[12] R. A. Rueppel, "Linear complexity and random sequences," in *Proc. Eurocrypt'85 (Lecture Notes in Computer Science)*. Berlin, Germany: Springer-Verlag, 1985, vol. 219, pp. 167–188.

[13] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*. Amsterdam, The Netherlands: North-Holland, 1986.

[14] B. M. M. de Weger, "Approximation lattices of p -adic numbers," *J. Num. Theory*, vol. 24, pp. 70–88, 1986.

[15] J. Xu and A. Klapper, "Feedback with carry shift registers over $z/(n)$," in *Proc. SETA'98*. New York: Springer-Verlag, 1998.

The Kullback–Leibler Divergence Rate Between Markov Sources

Ziad Rached, *Student Member, IEEE*,
Fady Alajaji, *Senior Member, IEEE*, and
L. Lorne Campbell, *Life Fellow, IEEE*

***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 nonnegative 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*—Classifcation, decision theory, Kullback–Leibler divergence rate, nonnegative matrices, pattern recognition, Perron–Frobenius theory, rate of convergence, Shannon entropy rate, time-invariant Markov sources.**

I. Introduction

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

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

so that

$$\begin{aligned} p^{(n)}(i^n) &:= Pr\{X_1 = i_1, \dots, X_n = i_n\} \\ &= p_{i_1} p_{i_1 i_2} \cdots p_{i_{n-1} i_n}, \qquad i_1, \dots, i_n \in \mathcal{X} \end{aligned}$$

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 $p = (p_1, \dots, p_M)$ and $q = (q_1, \dots, q_M)$ 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

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

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 earlier 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.

Here, we 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 nonnegative 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 nonnegative matrices are first briefly presented in Section II. In Section III, 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 (nonstationary in general) Markov sources is briefly addressed in Section IV. Numerical examples are presented in Section V, and conclusions are stated in Section VI.

II. Preliminaries

Matrices and vectors are *positive* if all their components are positive and *nonnegative* if all their components are nonnegative. Throughout, A denotes an $M \times M$ nonnegative matrix with elements a_{ij}. The ijth 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 $i \to j$.

Manuscript received July 19, 2001; revised November 13, 2003. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada. The material in this correspondence was presented in part at the Conference on Information Sciences and Systems, Princeton, NJ, March 2002.

Z. Rached was with the Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada. He is now with the Department of Mathematics and Statistics, Notre Dame University, Zouk Mosbeh, Keserouan, P. O. Box 72 Zouk Mikael, Lebanon.

F. Alajaji and L. L. Campbell are with the Department of Mathematics and Statistics, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: fady@mast.queensu.ca).

Communicated by İ. E. Telatar, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2004.826687

With these definitions, it is possible to partition the set of indexes $\{1, 2, \dots, M\}$ into disjoint sets, called *classes*. All essential indexes can be subdivided into *essential classes* in such a way that all the indexes belonging to one class communicate, but cannot lead to an index outside the class. Moreover, all inessential indexes (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 indexes 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 indexes form a single essential class; i.e., if every index communicates with every other index.

Proposition 1 [16, p. 15]: By renumbering the indexes (i.e., by performing row and column permutations), it is possible to put matrix A in the *canonical form*

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

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

Proposition 2 [9, p. 492]: Suppose A is irreducible and let R_i, $i = 1, \dots, M$ denote the sum of the ith row. Also, let

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

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 nonnegative 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

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

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 $i = 1, 2, \dots, h$, 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.

III. Kullback–Leibler Divergence Rate

A. 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 $(p \ll q)$ and that P is absolutely continuous with respect to Q $(P \ll Q)$; 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 $p^{(n)} \ll q^{(n)}$ 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 $\{X_1, X_2, \dots\}$ 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 $\pi = (\pi_1, \dots, \pi_M)$ 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 $P \ll Q$. Furthermore, since both $p \ll q$ and $P \ll Q$ hold, we have after expanding the logarithm that

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

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,

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

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. □

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 submatrices corresponding to essential classes and let P_i, $i = h + 1, \ldots, g$ be the submatrices 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

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

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,

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

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 $P_{i1}, P_{i2}, \ldots, P_{ii-1}$ is nonzero 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 $1/n$. □

B. kth-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 $n \ge 1$, 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

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

where

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

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

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

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 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

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

Therefore,

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

We have the following corollaries.

Corollary 1: Suppose that the Markov source $\{X_1, X_2, \dots\}$ 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 $\pi = (\pi_1, \dots, \pi_M)$ is the unique stationary distribution of P.

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

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

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

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

Proof: Note that P^i, $i = 1, 2, \dots$ 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,

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

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). □

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 $1/n$.

V. 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 $\{X_1, X_2, \dots\}$ (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 two essential classes, one inessential self-communicating class, and one inessential non-self-communicating class. Accordingly, the initial distributions are rewritten as

$$p = (1/7, 2/7, 3/7, 0, 1/7, 0, 0)$$

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

and

$$q = (3/8, 2/8, 2/8, 0, 1/8, 0, 0)$$

after permuting the first and third indexes and the second and sixth indexes. 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 two-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 indexes $\{1, 2, 3\}$ forms an essential class, while the singleton set $\{4\}$ forms a self-communicating nonessential 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 0.3301. 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.

VI. 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 nonnegative 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," *Inform. Contr.*, vol. 39, pp. 227–242, 1978.

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

[3] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 462–467, May 1968.

[4] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*. London, U.K.: 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 Lett.*, vol. 10, pp. 115–118, Apr. 2003.

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

[7] ——, *Discrete Stochastic Processes*. Boston, MA: Kluwer, 1996.

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

[9] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1985.

[10] T. T. Kadota and L. A. Shepp, "On the best finite set of linear observables for discriminating two Gaussian signals," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 278–284, Apr. 1967.

[11] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Trans. Commun. Technol.*, vol. COM-15, pp. 52–60, Feb. 1967.

[12] 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, pp. 61–67, Jan. 1980.

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

[14] K. Marton and P. C. Shields, "The positive-divergence and blowing-up properties," *Israel J. Math.*, vol. 86, pp. 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*. New York: Springer-Verlag, 1981.

[17] P. C. Shields, "Two divergence-rate counterexamples," *J. Theor. Probab.*, vol. 6, pp. 521–545, 1993.

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