A Formula for the Divergence Rate Between Discrete Markov Sources[1]

Z. Rached
Dept. of Math. & Stats.
Jeffery Hall
Queen's University
Kingston, ON K7L 3N6, Canada
rachedz@mast.queensu.ca

F. Alajaji
Dept. of Math. & Stats.
Jeffery Hall
Queen's University
Kingston, ON K7L 3N6, Canada
fady@mast.queensu.ca

L. L. Campbell
Dept. of Math. & Stats.
Jeffery Hall
Queen's University
Kingston, ON K7L 3N6, Canada
campblll@mast.queensu.ca

Abstract —

We consider two time-invariant Markov sources of arbitrary order and finite alphabet described by the probability distributions $p^{(n)}$ and $q^{(n)}$, respectively. We show that the Kullback-Leibler divergence rate, $\lim_{n\to\infty}\frac{1}{n}D(p^{(n)}\|q^{(n)})$, between $p^{(n)}$ and $q^{(n)}$ exists and is computable. We also examine its rate of convergence and illustrate it numerically. The main tools used to obtain these results are the theory of non-negative matrices and Perron-Frobenius theory. Finally, we provide a formula for the Shannon entropy rate $\lim_{n\to\infty}\frac{1}{n}H(p^{(n)})$ of Markov sources and examine its rate of convergence.

I. Introduction

Let $\{X_1, X_2, \ldots\}$ be a first-order time-invariant Markov source with finite alphabet $\mathcal{X} = \{1, \ldots, M\}$. Consider the following two different probability laws for this source. Under the first law,

$$Pr\{X_1 = i\} =: p_i \quad \text{and} \quad Pr\{X_{k+1} = j | X_k = i\} =: p_{ij},$$

$i, j \in \mathcal{X}$, so that

$$p^{(n)}(i^n) =: Pr\{X_1 = i_1, \ldots, X_n = i_n\} = p_{i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n},$$

$i_1, \ldots, i_n \in \mathcal{X}$, while under the second law the initial probabilities are q_i, the transition probabilities are q_{ij}, and the n-tuple probabilities are $q^{(n)}$. Let $p = (p_1, \ldots, p_M)$ and $q = (q_1, \ldots, q_M)$ denote the initial distributions under $p^{(n)}$ and $q^{(n)}$ respectively.

The Kullback-Leibler divergence [11] between two distributions $\hat{p}$ and $\hat{q}$ defined on $\mathcal{X}$ is given by

$$D(\hat{p}\|\hat{q}) = \sum_{i\in\mathcal{X}} \hat{p}_i \log \frac{\hat{p}_i}{\hat{q}_i},$$

where the base of the logarithm is arbitrary. The application of this measure can be found in many areas such as approximation of probability distributions [3], [10], signal processing [8], [9], pattern recognition [1], [2], etc. One natural direction for further studies is the investigation of the Kullback-Leibler divergence rate

$$\lim_{n\to\infty} \frac{1}{n} D(p^{(n)}\|q^{(n)})$$

between two probability distributions $p^{(n)}$ and $q^{(n)}$ defined on $\mathcal{X}^n$, where

$$D(p^{(n)}\|q^{(n)}) = \sum_{i^n\in\mathcal{X}^n} p^{(n)}(i^n) \log \frac{p^{(n)}(i^n)}{q^{(n)}(i^n)},$$

for sources with memory. In [6, p. 40], Gray proved that the Kullback-Leibler divergence rate exists between a stationary source $p^{(n)}$ and a time-invariant Markov source $q^{(n)}$. This result can also be found in [13, p. 27]. To the best of our knowledge, this is the only result available in the literature about the existence and the computation of the Kullback-Leibler divergence rate between sources with memory. In the sequel, we provide a computable expression for the Kullback-Leibler divergence rate between two arbitrary time-invariant finite alphabet Markov sources. Let us first recall some useful results about non-negative stochastic matrices (i.e., with the property that the sum of the entries in each row is equal to 1) most of which may be found in [4, Chapter 3], [5, Chapter 4], and [12, Chapter 1].

