (\mathbb{R}^d)^+ \) denote the positive cone in \( \mathbb{R}^d \) of vectors with all its components non-negative and let \((M_d)^+\) be the matrices that preserve this positive cone i.e. the all the elements of are non-negative. An algebraic measure \( \mu \) is called manifestly positive if there exists a \( d \in \mathbb{N} \) positive \( \tau, \sigma \in (\mathbb{R}^d)^+ \) and for each \( \alpha \in K \), a positive \( E_\alpha \in M_d^+ \) such that
\[
\mu((\omega_m, \omega_{m+1}, ..., \omega_n)) = \langle \tau | E_{w_m} E_{w_{m+1}} ... E_{w_n} \sigma \rangle
\]

Two important examples of manifestly positive algebraic measures are Markov chains and hidden Markov models.

1) **Markov chains:** Let \( \{X_i\}_{i \in \mathbb{N}} \) be a Markov chain taking values in \( K = \{0, ..., q-1\} \) and having stationary measure \( \mu \). Let
- \( \sigma \in (\mathbb{R}^q)^+ \) is the vector with all components equal to 1.
- \( \tau \in (\mathbb{R}^q)^+ \) such that \( \tau_\alpha = \mu((a)) \) i.e. the \( a^{th} \) component of the stationary distribution \( \tau \).
- \( E_\alpha \in M_d^+ \) : \( E_\alpha b,c = \delta_{a,b} \mu((a)) \) for all \( a,b,c \in K \) is the matrix with the only non-vanishing row to be the \( a^{th} \) row of the transition matrix \( E = \sum_{\alpha \in K} E_\alpha \).

Then it is easy to see that
\[
\mu((w_m, ..., w_n)) = \langle \tau | E_{w_m} ... E_{w_n} 1 \rangle
\]
\[
E^+ \tau = \tau
\]
\[
E \sigma = \sigma
\]

2) **Hidden Markov models:** In [1] a one-one correspondence was shown between functions of Markov processes (hidden Markov models) and the class of manifestly positive algebraic measures. Let \( X = \{X_i\}_{i \in \mathbb{N}} \) be a Markov chain that takes values on a finite alphabet \( L \) with transition matrix \( E \) stationary measure \( \nu \). Let \( F_a \) be the matrix with the only non-vanishing row equal to the \( a^{th} \) of \( E \). Thus
\[
\sum_{a \in L} F_a = E
\]

We can represent the measure \( \nu \) as in example 1 by the triplet \((\tau, 1, (F_a)_{a \in L})\). Let \( Y = \{Y_i\}_{i \in \mathbb{N}} \) be a noisy observation of the Markov chain with values in \( K = \{0, 1, ..., q-1\} \). Define the matrix \( \bar{R} = [r_{ab}] \) with \( r_{ab} = Pr[Y_i = a|X_i = b] \) and let
\[
E_a = \sum_{b \in L} r_{ab} F_b
\]

It is easy to verify that the manifestly positive representation of the stationary measure \( \mu \) associated with \( Y \) is \((\tau, \sigma, (E_\alpha)_{\alpha \in K})\). So that
\[
\mu((w_m, ..., w_n)) = \langle \tau | E_{w_m} \ldots E_{w_n} \sigma \rangle
\]
where \( \tau_\alpha = \mu(a) \) and \( \sigma \) is a vector with all components equal to 1.

**B. Entropy rate of manifestly positive algebraic measures**

It was shown that under certain irreducibility conditions (see Condition 1 given below) the mean entropy rate of a manifestly positive algebraic measure can be computed as an integral with respect to a measure on the simplex \( W = \{w = (w_1, w_2, ..., w_q) : w_i \geq 0, \sum w_i = 1\} \) similar to the Blackwell measure.

**Condition 1:**

i. There exists a \( c > 0 \) such that for all \( a, b \in K \) \( E_a E_b \geq cE_{a,b} \).
ii. There exists an \( a_0 \in K \) such that the invariant subspace corresponding to the largest eigenvalue of \( E_{a_0} \) is one dimensional.
iii. \( E \) is irreducible i.e. the invariant subspace corresponding to the largest eigenvalue 1 of \( E \) is one dimensional.

We have the following theorem for the entropy rate of a function of Markov process.

