The Kullback-Leibler Divergence Rate Between Markov Sources*

Ziad Rached Fady Alajaji L. Lorne Campbell

Abstract

In this work, we provide a computable expression for the Kullback-Leibler divergence rate, $\lim_{n\to\infty}\frac{1}{n}D(p^{(n)}\|q^{(n)})$, between two time-invariant finite-alphabet Markov sources of arbitrary order and arbitrary initial distributions described by the probability distributions $p^{(n)}$ and $q^{(n)}$, respectively. We illustrate it numerically and examine its rate of convergence. The main tools used to obtain the Kullback-Leibler divergence rate and its rate of convergence are the theory of non-negative matrices and Perron-Frobenius theory. Similarly, 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.

Index Terms: Decision theory, classifcation, pattern recognition, time-invariant Markov sources, Kullback-Leibler divergence rate, Shannon entropy rate, non-negative matrices, Perron-Frobenius theory, rate of convergence.

*IEEE Transactions on Information Theory: Submitted July 2001; revised November 2003. This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

Z. Rached was with the Department of Mathematics & Statistics, Queen's University, Kingston, ON K7L 3N6, Canada; he is now with the Department of Mathematics & Statistics, Notre Dame University, Zouk Mosbeh, Keserouan, P. O. Box 72 Zouk Mikael, Lebanon. F. Alajaji (e-mail: fady@mast.queensu.ca) and L. L. Campbell are with the Department of Mathematics & Statistics, Queen's University, Kingston, ON K7L 3N6, Canada.

1 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,

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

so that

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

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 [13] 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 the Kullback-Leibler divergence can be found in many areas such as approximation of probability distributions [3], [12], signal processing [10], [11], [5], 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 previous work, Gray [8] 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 [18, p. 27]. In [14], the authors noted that the Kullback-Leibler divergence rate between ergodic Markov sources exists. In [17], Shields presented two examples for non-Markovian sources for which the Kullback-Leibler divergence rate does not exist. Finally, in [5], Do provides an upper bound for the Kullback-Leibler divergence rate between stationary hidden Markov sources. To the best of our knowledge, these are the only results available in the literature about the existence and/or computation of the Kullback-Leibler divergence rate between sources with memory.

We herein provide an explicit computable expression for the Kullback-Leibler divergence rate between two arbitrary time-invariant (not necessarily stationary, irreducible) finite-alphabet Markov sources. This expression, which is proved in a straightforward manner using results from the theory of non-negative matrices and Perron-Frobenius theory, has a readily usable form, making it appealing for various analytical studies and applications involving the divergence rate for systems with memory.

The rest of this work is organized as follows. Preliminaries about the theory of non-negative matrices are first briefly presented in Section 2. In Section 3, an explicit formula for the divergence rate between arbitrary time-invariant finite-alphabet Markov sources is derived and its rate of convergence is investigated. A similar study for the expression and convergence rate of the Shannon entropy rate of time-invariant (non-stationary in general) Markov sources is briefly addressed in Section 4. Numerical examples are presented in Section 5 and conclusions are stated in Section 6.

2 Preliminaries

Matrices and vectors are *positive* if all their components are positive and *non-negative* if all their components are non-negative. Throughout, A denotes an $M \times M$ non-negative matrix with elements a_{ij}. The ij-th element of A^m is denoted by $a_{ij}^{(m)}$.

We write $i \to j$ if $a_{ij}^{(m)} > 0$ for some positive integer m, and we write $i \not\to j$ if $a_{ij}^{(m)} = 0$ 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 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 [16, p. 15] By renumbering the indices (i.e., by performing row and column permutations), it is possible to put matrix A in the *canonical form*

$$A = \begin{bmatrix} A_1 & \ldots & 0 & 0 & \ldots & 0 & \ldots & \ldots & 0 \\ 0 & \ldots & 0 & 0 & \ldots & 0 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & A_h & 0 & \ldots & 0 & \ldots & \ldots & 0 \\ A_{h+11} & \ldots & A_{h+1h} & A_{h+1} & \ldots & 0 & \ldots & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ A_{g1} & \ldots & A_{gh} & A_{gh+1} & \ldots & A_g & \ldots & \ldots & 0 \\ A_{g+11} & \ldots & A_{g+1h} & A_{g+1h+1} & \ldots & A_{g+1g} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ A_{l1} & \ldots & A_{lh} & A_{lh+1} & \ldots & A_{lg} & A_{lg+1} & \ldots & 0 \end{bmatrix}$$

where A_i, $i = 1, \ldots, g$, are irreducible square matrices, and in each row $i = h+1, \ldots, g$ at least one of the matrices $A_{i1}, A_{i2}, \ldots, A_{ii-1}$ is not zero. The matrix A_i for $i = 1, \ldots, h$ corresponds to the essential class C_i; while the matrix A_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. In every row $i = g+1, \ldots, l$ any of the matrices $A_{i1}, \ldots, A_{ii-1}$ may be zero.