II. Preliminaries

Matrices and vectors are *positive* if all their components are positive and *non-negative* if all their components are non-negative. Throughout this section, P denotes an $M \times M$ *stochastic* matrix with elements p_{ij}. The ij-th element of P^m is denoted by $p_{ij}^{(m)}$. We write $i \to j$ if $p_{ij}^{(m)} > 0$ for some positive integer m, and we write $i \not\to j$ if $p_{ij}^{(m)} = 0$ for every positive integer m. We say that i and j *communicate* and write $i \leftrightarrow j$ if $i \to j$ and $j \to i$. If $i \to j$ but $j \not\to i$ for some

[1] This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Premier's Research Excellence Award (PREA) of Ontario.

index j, then the index i is called *inessential* (or *transient*); otherwise, it is called *essential* (or *recurrent*). Thus if i is essential, $i \to j$ implies $i \leftrightarrow j$, and there is at least one j such that $i \to j$.

With these definitions, it is possible to partition the set of indices $\{1, 2, \ldots, M\}$ into disjoint sets, called *classes*. All essential indices can be subdivided into *essential classes* in such a way that all the indices belonging to one class communicate, but cannot lead to an index outside the class. Moreover, all inessential indices (if any) may be divided into two types of *inessential classes*: *self-communicating* classes and *non self-communicating* classes. Each self-communicating inessential class contains inessential indices which communicate with each other. A non self-communicating inessential class is a singleton set whose element is an index which does not communicate with any index (including itself). A matrix is *irreducible* if its indices form a single essential class; i.e., if every index communicates with every other index.

Proposition 1 [12, p. 14] By renumbering the indices (i.e., by performing row and column permutations), it is possible to put a stochastic matrix P in the *canonical form*

$$\begin{bmatrix} P_1 & \ldots & 0 & 0 & \ldots & 0 & \ldots & 0 \\ 0 & \ldots & 0 & 0 & \ldots & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & P_h & 0 & \ldots & 0 & \ldots & 0 \\ P_{h+11} & \ldots & P_{h+1h} & P_{h+1} & \ldots & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ P_{g1} & \ldots & P_{gh} & P_{gh+1} & \ldots & P_g & \ldots & 0 \\ P_{g+11} & \ldots & P_{g+1h} & P_{g+1h+1} & \ldots & P_{g+1g} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ P_{l1} & \ldots & P_{lh} & P_{lh+1} & \ldots & P_{lg} & \ldots & 0 \end{bmatrix}$$

where P_i, $i = 1, \ldots, g$, are irreducible square matrices, and in each row $i = h+1, \ldots, g$ at least one of the matrices $P_{i1}, P_{i2}, \ldots, P_{ii-1}$ is not zero. The matrix P_i for $i = 1, \ldots, h$ corresponds to the essential class C_i; while the matrix P_i for $i = h+1, \ldots, g$ corresponds to the self-communicating inessential class C_i. The other diagonal block sub-matrices which correspond to non self-communicating classes C_i, $i = g+1, \ldots, l$, are 1×1 zero matrices.

A *right eigenvector*, b, corresponding to an *eigenvalue* λ, is a nonzero vector such that $Pb = \lambda b$. A *left eigenvector*, a, corresponding to λ, is a nonzero vector such that $aP = \lambda a$. Note that a is a row vector while b is a column vector.

Proposition 2 [5, p. 115] If P is irreducible, then P has a real positive eigenvalue $\lambda = 1$ that is greater than or equal to the magnitude of each other eigenvalue. There is a positive left (right) eigenvector, $a(b)$, corresponding to λ, unique within a scale factor.

Remark: The left eigenvector a is the unique stationary distribution π of P associated with the largest positive real eigenvalue $\lambda = 1$ and $b^t = (1, \ldots, 1)$, where t denotes the transpose operation.