**Theorem II.2 (12):** The mean entropy rate \( H(\mu) \) of a manifestly positive algebraic measure \( \mu \) (satisfying Condition 1) is given by
\[
H(\mu) = \sum_{a \in K} \int_W h_a(w) \phi(dw)
\]
where \( h_a(w) = -\langle w | E_a \sigma \rangle \log \langle w | E_a \sigma \rangle \) and \( \phi(dw) \) is a probability measure on the simplex \( W \).

In [1] an equation for \( \phi(dw) \) is derived in terms of a Markov operator \( T_\mu \) on \( C(W) \), the space of continuous functions on the simplex \( W \). In addition the support of the measure is also characterized. For functions of Markov processes Blackwell [4] obtained a formula similar to equation (8) however there
was no clear connection of the measure with the Markov operator and the support of the measure was also not explicitly characterized. The Markov operator \( T_\mu \) can be described in terms of the collection \( \{ E_\alpha | \alpha \in K \} \) as follows. We first define \( \Gamma_\alpha : \mathcal{W}_0 \to \mathcal{W}_0 \) with \( \mathcal{W}_0 = \mathbb{W} \cup \{ 0 \} \) by

\[
\Gamma_\alpha(\nu) = \begin{cases} 
\frac{E_\nu}{(\nu E_\alpha | \sigma)} & \text{if } (\nu E_\alpha | \sigma) \neq 0 \\
0 & \text{otherwise}
\end{cases}
\]  

(9)

Let \( \mathcal{C}_0(\mathcal{W}_0) \) be the set of continuous real valued functions that vanish on 0, then \( T_\mu : \mathcal{C}_0(\mathcal{W}_0) \to \mathcal{C}_0(\mathcal{W}_0) \) is defined by

\[
(T_\mu f)(w) = \sum_{\alpha \in K}(w|E_\alpha \sigma)f(\Gamma_\alpha w)
\]  

(10)

By the Riesz-Markov theorem (Theorem IV. 14 [14]) there is a one to one correspondence between the space of measures and normalized positive linear functionals on \( \mathcal{C}(\mathcal{W}) \). Therefore

\[
\phi(f) = \int_W f(w)\phi(dw) \quad f \in \mathcal{C}(\mathcal{W})
\]  

(11)

In [1] the measure \( \phi(dw) \) was characterized as the unique measure on \( \mathcal{C}(\mathcal{W}) \) that is invariant under \( T_\mu \), i.e.,

\[
\phi(f) = \phi(T_\mu f) \quad f \in \mathcal{C}(\mathcal{W})
\]  

(12)

The support of \( \phi(dw) \) is given by

\[
\text{supp}(\phi) = \Delta \equiv \{ \Gamma_\omega \omega | \omega \in K^n, n \in \mathbb{N} \}
\]  

(13)

where \( \Gamma_\omega = \Gamma_{\omega_n-1} \cdots \Gamma_{\omega_0} \) and \( \hat{\tau} \) is the only non-trivial fixed point of \( \Gamma_{\omega_0} \).

III. NOISE MODEL AND SUPPORT OF THE MEASURE \( \phi \)

The entropy rate formulas derived in this paper are for a general \( q \) state hidden Markov process described by a particular noise model. In this section we describe the noise model and the support of the measure \( \phi(dw) \) given by equation (13) for this noise model. The noise model that we work assumes that noise does not affect exactly one of the input symbols, say the symbol 0. If the symbol 0 is transmitted then it is always unambiguously received at the other end. On the other hand if any of the other symbol is transmitted then it is either received without any error or received as the symbol 0 with a small error probability. That is \( P(Y_i = 0) = 0 = 1, P(Y_i = 0|X_i = a) = \epsilon_a \) and \( P(Y_i = b|X_i = a) = 1 - \epsilon_a \) for \( a = 1, ..., q - 1 \) and \( P(Y_i = b|X_i = a) = 0 \) when \( 0 \neq b \neq a \). See figure 1 for a description of the model in the special case of \( q = 3 \). In this paper without loss of generality we shall assume that the unambiguous symbol is 0. Let the matrices \( \{ F_\alpha \} \) be the matrices that describe the uncorrupted Markov source as in equation (5). For this noise model we write the matrices \( \{ E_\alpha \} \) given by equation (6) as

\[
E_0 = F_0 + \sum_{a=1}^{a-1} \epsilon_a F_a
\]  

(14)

\[
E_a = (1 - \epsilon_a)F_a \quad \text{for } a = 1, ..., q - 1
\]

\[
\sum_{a \in K} E_a = E
\]