Proposition 2 [9, p. 492] Suppose A is irreducible and let R_i, $i = 1, \ldots, M$ denote the sum of the i-th row. Also, let $R_{\max} = \max\{R_1, \ldots, R_M\}$ and $R_{\min} = \min\{R_1, \ldots, R_M\}$. Then the largest positive real eigenvalue λ satisfies

$$R_{\min} \leq \lambda \leq R_{\max}.$$

The following lemma follows by appropriately modifying the proof of the above proposition and applying the Frobenius theorem [7, p. 115].

Lemma 1 If A is irreducible and the row sums are not all identical, then the largest positive real eigenvalue λ satisfies,

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

With the aid of [9, Theorem 8.6.1, p. 524] and Proposition 1, it can be shown that for an arbitrary stochastic matrix P (i.e., with non-negative entries and every row-sum equal to one), the Cesáro limit, $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} P^i$, exists and is computable.

Proposition 3 [4, p. 129] Let P be the probability transition matrix for an arbitrary Markov source with associated canonical form as in Proposition 1:

$$P = \begin{bmatrix} \Gamma & 0 \\ B & C \end{bmatrix},$$

where

$$\Gamma = \begin{bmatrix} P_1 & \ldots & 0 \\ 0 & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & P_h \end{bmatrix}, \quad B = \begin{bmatrix} P_{h+11} & \ldots & P_{h+1h} \\ \ldots & \ldots & \ldots \\ P_{g1} & \ldots & P_{gh} \\ P_{g+11} & \ldots & P_{g+1h} \\ \ldots & \ldots & \ldots \\ P_{l1} & \ldots & P_{lh} \end{bmatrix},$$

and

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

Let a_i (b_i) be the left (right) eigenvector of P_i associated with $\lambda = 1$ such that $a_i b_i = 1$, for $i = 1, \dots, h$, and define

$$D = \begin{bmatrix} b_1 a_1 & \dots & 0 \\ 0 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & b_h a_h \end{bmatrix}.$$

We then 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.

Remark: For each $i = 1, 2, \cdots, h$, the above left eigenvector a_i is the unique stationary distribution π of P_i and $b_i^t = (1, \dots, 1)$, where t denotes the transpose operation.

3 Kullback-Leibler Divergence Rate

3.1 First-Order Markov Sources

We first assume that the time-invariant 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 two initial distributions and P and Q be two probability transition matrices for the source, yielding n-tuple distributions $p^{(n)}$ and $q^{(n)}$ respectively. We assume that p is absolutely continuous with respect to q ($p << q$) and that P is absolutely continuous with respect to Q ($P << Q$); i.e., $q_i = 0 \Rightarrow p_i = 0$ and $q_{ij} = 0 \Rightarrow p_{ij} = 0$, for all $i, j \in \mathcal{X}$. These conditions ensure that $p^{(n)} << q^{(n)}$ for each n and cover most cases of interest regarding the computation of the divergence rate. We then have the following results.

Theorem 1 Suppose that the Markov source $\{X_1, X_2, \ldots\}$ is irreducible under P and Q. Let

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

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

Proof: First note that $S(X_2|X_1 = i)$ is well defined for all $i \in \mathcal{X}$ since $P << Q$. Furthermore, since both $p << q$ and $P << Q$ hold, we have after expanding the logarithm that

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

where

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

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

$$\lim_{n \to \infty} \frac{1}{n} p(I + P + \cdots + P^{n-2}) V = pLV,$$

where

$$L = ba = (1, \ldots, 1)^t (\pi_1, \ldots, \pi_M) = \begin{bmatrix} \pi_1 & \pi_2 & \ldots & \pi_M \\ \pi_1 & \pi_2 & \ldots & \pi_M \\ \vdots & \vdots & \vdots & \vdots \\ \pi_1 & \pi_2 & \ldots & \pi_M \end{bmatrix},$$

where a (b) is the left (right) eigenvector of P associated with the largest real eigenvalue $\lambda = 1$ such that $ab = 1$ (note that since P is irreducible, then it is already in its canonical form; so $g = h = 1$ in Proposition 3 with the Cesáro limit trivially reducing to $D = ba$). Thus

$$\lim_{n\to\infty} \frac{1}{n} D(p^{(n)} \| q^{(n)}) = p \begin{bmatrix} \pi_1 & \pi_2 & \dots & \pi_M \\ \pi_1 & \pi_2 & \dots & \pi_M \\ \vdots & \vdots & \vdots & \vdots \\ \pi_1 & \pi_2 & \dots & \pi_M \end{bmatrix} V = \sum_{i\in\mathcal{X}} \pi_i S(X_2|X_1 = i).$$