Proposition 3 [7, p. 524] Let P be the probability transition matrix for an irreducible Markov source. Also, let $a(b)$ be the left (right) eigenvector associated with the largest positive real eigenvalue $\lambda = 1$ such that $ab = 1$. Also, let $L = ba$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P^i = L.$$

Moreover, there exists a finite positive constant C such that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} P^i - L \right\|_\infty \le \frac{C}{n},$$

for all $n = 1, 2, \ldots$ and $\|\cdot\|_\infty$ is the l_∞ *norm*, where the l_∞ norm of an $M \times M$ matrix A is defined by $\|A\|_\infty \triangleq \max_{1\le i,j\le M} |a_{ij}|$.

With the aid of Proposition 1 and Proposition 3, it can be shown that the cesáro limit $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} P^i$ of an arbitrary (not necessarily irreducible) stochastic matrix P exists and is computable.

Proposition 4 [4, p. 129] Let P be the probability transition matrix for an arbitrary Markov source with associated canonical form as in Proposition 1. Let $a_i(b_i)$ be the left (right) eigenvector associated with $\lambda = 1$ such that $a_i b_i = 1$, for $i = 1, \ldots, h$. Let

$$D = \begin{bmatrix} b_1 a_1 & \ldots & 0 \\ 0 & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & b_h a_h \end{bmatrix}, \quad B = \begin{bmatrix} P_{h+11} & \ldots & P_{h+1h} \\ \ldots & \ldots & \ldots \\ P_{g1} & \ldots & P_{gh} \\ P_{g+11} & \ldots & P_{g+1h} \\ \ldots & \ldots & \ldots \\ P_{l1} & \ldots & P_{lh} \end{bmatrix}.$$

Also, let

$$C = \begin{bmatrix} P_{h+1} & \ldots & 0 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ P_{gh+1} & \ldots & P_g & \ldots & \ldots & 0 \\ P_{g+1h+1} & \ldots & P_{g+1g} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ P_{lh+1} & \ldots & P_{lg} & P_{lg+1} & \ldots & 0 \end{bmatrix}.$$

We have the following:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P^i = \begin{bmatrix} D & 0 \\ (I - C)^{-1} BD & 0 \end{bmatrix},$$

where I is the identity matrix.

Proposition 5 **[7, p. 492]** Let A be a non-negative matrix. The spectral radius $\rho(A) \triangleq \max\ \{|\lambda| : \lambda \text{ eigenvalue of } A\}$ satisfies

$$\min\{\textit{row sum}\} \le \rho(A) \le \max\{\textit{row sum}\}.$$

The following lemma follows by appropriately modifying the proof of the above proposition.

Lemma 1 If A is non-negative and irreducible and the row sums are not all identical, then the spectral radius $\rho(A)$ satisfies

$$\min\{\textit{row sum}\} < \rho(A) < \max\{\textit{row sum}\}.$$

Proof: Let λ be the largest positive real eigenvalue of A with associated strictly positive left eigenvector a. Without loss of generality a can be normalized, i.e., the sum of its components is equal to 1. Let $\mathbf{1}^t$ be the row vector

$$\mathbf{1}^t = (1, \dots, 1).$$

Note that $a\mathbf{1} = 1$. We have $aA = \lambda a$. Hence $aA\mathbf{1} = \lambda a \mathbf{1} = \lambda$. On the other hand

$$aA\mathbf{1} = a \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_M \end{bmatrix},$$

where R_i, $i = 1, \dots, M$ denotes the sum of the i-th row. Let

$$R_{\max} = \max\{R_1, \dots, R_M\}.$$

Then

$$aA\mathbf{1} < a \begin{bmatrix} R_{\max} \\ R_{\max} \\ \vdots \\ R_{\max} \end{bmatrix} = \sum_{i=1}^{M} a_i R_{\max} = R_{\max}.$$

Therefore

$$\lambda < R_{\max}$$

Similarly, we can show that

$$\lambda > R_{\min},$$

where

$$R_{\min} = \min\{R_1, \dots, R_M\}.$$

Finally we conclude that

$$R_{\min} < \rho(A) < R_{\max}.$$

□

Proposition 6 **[7, p. 494]** If a non-negative matrix A has a right positive eigenvector b, then for all $n = 1, 2, \dots$ and for all $i = 1, \dots, M$ we have

$$\sum_{j=1}^{M} a_{ij}^{(n)} \le \left[\frac{\max_{1 \le k \le M} b_k}{\min_{1 \le k \le M} b_k} \right] \rho^n(A).$$

The following corollary follows directly from the previous proposition by observing that, $a_{ij}^{(n)} \le \sum_{j=1}^{M} a_{ij}^{(n)}$ for all $i = 1, \dots, M$ and $j = 1, \dots, M$.