Let \( e_i \) denote the transpose of the \((i-1)^{st}\) row of \( E = [e_{ij}] \). Let \( p = \min_{ij} e_{ij} \) and \( P = \max_{ij} e_{ij} \). Our only assumption will be

**Assumption 1:**

\[ 0 < p \leq P < 1 \quad \epsilon_0 = 1 \quad \epsilon_a > 0 \quad \forall a \in \{1, ..., q - 1\}. \]

From equation (9) one can see that for each \( \nu \in \mathcal{W} \)

\[
\Gamma_0 \nu = \frac{\sum_{a=0}^{q-1}\epsilon_a e_{\alpha a}e_a}{\sum_{a=0}^{q-1}\epsilon_a e_{\alpha a}} = \sum_{a=0}^{q-1} \alpha_a e_a
\]  

(15)

Because of Assumption 1 we get

\[
p \leq (\Gamma_0 \nu) \leq P \quad \forall \alpha \in K.
\]  

(16)

Also one gets from equation (9) for all \( a \in \{1, ..., q - 1\} \).

\[
\Gamma_a \nu = \begin{cases} 
\epsilon_a & \text{if } \nu_a \neq 0 \\
0 & \text{otherwise}
\end{cases}
\]  

(17)

A key factor that simplifies our analysis of the entropy rate is that the support of \( \phi(dw) \) given by equation (13) simplifies significantly with this noise model. We have the following proposition about the support of the measure \( \phi \).

**Proposition III.1:** The support of the measure \( \phi \) is given by

\[
\Delta = \{ \Gamma_\omega \omega | \omega \in K^n, n \in \mathbb{N} \}
\]

(18)

**Proof:** We know from equation (13) that the support of the measure \( \phi \) is given by

\[
\Delta = \{ \hat{\tau} \} \cup \{ \Gamma_m e_j | j \in \{1, ..., q - 1\}; m \in \mathbb{N}_0 \}
\]

(19)

One can show (1) equation (23) that for any \( \nu \in \mathcal{W}^+ \) there exists constants \( C_1 \) and \( P \) such that

\[
|\Gamma_0^{P+1} \nu - \hat{\tau}|_1 \leq C_1 \rho^k
\]

This implies that all subsequences in \( \Delta \) converge to \( \hat{\tau} \) and hence \( \hat{\tau} \) is the only accumulation point in \( \Delta \). Therefore

\[
\Delta = \{ \Gamma_m e_j | j \in \{1, ..., q - 1\}; m \in \mathbb{N}_0 \}
\]
Moreover for any \( j \in \{1, \ldots, q - 1\} \)
\[
\lim_{m \to \infty} \Gamma_0^m e_j = \tilde{\tau} \tag{20}
\]

We note that \( \Delta \) may consist finite or infinitely many distinct points in \( \mathcal{W} \). For instance in the case if \( e_0 = e_1 = \cdots = e_{q-1} \) then it is easy to see that \( \tilde{\tau} = e_0 = \cdots = e_{q-1} \) and \( \Delta \) consists of a single element i.e. \( \Delta = \{e_0\} \). It is clear from \( \text{[19]} \) that if \( \Delta \) is a finite set then \( \tilde{\tau} = \Gamma_0^m e_j \) for some \( n \in \mathbb{N} \) and some \( j \in \{1, \ldots, q - 1\} \). On the other hand \( \Delta \) can be countably infinite as can be seen from the following lemma.

**Lemma III.2:** If \( E_0^* \) is one to one then \( \Delta \) is countably infinite set of distinct elements.

**Proof:** We will first show that \( E_0^* \) is one-one \( \implies \Gamma_0 : \mathcal{W} \to \mathcal{W} \) is one-one.