$\square$

Theorem 2 Suppose that the Markov source $\{X_1, X_2, \ldots\}$ under $p^{(n)}$ and $q^{(n)}$ is arbitrary[1] (not necessarily irreducible, stationary, etc.). Let the canonical form of P be as in Proposition 1. Also, let B, D and C be as defined in Proposition 3. Then, the Kullback-Leibler divergence rate between $p^{(n)}$ and $q^{(n)}$ is given by

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

where

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

and I is the identity matrix with same dimensions as the matrix C.

Proof: As in the previous theorem, we have that

$$\frac{1}{n} D(p^{(n)} \| q^{(n)}) = \frac{1}{n} p(I + P + \cdots + P^{n-2})V \tag{3}$$
$$+ \frac{1}{n} \sum_{i\in\mathcal{X}} p_i \log \frac{p_i}{q_i}. \tag{4}$$

Then, the desired result follows immediately from Proposition 3.

$\square$

[1] Since p and P are assumed to be absolutely continuous with respect to q and Q respectively, it follows that $p^{(n)}$ is absolutely continuous with respect to $q^{(n)}$. Hence, some restriction on their behavior is induced. For instance, if P is irreducible, Q must be irreducible. However, it is possible to have Q irreducible and P reducible. So, in general, Q and P do not necessarily have the same number of classes.

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

Proof: Clearly, the rate of convergence of (4) to 0 is of the order $1/n$. In [9, Theorem 8.6.1, p. 524], it is proved that the rate of convergence of the Cesáro sum of an irreducible stochastic matrix is of the order $1/n$. On the other hand, if P is not irreducible, let P_i, $i = 1, \ldots, h$, be the sub-matrices corresponding to essential classes and let P_i, $i = h+1, \ldots, g$ be the sub-matrices corresponding to inessential classes as in Proposition 1. For $i = 1, \ldots, h$, each P_i is stochastic and irreducible; so its Cesáro-sum is of the order $1/n$ by [9, Theorem 8.6.1, p. 524]. Now, for $i = h+1, \ldots, g$, every P_i is irreducible and hence, by [15, Corollary 1], we have that

$$P_i^n \leq \lambda_i^n G_i, \quad i = h+1, \ldots, g, \tag{5}$$

where λ_i is the largest positive real eigenvalue of P_i, and G_i is a matrix with identical entries that are independent of n. Therefore

$$\frac{1}{n}\sum_{j=1}^{n} P_i^j \;\; \leq \;\; \frac{1}{n}\sum_{j=1}^{n} \lambda_i^j G_i \;\; = \;\; \frac{1}{n}\frac{\lambda_i(1-\lambda_i^n)}{1-\lambda_i} G_i,$$

for $i = h+1, \ldots, g$. If P_i has all row sums identical, then $\lambda_i < 1$ by Proposition 2, the fact that P is stochastic and the fact that, in the canonical form of P, at least one of the matrices $P_{i1}, P_{i2}, \cdots, P_{ii-1}$ is non-zero when $i = h+1, \ldots, g$ (so that the row sums of P_i are strictly less than one). Otherwise, $\lambda_i < 1$ by Lemma 1. Hence, the Cesáro sum of P_i, $i = h+1, \ldots, g$ is of the order $1/n$. By considering the Cesáro sum of the canonical form of P, we get that the rate of convergence of (3) is of the order $1/n$. Therefore the rate of convergence of the Kullback-Leibler divergence rate is of the order $1/n$. □

3.2 *k*-th Order Markov Sources

We next suppose that the Markov source $\{X_n\}$ has an arbitrary order k, and let $\tilde{p}^{(n)}$ and $\tilde{q}^{(n)}$ be two possible n-tuple distributions for $\{X_n\}$. Define $\{W_n\}$ as the process obtained by k-step blocking the Markov source $\{X_n\}$; 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 $p = (p_1, \ldots, p_{M^k})$ and $q = (q_1, \ldots, q_{M^k})$ denote the initial distributions of W_1 and let $P = [p_{ij}]$ and $Q = [q_{ij}]$, (with $i, j = 1, \ldots, M^k$) denote the probability transition matrices for $\{W_n\}$, resulting in n-tuple distributions $p^{(n)}$ and $q^{(n)}$ respectively.