Corollary 1 If A is non-negative and irreducible, then $A^n \le \rho^n(A) C$ (i.e., $a_{ij}^{(n)} \le \rho^n(A) c_{ij}$), for all $n = 1, 2, \dots$, where $C = (\frac{\max_{1 \le k \le M} b_k}{\min_{1 \le k \le M} b_k})$ is a matrix with identical entries that are independent of n.

III. Kullback-Leibler divergence rate

A First-order Markov sources

We assume first that the Markov source $\{X_1, X_2, \dots\}$ is of order one. Later, we generalize the results for sources of arbitrary order k. Let p and q be the initial distributions with respect to $p^{(n)}$ and $q^{(n)}$ respectively. Also, let P and Q be the probability transition matrices with respect to $p^{(n)}$ and $q^{(n)}$ respectively. Without loss of generality, we may assume that p and P are absolutely continuous with respect to q and Q respectively (i.e., $q_i = 0 \Rightarrow p_i = 0$ and $q_{ij} = 0 \Rightarrow p_{ij} = 0$ for all $i, j \in \mathcal{X}$). We have the following results.

Theorem 1 Suppose that the Markov source $\{X_1, X_2, \dots\}$ under $p^{(n)}$ and $q^{(n)}$ is irreducible. Let

$$V^t = (S(X_2|X_1 = 1), \dots, S(X_2|X_1 = M)),$$

where

$$S(X_2|X_1 = i) \triangleq \sum_{j \in \mathcal{X}} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

The Kullback-Leibler divergence rate between $p^{(n)}$ and $q^{(n)}$ is given by

$$\lim_{n \to \infty} \frac{1}{n} D(p^{(n)} \| q^{(n)}) = \sum_{i \in \mathcal{X}} \pi_i S(X_2|X_1 = i),$$

where $\pi = (\pi_1, \dots, \pi_M)$ is the unique stationary distribution of P.

Proof: We have that

$$\frac{1}{n} D(p^{(n)} \| q^{(n)}) = \frac{1}{n} \sum_{i \in \mathcal{X}} [p(X_1 = i) + \cdots + p(X_{n-1} = i)] S(X_2|X_1 = i) + \frac{1}{n} \sum_{i \in \mathcal{X}} p(X_1 = i) \log \frac{p(X_1 = i)}{q(X_1 = i)},$$

which can be also written as

$$\frac{1}{n}D(p^{(n)}\|q^{(n)}) = \frac{1}{n}p(I+P+\cdots+P^{n-2})V \quad (1)$$
$$+\frac{1}{n}\sum_{i\in\mathcal{X}} p_i \log\frac{p_i}{q_i}. \quad (2)$$

Note that (2) approaches 0 as $n\to\infty$. Hence, by Proposition 3, we obtain the desired result. $\square$

Theorem 2 Suppose that the Markov sources $p^{(n)}$ and $q^{(n)}$ are arbitrary (not necessarily irreducible, stationary, etc.). Let $V^t = (S(X_2|X_1=1),\ldots,S(X_2|X_1=M))$, where

$$S(X_2|X_1=i) \triangleq \sum_{j\in\mathcal{X}} p_{ij}\log\frac{p_{ij}}{q_{ij}}.$$

Let the canonical form of P be as in Proposition 1. Also, let B, D and C be as defined in Proposition 4. The Kullback-Leibler divergence rate between $p^{(n)}$ and $q^{(n)}$ is given by

$$\lim_{n\to\infty}\frac{1}{n}D(p^{(n)}\|q^{(n)}) = p\begin{bmatrix} D & 0 \\ (I-C)^{-1}BD & 0 \end{bmatrix}V.$$