If \( \Gamma_0 \nu = \Gamma_0 \eta \) for \( \nu, \eta \in \mathcal{W} \) then
\[
\frac{E_0^* \nu}{\langle \nu | E_0 \sigma \rangle} = \frac{E_0^* \eta}{\langle \eta | E_0 \sigma \rangle}
\]
Therefore \( E_0^* (\nu - C \eta) = 0 \) where \( C = \frac{\langle \nu | E_0 \sigma \rangle}{\langle \eta | E_0 \sigma \rangle} \). But \( E_0^* \) is one-one implies
\[
\nu = C \eta
\]
and since \( \nu, \eta \in \mathcal{W} \) therefore \( C = 1 \) so \( \nu = \eta \). Since \( E_0^* \) is one-the vectors \( e_a \in \mathcal{W}, a \in K \) form a linearly independent set. We will show that all the elements of \( \Delta \) are distinct i.e. for any \( m, n \geq 0 \)
\[
\Gamma_0^m e_i \neq \Gamma_0^n e_j \tag{21}
\]
Assume \( \Gamma_0^m e_i = \Gamma_0^n e_j \) for some \( m, n \geq 0 \) and for \( i, j \in \{1, \ldots, q - 1\} \). If \( m \) and \( n \) are both zero then \( e_i = e_j \) but this contradicts the fact that \( e_j \)'s form a linearly independent set. Assume wlog that \( m > n \). Since \( \Gamma_0 \) is one-one we arrive at \( \Gamma_0^k e_i = e_j \) for \( k = m - n \). Again using equation \( \text{[15]} \) we arrive at a contradiction that \( e_j \)'s form a linearly independent set.

From above, \( \Gamma_0 e_a \neq e_a \) for any \( a \in K \) so \( e_a \neq \tilde{\tau} \). If \( \Gamma_0^m e_a = \tilde{\tau} \) for some \( m \geq 1 \) then \( \Gamma_0(\Gamma_0^{m-1} e_a) = \tilde{\tau} \). Since \( \tilde{\tau} \) is a fixed point of \( \Gamma_0 \) and \( \Gamma_0 \) is one to one we conclude that \( (\Gamma_0^{m-1} e_a) = \tilde{\tau} \). Repeating the argument we get \( e_a = \tilde{\tau} \) which again is a contradiction.

**Lemma III.3:** If \( \Delta \) is countably infinite then \( \phi(\tilde{\tau}) = 0 \)

**Proof:** Since \( \Delta \) is countably infinite there exists a \( j \in \{1, \ldots, q - 1\} \) such that \( \lim_{m \to \infty} \Gamma_0^m e_j = \tilde{\tau} \). Consider the set
\[
\Delta_{j,m} = \{ \Gamma_0^k e_j | k \geq m \}
\]
for \( m \in \{1, 2, \ldots\} \). This is a decreasing sequence of sets with
\[
\bigcap_{m \in \mathbb{N}} \Delta_{j,m} = \{ \tilde{\tau} \}
\]
Therefore \( \phi(\tilde{\tau}) = \lim_{m \to \infty} \phi(\Delta_{j,m}) \)

Let \( f_m \in \mathcal{C}(\mathcal{W}) \) be defined so that
\[
f_m(\nu) = \begin{cases} 1 & \nu \in \Delta_{j,m} \\ 0 & \text{otherwise} \end{cases}
\]
So,
\[
\phi(\Delta_{j,m}) = \int_{\Delta} f_m(\nu) d\phi
\]
By \( \text{[10]} \) and \( \text{[12]} \) we get
\[
\phi(\Delta_{j,m}) = \int_{\Delta_{m-1}} \langle \nu | E_0 \sigma \rangle f_m(\Gamma_0 \nu) d\phi
\]
Therefore we get
\[
\phi(\Delta_{j,m}) = \int_{\Delta_{m-1}} \langle \nu | E_0 \sigma \rangle d\phi
\]
where we used \( \text{[17]} \) and the fact that \( \Gamma_0 \nu \in \Delta_m \) only if \( \nu \in \Delta_{m-1} \). We have
\[
\phi(\Delta_{j,m}) = \int_{\Delta_{m-1}} \langle \nu | E_0 \sigma \rangle d\phi
\]
\[
= \int_{\Delta_{m-1}} \sum_{a=0}^{q-1} e_a \nu_a d\phi
\]
Since \( \sum_a \nu_a = 1, e_0 = 1 \) and \( e_a < 1 \) for \( a \neq 0 \)
\[
r = \sum_{a=0}^{q-1} e_a \nu_a < 1.
\]
So,
\[
\phi(\Delta_{j,m}) \leq r \phi(\Delta_{j,m-1}) \implies \phi(\Delta_{j,m}) \leq r^{m-1} \phi(\Delta_1) = 0.
\]

The next lemma shows that our Assumption 1 on the Markov transition matrix and the noise parameters is enough to satisfy Condition 1.