We first note that since $\tilde{p}^{(n+k-1)}(x^{n+k-1}) = p^{(n)}(w^n)$ and $\tilde{q}^{(n+k-1)}(x^{n+k-1}) = q^{(n)}(w^n)$ for all $n \geq 1$, then $D(\tilde{p}^{(n+k-1)}\|\tilde{q}^{(n+k-1)}) = D(p^{(n)}\|q^{(n)})$. Therefore, the divergence rates for $\{X_n\}$ and $\{W_n\}$ are identical since $(n+k-1)/n \to 1$ as $n \to \infty$. Now clearly, $D(p^{(n)}\|q^{(n)})$ can be written as

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

where

$$V^t = (S(W_2|W_1 = 1), \ldots, S(W_2|W_1 = M^k)).$$

It then directly follows that Theorems 2 and 3 also hold for a Markov source of arbitrary order k.

4 Shannon Entropy Rate

The existence and the computation of the Shannon entropy rate of an arbitrary time-invariant finite-alphabet Markov source can be directly deduced from the existence and the computation of the Kullback-Leibler divergence rate. Indeed, if $q^{(n)}$ is stationary memoryless with uniform marginal distribution, then

$$D(p^{(n)}\|q^{(n)}) = n \log M - H(p^{(n)}).$$

Therefore

$$\lim_{n\to\infty} \frac{1}{n}D(p^{(n)}\|q^{(n)}) = \log M - \lim_{n\to\infty} \frac{1}{n}H(p^{(n)}). \qquad (6)$$

We have the following corollaries.

Corollary 1 Suppose that the Markov source $\{X_1, X_2, \ldots\}$ under P is irreducible. Let

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

Then, 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.

Proof: Obtained directly by plugging $q_{ij} = 1/M$ in Theorem 1 and using (6).

$\square$

Corollary 2 Let the canonical form of P be as in Proposition 1. Also, let B, D and C be as defined in Proposition 3. 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 I is the identity matrix with the same dimensions as the matrix C.

Proof: Note that P^i, $i = 1, 2, \ldots$ is a stochastic matrix[2]. Hence,

$$\begin{aligned} \lim_{n \to \infty} \frac{1}{n} (I + P + \cdots + P^{n-2}) \mathbf{1}^t &= \lim_{n \to \infty} \frac{n-1}{n} \mathbf{1}^t \\ &= \mathbf{1}^t \end{aligned}$$

which yields that

$$\lim_{n \to \infty} \frac{1}{n} (I + P + \cdots + P^{n-2})$$

is a stochastic matrix. Therefore,

$$\begin{bmatrix} D & 0 \\ (I - C)^{-1} BD & 0 \end{bmatrix}$$

[2] We have that $P\mathbf{1}^t = \mathbf{1}^t$, where $\mathbf{1} = (1, \ldots, 1)$ and t is the transpose operation. Using this fact and the fact that $P^i = PP^{i-1}$, the result follows by mathematical induction on i.

is also a stochastic matrix. Hence,

$$\begin{aligned} p\begin{bmatrix} D & 0 \\ (I-C)^{-1}BD & 0 \end{bmatrix}\begin{bmatrix} \log M \\ \vdots \\ \log M \end{bmatrix} &= p\begin{bmatrix} \log M \\ \vdots \\ \log M \end{bmatrix} \\ &= \log M. \end{aligned}$$

Then, the corollary follows directly by plugging $q_{ij} = \frac{1}{M}$ in Theorem 2 and using (6).

$\square$

Remark: It was mentioned in [6, p. 68] that the Shannon entropy rate for an arbitrary time-invariant finite-alphabet Markov source exists, but no computational details nor an explicit analytical expression for the entropy rate (as shown above) were provided.

Corollary 3 The rate of convergence of the Shannon entropy rate of $p^{(n)}$ is of the order $1/n$.

5 Numerical Examples

In this section, we use the natural logarithm for simplicity.