Proof: We have that

$$\frac{1}{n}D(p^{(n)}\|q^{(n)}) = \frac{1}{n}p(I+P+\cdots+P^{n-2})V + \frac{1}{n}\sum_{i\in\mathcal{X}} p_i\log\frac{p_i}{q_i}.$$

Then, the desired result follows immediately from Proposition 4. $\square$

Theorem 3 The rate of convergence of the Kullback-Leibler divergence rate between $p^{(n)}$ and $q^{(n)}$ is of the order $1/n$.

Proof: Clearly, the rate of convergence of (2) to 0 is of the order $1/n$. In Proposition 3, it is proved that the rate of convergence of the cesáro sum of an irreducible matrix is of the order $1/n$. On the other hand, if P is not irreducible, let P_i, $i=h+1,\ldots,g$ be the sub-matrices corresponding to inessential classes as in Proposition 1. Every P_i is irreducible and hence, by Corollary 1, we have that

$$P_i^n \le \rho^n(P_i)G_i, \quad i=h+1,\ldots,g \quad (3)$$

where G_i is a matrix with identical entries that are independent of n. If P_i has all row sums identical then $\rho(P_i)<1$ by Proposition 5. Otherwise, $\rho(P_i)<1$ by Lemma 1. Hence, by (3), P_i^n converges exponentially fast to the zero matrix of the same dimensions for each $i=h+1,\ldots,g$. By considering the cesáro sum of the canonical form of P, we get that the rate of convergence of (1) is of the order $1/n$. Therefore the rate of convergence of the Kullback-Leibler divergence rate is of the order $1/n$. $\square$

B k-th order Markov sources

Now, suppose that the Markov source has an arbitrary order k. Define $\{W_1,W_2,\ldots\}$ as the process obtained by k-step blocking the Markov source $\{X_1,X_2,\ldots\}$; i.e.,

$$W_n := (X_n, X_{n+1},\ldots,X_{n+k-1}).$$

Then $\{W_n\}$ is a first order Markov source with M^k states. Let $p_{w_{n-1}w_n} := Pr(W_n=w_n|W_{n-1}=w_{n-1})$. Let $p=(p_1,\ldots,p_{M^k})$ and $q=(q_1,\ldots,q_{M^k})$ denote the arbitrary initial distributions of W_1 under $p^{(n)}$ and $q^{(n)}$ respectively. Also, let p_{ij} and q_{ij} denote the transition probability that W_n goes from index i to index j under $p^{(n)}$ and $q^{(n)}$ respectively, $i,j=1,\ldots,M^k$. Then clearly $D(p^{(n)}\|q^{(n)})$ can be written as

$$\frac{1}{n}D(p^{(n)}\|q^{(n)}) = \frac{1}{n}p(I+P+\cdots+P^{n-2})V + \frac{1}{n}\sum_{i\in\mathcal{X}^k} p(W_1=i)\log\frac{p(W_1=i)}{q(W_1=i)}.$$

It follows directly that the previous results also hold for a Markov source of arbitrary order.

IV. Shannon entropy rate

The existence and the computation of the Shannon entropy rate of an arbitrary time-invariant finite-alphabet Markov source can be deduced from the existence and the computation of the Kullback-Leibler divergence rate. We have the following corollaries.

Corollary 2 Suppose that the Markov source $\{X_1,X_2,\ldots\}$ under $p^{(n)}$ is irreducible. Let

$$V^t = (H(X_2|X_1=1),\ldots,H(X_2|X_1=M)),$$

where

$$H(X_2|X_1=i) \triangleq -\sum_{j\in\mathcal{X}} p_{ij}\log p_{ij}.$$

The Shannon entropy rate of $p^{(n)}$ is given by

$$\lim_{n\to\infty}\frac{1}{n}H(p^{(n)}) = \sum_{i\in\mathcal{X}}\pi_i H(X_2|X_1=i),$$

where $\pi=(\pi_1,\ldots,\pi_M)$ is the unique stationary distribution of P.