**Lemma III.4:** If \( 0 < p \leq P < 1 \) and \( e_a > 0 \) for all \( a \in K \) then the matrices \( (E_a)_{a \in K} \) satisfy Condition 1.

**Proof:** Condition 1 i. can be verified by a simple computation by choosing \( c = c_0^{P-1} \) where \( c = \min_i \epsilon_i \). ii. and iii. follow from the Perron-Frobenius theorem and the fact that \( E \) is a Markov transition matrix.

IV. ENTROPY RATE FORMULAS FOR A HIDDEN MARKOV PROCESS

In this section we apply the results on manifestly positive algebraic measures in \( [\text{[1]} \] \) to the case of the hidden Markov model described in section III. We divide our support set \( \Delta \) into disjoint sets in the following way
\[
\Delta_1 = \{ \Gamma_0^m e_1 : m \in \mathbb{N}_0 \}
\]
\[
\Delta_2 = \{ \Gamma_0^m e_2 : m \in \mathbb{N}_0 \} \setminus \{ \Delta_1 \}
\]
\[
\Delta_3 = \{ \Gamma_0^m e_3 : m \in \mathbb{N}_0 \} \setminus (\Delta_1 \cup \Delta_2)
\]
\[
\vdots
\]
\[
\Delta_{q-1} = \{ \Gamma_0^m e_{q-1} : m \in \mathbb{N}_0 \} \setminus \bigcup_{i=1}^{q-2} \Delta_i
\]
Define
\[ c_{j,m} = \prod_{i=1}^{m} (\Gamma_{0}^{m-i} e_i | E_0 \sigma) \] (22)

Let \( A \) be the \( q \times q - 1 \) matrix defined by entries. If \( i \neq q \) then
\[ A_{ij} = \begin{cases} -\delta_{ij} + \sum_{m=0}^{q-1} |\Gamma_{0}^{m} e_i| E_i \sigma c_{j,m} & \text{if } q \neq 2 \\ 0 & \text{if } q = 2 \end{cases} \] (23)

\[ A_{qj} = \sum_{m=0}^{q-1} c_{j,m} \] (24)
\[ \Phi = [\phi(e_1), ..., \phi(e_{q-1})]' \in \mathbb{R}^{q-1} \] (25)
\[ b = [0, 0, ..., 1]' \in \mathbb{R}^{q} \]

**Theorem IV.1:** Under Assumption 1 the entropy rate of the measure \( \mu \) associated with the hidden Markov process with the noise model described in section III is given by
\[ H(\mu) = \sum_{j=1}^{q-1} \sum_{m=0}^{q-1} h_a(\Gamma_{0}^{m} e_j) c_{j,m} \Phi_j \] (26)

where \( \Phi_j \) is the \( j \)-th coordinate of the solution to the matrix equation \( A \Phi = b \).

**Proof:** Let
\[ f_{j,m}(\nu) = \begin{cases} 1 & \text{if } \nu = \Gamma_{0}^{m} e_j \\ 0 & \text{otherwise} \end{cases} \]
for \( m \in \mathbb{N} \) and \( j = 1, ..., q - 1 \). We have
\[ \phi(\Gamma_{0}^{m} e_j) = \int_{W} f_{j,m}(\nu) d\phi \]

Using (22) we get
\[ \phi(e_i) = \sum_{j=1}^{q-1} \sum_{m=0}^{\infty} \phi(e_j) \sum_{m=0}^{q-1} (\Gamma_{0}^{m} e_j | E_i \sigma) c_{j,m} \] (28)

Now the measure of the whole set \( \phi(\Delta) = 1 \) so
\[ \sum_{j=1}^{q-1} \sum_{m=0}^{\infty} \phi(\Gamma_{0}^{m} e_j) = \sum_{j=1}^{q-1} \phi(e_j) \sum_{m=0}^{q-1} c_{j,m} = 1 \] (29)

Equations (28) and (29) form a system of \( q \) linear equations \( A \Phi = b \) with \( A \in \mathbb{M}_{q,q-1}, \Phi \in \mathbb{R}^{q-1} \) and \( b \in \mathbb{R}^{q} \) as given by (23), (24), (25). By lemma IV.2 \( \sum_{m=0}^{q-1} c_{j,m} < \infty \) which ensures that each \( A_{ij} \) is bounded. Now from the integral formula for the entropy rate given by (8) and the support of the measure given by proposition (III.1) and we get that
\[ H(\mu) = \sum_{j=1}^{q-1} \sum_{m=0}^{q-1} h_a(\Gamma_{0}^{m} e_j) \phi(\Gamma_{0}^{m} e_j) \]