Example 1: Let P and Q be two possible probability transition matrices for a first order Markov source $\{X_1, X_2, \ldots\}$ (not stationary and not irreducible) defined as follows:

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

and

$$Q = \begin{bmatrix} 1/3 & 0 & 0 & 2/3 & 0 & 0 & 0 \\ 0 & 0 & 2/7 & 1/7 & 4/7 & 0 & 0 \\ 0 & 0 & 1/5 & 0 & 0 & 4/5 & 0 \\ 1/6 & 0 & 0 & 5/6 & 0 & 0 & 0 \\ 1/5 & 2/5 & 0 & 0 & 2/5 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1/4 & 0 & 1/4 & 0 & 1/2 & 0 & 0 \end{bmatrix}.$$

Let $p = (3/7, 0, 1/7, 0, 1/7, 2/7, 0)$ and $q = (2/8, 0, 3/8, 0, 1/8, 2/8, 0)$ be two possible initial distributions under $p^{(n)}$ and $q^{(n)}$, respectively. In canonical form, P and Q can be rewritten as

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

and

$$Q = \begin{bmatrix} 1/5 & 4/5 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 2/3 & 0 & 0 & 0 \\ 0 & 0 & 1/6 & 5/6 & 0 & 0 & 0 \\ 0 & 0 & 1/5 & 0 & 2/5 & 2/5 & 0 \\ 2/7 & 0 & 0 & 1/7 & 4/7 & 0 & 0 \\ 1/4 & 0 & 1/4 & 0 & 1/2 & 0 & 0 \end{bmatrix},$$

simply by permuting the first and third rows (columns) and the second and sixth rows (columns). Note that P has 2 essential classes, 1 inessential self-communicating class and 1 inessential non self-communicating class. Accordingly, the initial distributions are rewritten as $p = (1/7, 2/7, 3/7, 0, 1/7, 0, 0)$ and $q = (3/8, 2/8, 2/8, 0, 1/8, 0, 0)$, after permuting the first and third indices and the second and sixth indices. We obtain the following.

n	$\frac{1}{n}D(p^{(n)}\|q^{(n)})$
10	0.05323
50	0.03626
100	0.03415

By Theorem 2, the Kullback-Leibler divergence rate is equal to 0.032. Clearly, as n gets larger, $\frac{1}{n}D(p^{(n)}\|q^{(n)})$ is closer to the Kullback-Leibler divergence rate. We also obtain the following.

n	$\frac{1}{n}H(p^{(n)})$
10	0.54366
50	0.50877
100	0.50442

By Corollary 2, the Shannon entropy rate is equal to 0.50008. Similarly, as n gets larger, the value of $\frac{1}{n}H(p^{(n)})$ moves closer to the Shannon entropy 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. Note also that both $p^{(n)}$ and $q^{(n)}$ are not stationary. 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 increases, $\frac{1}{n}D(p^{(n)}\|q^{(n)})$ gets closer to the Kullback-Leibler divergence rate. We also obtain the following.

n	$\frac{1}{n}H(p^{(n)})$
10	0.4618
50	0.4175
100	0.4116

By Corollary 2, the Shannon entropy rate is equal to 0.4057. Similarly, $\frac{1}{n}H(p^{(n)})$ approaches the Shannon entropy rate with increasing n.

6 Conclusion

In this work, we derived a formula for the Kullback-Leibler divergence rate between two time-invariant finite-alphabet Markov sources of arbitrary order and arbitrary initial distributions. We also investigated its rate of convergence. Similarly, we examined the computation and the existence of the Shannon entropy rate for Markov sources and investigated its rate of convergence. The main tools used in obtaining these results are the theory of non-negative matrices and Perron-Frobenius theory. One interesting and challenging direction for future work is the investigation of the Kullback-Leibler divergence rate for general hidden Markov sources.

References

[1] M. B. Bassat, "f-entropies, probability of error, and feature selection," *Information and Control*, 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 Transactions on Information Theory*, vol. 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] M. N. Do, "Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models," *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 115-118, April 2003.

[6] R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, 1968.

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

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

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

[10] T. T. Kadota and L. A. Shepp, "On the best finite set of linear observables for discriminating two Gaussian signals," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 278-284, Apr. 1967.

[11] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52-60, Feb. 1967.

[12] D. Kazakos and T. Cotsidas, "A decision theory approach to the approximation of discrete probability densities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 1, pp. 61-67, Jan. 1980.

[13] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79-86, 1951.

[14] K. Marton and P. C. Shields, "The positive-divergence and blowing-up properties," *Israel Journal of Mathematics*, vol. 86, 331-348, 1994.

[15] Z. Rached, F. Alajaji and L. L. Campbell, "Rényi's divergence and entropy rates for finite alphabet Markov sources," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1553-1561, May 2001.

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

[17] P. C. Shields, "Two divergence-rate counterexamples," *Journal of Theoretical Probability*, vol. 6, 521-545, 1993.

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