Corollary 3 Let the canonical form of P be as in Proposition 1. Also, let B, D and C be as defined in Proposition 4. Then, the Shannon entropy rate is given by

$$\lim_{n\to\infty}\frac{1}{n}H(p^{(n)}) = p\begin{bmatrix} D & 0 \\ (I-C)^{-1}BD & 0 \end{bmatrix}V,$$

where $V^t=(H(X_2|X_1=1),\ldots,H(X_2|X_1=M))$, and

$$H(X_2|X_1=i) \triangleq -\sum_{j\in\mathcal{X}} p_{ij}\log p_{ij}.$$

Corollary 4 The rate of convergence of the Shannon entropy rate of $p^{(n)}$ is of the order $1/n$.

V. Numerical examples

In this section, we use the natural logarithm.

Example 1: Let P and Q be two possible probability transition matrices for $\{X_1, X_2, \ldots\}$ defined as follows:

$$P = \begin{bmatrix} 1/4 & 0 & 0 & 1/2 & 1/4 \\ 2/3 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1/5 & 0 & 4/5 \\ 4/7 & 0 & 3/7 & 0 & 0 \\ 0 & 0 & 3/4 & 0 & 1/4 \end{bmatrix},$$

and

$$Q = \begin{bmatrix} 2/5 & 0 & 0 & 2/5 & 1/5 \\ 4/5 & 0 & 1/5 & 0 & 0 \\ 0 & 0 & 2/3 & 0 & 1/3 \\ 5/6 & 0 & 1/6 & 0 & 0 \\ 0 & 0 & 3/8 & 0 & 5/8 \end{bmatrix}.$$

Let $p = (2/7, 4/7, 1/7, 0, 0)$ and $q = (1/5, 1/5, 3/5, 0, 0)$ be two possible initial distributions under $p^{(n)}$ and $q^{(n)}$, respectively. In canonical form, P and Q can be rewritten as

$$P = \begin{bmatrix} 1/5 & 4/5 & 0 & 0 & 0 \\ 3/4 & 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/2 & 0 \\ 3/7 & 0 & 4/7 & 0 & 0 \\ 1/3 & 0 & 2/3 & 0 & 0 \end{bmatrix},$$

and

$$Q = \begin{bmatrix} 2/3 & 1/3 & 0 & 0 & 0 \\ 3/8 & 5/8 & 0 & 0 & 0 \\ 0 & 1/5 & 2/5 & 2/5 & 0 \\ 1/6 & 0 & 5/6 & 0 & 0 \\ 1/5 & 0 & 4/5 & 0 & 0 \end{bmatrix},$$

simply by permuting the second and fifth rows (columns) and the first and third rows (columns). Note that P has 1 essential class, 1 inessential self-communicating class and 1 inessential non self-communicating class. Accordingly, the initial distributions are rewritten as $p = (1/7, 0, 2/7, 0, 4/7)$ and $q = (3/5, 0, 1/5, 0, 1/5)$, after permuting the first and third indices and the second and fifth indices. We obtain the following.

n	$\frac{1}{n}D(p^{(n)}\|q^{(n)})$
10	0.3473
50	0.3671
100	0.3698

By Theorem 2, the Kullback-Leibler divergence rate is equal to 0.3725. Clearly, as n gets large $\frac{1}{n}D(p^{(n)}\|q^{(n)})$ is closer to the Kullback-leibler divergence rate.