From (27) we get
\[ H(\mu) = \sum_{j=1}^{q-1} \sum_{m=0}^{\infty} h_a(\Gamma_{0}^{m} e_j) c_{j,m} \Phi_j \]

**Lemma IV.2:** For all \( j = 1, ..., q - 1 \) and \( m \in \mathbb{N}_0 \)
\[ c_{j,m} < \gamma^m \text{ where } \gamma = \max_{j,k} \sum_{a=0}^{q-1} e_a(\Gamma_{0}^{k} e_j)_a < 1 \] (30)

**Proof:**
\[ c_{j,m} = \prod_{i=1}^{m} (\Gamma_{0}^{n-i} e_j | E_0 \sigma) \] (31)
\[ = \prod_{i=1}^{m} \sum_{a=0}^{q-1} e_a(\Gamma_{0}^{k} e_j)_a \]

We know by (16) \( p \leq \Gamma_{0}^{k} e_j | a \leq P \) for each \( a \in K \). Since \( \Gamma_{0}^{k} e_j \in W, \sum_{a=0}^{K} (\Gamma_{0}^{k} e_j)_a = 1 \). Now, by Assumption 1. \( e_a < 1 \) if \( a = \{1, ..., q - 1\} \) and \( e_0 = 1 \).

So \( \sum_{a=0}^{q-1} e_a(\Gamma_{0}^{k} e_j)_a < 1 \forall k \in \mathbb{N}_0 \)

Moreover by (20)
\[ \lim_{k \to \infty} \Gamma_{0}^{k} e_j = \tau \in W^+ \]
Therefore,
\[ \sup_k \sum_{a=0}^{q-1} \epsilon_a \Gamma_0^k e_j | e_j | a < 1 \]
and
\[ \gamma = \max_j \sup_k \sum_{a=0}^{q-1} \epsilon_a \Gamma_0^k e_j | e_j | a < 1 \]

We also prove the following bound for \( c_{im} \) which will be useful in the numerical estimates in section V.

**Lemma V.3:** Let \( \epsilon = \max_{a \in \{1, \ldots, q-1\}} \epsilon_a \). If \( \epsilon < p \) then
\[ c_{j,m} < r^m \text{ where } \]
\[ r = 1 - (q-1)(p - \epsilon P) < 1 \]

**Proof:** \( \epsilon < p \implies r = (1 - (q - 1)p + (q - 1)\epsilon P) < 1 \).

\[ c_{j,m} = \prod_{i=1}^m (\Gamma_0^{m-i} e_j | e_j | 0 + \sum_{a=1}^{q-1} \epsilon_a \Gamma_0^{m-a} e_j | e_j | a) \leq \prod_{i=1}^m (1 - (q-1)p + (q-1)\epsilon P) = r^m \]

V. APPROXIMATING THE ENTROPY RATE

In this section we present some numerics approximations to the entropy rate formulas that were derived in section IV. We show that if we take only the first \( N \) terms of the matrix \( A \) given by (23) and (24) for the entropy rate calculations then this gives an approximation of order \( O(\gamma^{N+1}) \) where is given by (30).

Let
\[ A = \hat{A} + R \]
where the entries of \( R \) are the \( N^{th} \) tails of the entries of \( A \).

Let \( \hat{\Phi} \) be the least square solution to
\[ \hat{A} \Phi = b \]

Therefore
\[ \hat{\Phi} = \hat{A}^\dagger b \]

where \( \hat{A}^\dagger \) is the pseudo-inverse of \( A \). We first prove the following lemma

**Lemma VI.1:**
\[ \| \Phi - \hat{\Phi} \|_1 \leq q \| \hat{A}^\dagger \|_1 \frac{1}{1 - \gamma} \]

**Proof:** From (33) and (34)
\[ A \Phi = \hat{A} \Phi + R \Phi = b \]

Substituting (35)
\[ \hat{A} \Phi - \hat{A} \Phi = - R \Phi - \hat{A} \Phi \]
\[ \hat{A} (\Phi - \hat{\Phi}) = b - R \Phi - \hat{A} \Phi \]
\[ \| \Phi - \hat{\Phi} \|_1 = \| \hat{A}^\dagger b - \hat{A}^\dagger R \Phi - \hat{A} \Phi \| \]
\[ \| \Phi - \hat{\Phi} \|_1 \leq \| \hat{A}^\dagger \|_1 \| R \Phi \|_1 \]

But since \( R \) consists of the tail of the entries each entry in \( R \) is bounded by \( \sum_{m=N+1}^\infty \gamma^m = \frac{\gamma^{N+1}}{1 - \gamma} \).

\[ \| R \Phi \|_1 \leq \frac{\gamma^{N+1}}{1 - \gamma} \left( \begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \right) \leq \frac{\gamma^{N+1} q}{1 - \gamma} \]

\[ \| \Phi - \hat{\Phi} \|_1 \leq \| \hat{A}^\dagger \|_1 \| R \Phi \|_1 \leq \frac{q \| \hat{A}^\dagger \|_1}{1 - \gamma} \gamma^{N+1} \]

**Theorem V.2:** Under Assumption 1 the entropy rate of the measure \( \mu \) associated with the hidden Markov process described in section III can be approximated to \( O(\gamma^{N+1}) \) by
\[ H_N(\mu) = \sum_{j=1}^{q-1} \sum_{m=0}^\infty h_a(\Gamma_0^m e_j) c_{j,m} \hat{\Phi}_j \]

**Proof:**
\[ \text{err}(N) = | H(\mu) - H_N(\mu) | \]

\[ \text{err}(N) = \left| \sum_{j=1}^{q-1} \sum_{m=0}^\infty \sum_{a=0}^\infty h_a(\Gamma_0^m e_j) c_{j,m} \hat{\Phi}_j + \sum_{j=1}^{q-1} \sum_{m=0}^\infty \sum_{a=0}^\infty h_a(\Gamma_0^m e_j) c_{j,m} (\Phi_j - \hat{\Phi}_j) \right| \]
\[ \leq q \sum_{j=1}^{q-1} \sum_{m=0}^\infty \gamma^m | \Phi_j + q \sum_{j=1}^{q-1} \sum_{m=0}^\infty \gamma^m | \Phi_j - \hat{\Phi}_j | \]
\[ = q \gamma^{N+1} \frac{1}{1 - \gamma} | \Phi_j | + q \gamma^{N+1} \frac{1}{1 - \gamma} | \Phi - \hat{\Phi} |_1 \]

By lemma VI.1
\[ \leq q \gamma^{N+1} \frac{1}{1 - \gamma} + q \gamma^{N+1} \frac{1}{1 - \gamma} \left( q \| \hat{A}^\dagger \|_1 \right) \]

Therefore
\[ \text{err}(N) \leq q \gamma^{N+1} \frac{1}{1 - \gamma} (1 + q \| \hat{A}^\dagger \|_1) \leq O(\gamma^{N+1}) \]

Next, we present a numerical example for approximating the entropy rate formulas given by theorem V.2 We can get estimates on \( \text{err}(N) \) using the bound (36), but \( \gamma \) is difficult to compute. However, note that lemma IV.3 implies that whenever \( \epsilon < p \) the bound
\[ \text{err}(N) \leq q \gamma^{N+1} \frac{1}{1 - \gamma} \left( 1 + q \| \hat{A}^\dagger \|_1 \right) \]
is also obtained where \( r \) is given by (32). In example that follows, in order to estimate \( \text{err}(N) \) we work with an additional assumption that \( \epsilon < p \).

1) **Example 1:** In the example we let \( q = 2 \), \( \epsilon = 0.01 \). The transition matrix we use is:

\[
E = \begin{pmatrix}
0.85 & 0.15 \\
0.28 & 0.72
\end{pmatrix}
\]

Table I shows the estimate entropy rate \( H_N(\mu) \) using the formula given by theorem V.2. Figure 2 shows the support of the measure \( \hat{\phi} \). The value of \( \epsilon \) is chosen to be 0.2 so as to make the plot clearly visible.