Example 2: Suppose that the Markov source is of order 2 under $p^{(n)}$ and $q^{(n)}$ respectively. Let $\{W_1, W_2, \ldots\}$ be the process obtained by 2-step blocking the Markov source. Let P and Q be two possible transition matrices for $\{W_1, W_2, \ldots\}$ defined as follows:

$$P = \begin{bmatrix} 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2/5 & 3/5 & 0 & 0 \\ 0 & 0 & 1/6 & 5/6 \end{bmatrix},$$

and

$$Q = \begin{bmatrix} 3/4 & 1/4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 7/8 & 1/8 & 0 & 0 \\ 0 & 0 & 2/3 & 1/3 \end{bmatrix},$$

Let $p = (1/8, 3/8, 2/8, 2/8)$ and $q = (1/7, 2/7, 3/7, 1/7)$ denote two possible initial distributions of W_1 under $p^{(n)}$ and $q^{(n)}$ respectively. The set of indices $\{1, 2, 3\}$ forms an essential class, while the singleton set $\{4\}$ forms a self-communicating non-essential class. Hence, P and Q are not irreducible. We obtain the following.

n	$\frac{1}{n}D(p^{(n)}\|q^{(n)})$
10	0.2982
50	0.3253
100	0.3277

By Theorem 2, the Kullback-Leibler divergence rate is equal to .3301. Clearly, as n gets large $\frac{1}{n}D(p^{(n)}\|q^{(n)})$ is closer to the Kullback-leibler divergence rate.

Example 3: Consider the Markov source under $p^{(n)}$ as in Example 1. We obtain the following.

n	$\frac{1}{n}H(p^{(n)})$
10	0.5437
50	0.5088
100	0.5044

By Corollary 3, the Shannon entropy rate is equal to 0.5001. Clearly, as n gets large $\frac{1}{n}H(p^{(n)})$ is closer to the Shannon entropy rate.

Example 4: Consider the following second order Markov source with probability transition matrix

$$P \begin{bmatrix} 1/3 & 2/3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/4 & 3/4 \end{bmatrix},$$

and initial distribution $p = (1/5, 2/5, 0, 2/5)$. The set of indices $\{1, 2, 3\}$ forms an essential class, while the singleton set $\{4\}$ forms a self-communicating non-essential class. Hence, P is not irreducible. We obtain the following.

n	$\frac{1}{n}H(p^{(n)})$
10	0.4641
50	0.4339
100	0.4298

By Corollary 3, the Shannon entropy rate is equal to 0.4256. Clearly, as n gets large $\frac{1}{n}H(p^{(n)})$ is closer to the Shannon entropy rate.

VI. Conclusion and Future work

In this work, we derived a formula for the Kullback-Leibler divergence rate between two time-invariant finite-alphabet arbitrary Markov sources (not necessarily, irreducible, stationary, etc.). We illustrated numerically and investigated its rate of convergence. Finally, we examined the computation and the existence of the Shannon entropy rate for Markov sources and investigated its rate of convergence. A possible future direction is the investigation of the results for Hidden Markov sources and for more general sources with memory such as stationary ergodic sources.

References

[1] M. B. Bassat, "f-entropies, probability of error, and feature selection," *Inform. Contr.*, vol. 39, pp. 227–242, 1978.

[2] C. H. Chen, *Statistical Pattern Recognition*, Rochelle Park, NJ: Hayden Book Co., Ch. 4, 1973.

[3] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 3, pp. 462–467, May 1968.

[4] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*, Methuen and Co Ltd, 1965.

[5] R. G. Gallager, *Discrete Stochastic Processes*, Kluwer, Boston, 1996.

[6] R. M. Gray, *Entropy and Information Theory*, Springer-Verlag, New York, 1990.

[7] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[8] T. T. Kadota and L. A. Shepp, "On the best finite set of linear observables for discriminating two Gaussians signals," *IEEE Trans. Inform. Theory*, vol. IT-13, no. 2, pp. 278–284, Apr. 1967.

[9] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. COM-15, no. 1, pp. 52–60, Feb. 1967.

[10] D. Kazakos and T. Cotsidas, "A decision theory approach to the approximation of discrete probability densities," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, vol. 1, pp. 61–67, Jan. 1980.

[11] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[12] E. Seneta, *Non-Negative Matrices and Markov Chains*, Springer-Verlag New York Inc., 1981.

[13] Z. Ye and T. Berger, *Information Measures For Discrete Random Fields*, Science Press, Beijing, New York, 1998.