2) **Example 2:** In the example we let \( q = 3 \), \( \epsilon_1 = 0.01 \) and \( \epsilon_2 = 0.02 \). The transition matrix we use is:

\[
E = \begin{pmatrix}
0.4 & 0.25 & 0.35 \\
0.25 & 0.45 & 0.3 \\
0.2 & 0.55 & 0.25
\end{pmatrix}
\]

Table II shows the estimate entropy rate \( H_N(\mu) \) using the formula given by theorem V.2. Figure 3 show a plot of the entropy rate \( H_N(\mu) \) versus \( N \). It is seen that the formulas for \( H_N(\mu) \) converge very quickly with \( N \). Figure 4 shows the plot of the bound on \( \text{err}(N) \) given by equation (37) versus \( N \). It can be seen that we get a very good bound on \( \text{err}(N) \) within a few terms.

### Table I

| \( N \) | \( H_N(\mu) \) | \( \text{err}(N) \) bound |
|---|---|---|
| 10 | 0.7139684740464 | 0.0566 |
| 20 | 0.7027846315804 | 0.3408 |
| 30 | 0.7088402087809 | 0.7408 |
| 40 | 0.704584593354 | 0.1011 |
| 50 | 0.7038443295765 | 0.0350 |
| 60 | 0.7003697902382 | 0.0076 |
| 70 | 0.7006689107994 | 0.0016 |
| 80 | 0.7003661007983 | 3.0025 \times 10^{-6} |
| 90 | 0.7003662028038 | 7.837 \times 10^{-9} |
| 100 | 0.7003661807786 | 1.7044 \times 10^{-9} |

### Table II

| \( N \) | \( H_N(\mu) \) | \( \text{err}(N) \) bound |
|---|---|---|
| 10 | 0.9596105211351 | 0.3561 |
| 20 | 0.9596105211352 | 0.0030 |
| 30 | 0.9596112616404 | 2.676 \times 10^{-7} |
| 40 | 0.9596112616404 | 2.3103 \times 10^{-9} |
| 50 | 0.9596112616404 | 2.0103 \times 10^{-9} |

Fig. 3. Plot of Entropy rate \( H_N(\mu) \) computed using new formulas versus \( N \). The estimate converges very quickly with \( N \).

Fig. 4. Plot of the error \( \text{err}(N) \) versus \( N \). A reasonable bound on the error is obtained with fewer than 50 terms of the sum.

### References

[1] M. Fannes, B. Nachtergaele, and L. Slegers, “Functions of Markov processes and algebraic measure,” *Reviews in Mathematical Physics*, vol. 4, p. 39, 1992.

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

[3] Y. Ephraim and M. N, “Hidden Markov processes,” *IEEE Trans. Inf. Theory*, vol. 48, pp. 1518–1569, 2002.

[4] D. Blackwell, “The entropy of functions of finite-state Markov chains,” *Trans. 1st Prague Conf. Information Theory, Statistical Decision Functions, Random Processes*, pp. 13–20, 1957.

[5] P. Jacquet, G. Serroussi, and W. Szpankowski, “On the entropy of a hidden Markov process,” *Proceedings of Data Compression Conference, Snowbird, UT*, pp. 362–371, 2004.
[6] T. Holliday, A. Goldsmith, and P. Glynn, “On entropy and lyapunov exponents for finite state channels,” IEEE transactions on information theory, vol. 52, 2006.

[7] C. Nair, E. Ordentlich, and T. Weissman, “Proceedings of isit adelaide,” IEEE Transactions on Information Theory, 2005.

[8] E. Ordentlich and T. Weissman, “On the optimality of symbol by symbol filtering and denoising,” IEEE Transactions on Information Theory, vol. 52, pp. 19–40, 2006.

[9] ——, “New bounds on the entropy rate of a hidden Markov process,” Proc. San Antonio Inf. Theory Workshop, 2004.

[10] O. Zuk, E. Domany, I. Kanter, and M. Aizenmann, “From finite-system entropy to entropy rate of a hidden Markov process,” IEEE Signal Processing Letters, vol. 13, 2004.

[11] A. Schonhuth, “On analytic properties of entropy rate,” IEEE Trans. Inf. Theory, vol. 55, pp. 2119–2127, 2009.

[12] G. Han and B. Marcus, “Asymptotics of entropy rate in special families of hidden Markov chains,” IEEE Trans. Inf. Theory, vol. 56, pp. 1287–1295, 2010.

[13] O. Zuk, I. Kanter, and E. Domany, “The entropy of a binary hidden Markov process,” J. Stat. Phys., vol. 121, 2005.

[14] M. Reed and B. Simon, Methods of modern mathematical physics:Vol 1 Functional analysis. Academic press, 1980.