International Symposium on Information Theory and Its Applications
Honolulu, Hawaii, U.S.A., November 5–8, 2000

On the Rényi Divergence Rate for Finite Alphabet Markov Sources

Ziad Rached, Fady Alajaji and L. L. Campbell

Dept. of Mathematics & Statistics
Queen's University
Kingston, ON K7L 3N6, Canada
Email: rachedz@mast.queensu.ca

Abstract

In this work, we examine the existence and the computation of the Rényi divergence rate, $\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$, $0<\alpha<1$ between two time-invariant finite-alphabet Markov sources of arbitrary order and arbitrary initial distributions described under the probability distributions $p^{(n)}$ and $q^{(n)}$, respectively. This yields a generalization of the result of Nemetz where he assumed that the initial probabilities under $p^{(n)}$ and $q^{(n)}$ are strictly positive. The main tools used to obtain the Rényi divergence rate result are the theory of non-negative matrices and Perron-Frobenius theory. We also investigate the limits of the Rényi divergence rate as $\alpha\to 1$ and as $\alpha\to 0$.

Index Terms: Shannon theory, time-invariant Markov sources, Rényi's divergence rate, non-negative matrices, Perron-Frobenius theory.

1. Introduction

Without loss of generality, we will deal with first-order Markov sources since any k-th order Markov source can be converted to a first-order Markov source by k-step blocking it. Throughout, $\{X_1,X_2,\ldots\}$ denotes 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,

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

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

$$\begin{aligned} p^{(n)}(i^n) &=: Pr\{X_1=i_1,\ldots,X_n=i_n\} \\ &= p_{i_1}p_{i_1i_2}\cdots p_{i_{n-1}i_n}, \quad i_1,\ldots,i_n\in\mathcal{X}, \end{aligned}$$

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 $p=(p_1,\ldots,p_M)$ and $q=(q_1,\ldots,q_M)$ denote the initial distributions under $p^{(n)}$ and $q^{(n)}$ respectively.

The Rényi divergence [8] of order α between two distributions $\hat{p}$ and $\hat{q}$ defined on $\mathcal{X}$ is given by

$$D_\alpha(\hat{p}\|\hat{q})=\frac{1}{\alpha-1}\log\left(\sum_{i\in\mathcal{X}}\hat{p}_i^\alpha\hat{q}_i^{1-\alpha}\right),$$

where $0<\alpha<1$. The base of the logarithm is arbitrary. As $\alpha\to 1$, the Rényi divergence approaches the Kullback-Leibler divergence (relative entropy) given by

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

The Rényi divergence was originally introduced for the analysis of memoryless sources. One natural direction for further studies is the investigation of the Rényi divergence rate

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

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

$$D_\alpha(p^{(n)}\|q^{(n)})=\frac{1}{\alpha-1}\log\left(\sum_{x^n\in\mathcal{X}^n}[p^{(n)}(x^n)]^\alpha[q^{(n)}(x^n)]^{1-\alpha}\right)$$

for sources with memory, in particular, Markov sources. Nemetz addressed this problem in [5], where he evaluated the Rényi divergence rate $\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$ between two Markov sources characterized by $p^{(n)}$ and $q^{(n)}$, respectively, under the restriction that the initial probabilities p and q are

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

strictly positive (i.e., all p_i's and q_i's are strictly positive).

In this work, we generalize the Nemetz result by establishing a computable expression for the Rényi divergence rate between Markov sources with *arbitrary* initial distributions. We also investigate the questions of whether the Rényi divergence rate reduces to the Kullback-Leibler divergence rate as $\alpha \to 1$ and the interchangeability of limits between n and α as $n \to \infty$ and as $\alpha \to 0$. To the best of our knowledge, these issues have not been addressed before. We provide sufficient (but not necessary) conditions on the underlying Markov source distributions $p^{(n)}$ and $q^{(n)}$ for which the interchangeability of limits as $n \to \infty$ and $\alpha \to 1$ is valid. We also provide a counterexample where the interchangeability of limits as $n \to \infty$ and $\alpha \to 1$ does not hold. We also show that the interchangeability of limits as $n \to \infty$ and $\alpha \to 0$ always hold.

The Rényi divergence rate has played a significant role in certain hypothesis testing questions [3, 5]. Before stating our main results, we recall some facts about non-negative matrices which may be found in [9, Chapter 1].

2. Non-negative matrices

Matrices and vectors are *positive* if all their components are positive and *non-negative* if all their components are non-negative. Let A denotes an $M \times M$ non-negative matrix ($A \geq 0$) 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* (*transient*). An index which leads to no index at all (this arises when A has a row of zeros) is also called inessential. Otherwise, the index i is called *essential* (*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 (if any) 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 By renumbering the indices (i.e., by performing row and column permutations), it is possible to put a non-negative matrix A in the *canonical form*

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

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 corresponds to the essential class C_i, $i = 1, \ldots, h$, while the matrix A_i correspond to the inessential class C_i, $i = h+1, \ldots, g$. 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.

A class C_j is *reachable* from another class C_i if $A_{ij} \neq 0$, or if for some $i_1, \ldots, i_c$, $A_{ii_1} \neq 0$,$A_{i_1 i_2} \neq 0$,$\ldots$,$A_{i_c, j} \neq 0$, where c is at most $l-1$ (since there are l classes). Thus, c can be viewed as the number of steps needed to reach class C_j starting from class C_i. Note that from the canonical form of A, the class C_j is reachable from class C_i if $A_{ij}^{(c)} \neq 0$ for some $c = 1, \ldots, l-1$, where $A_{ij}^{(c)}$ is the ij-th submatrix of A^c.

Proposition 2 If a non-negative matrix A is irreducible, then A has a real positive eigenvalue λ that is greater than or equal to the magnitude of each other eigenvalue. There is a positive left (right) eigenvector, $\mathbf{a}$ ($\mathbf{b}$), corresponding to λ, where $\mathbf{a}$ is a row vector and $\mathbf{b}$ is a column vector.

3. Main results

Define a new matrix $R = (r_{ij})$ by

$$r_{ij} = p_{ij}^{\alpha} q_{ij}^{1-\alpha}, \quad i, j = 1, \ldots, M.$$

Also, define two new $1 \times M$ vectors $\mathbf{s} = (s_1, \ldots, s_M)$ and $\mathbf{1}$ by

$$s_i = p_i^{\alpha} q_i^{1-\alpha}, \quad \mathbf{1} = (1, \ldots, 1).$$

Then clearly $D_\alpha(p^{(n)}\|q^{(n)})$ can be written as

$$D_\alpha(p^{(n)}\|q^{(n)}) = \frac{1}{\alpha-1}\log \mathbf{s}R^{n-1}\mathbf{1}^t, \tag{1}$$

where $\mathbf{1}^t$ denotes the transpose of the vector $\mathbf{1}$. Without loss of generality, we will herein assume that there exists at least one $i \in \{1,\ldots,M\}$ for which $s_i > 0$, because otherwise, $D_\alpha(p^{(n)}\|q^{(n)})$ is infinite. We have the following lemma.

Lemma 1 If the matrix R is irreducible, then the Rényi divergence rate between $p^{(n)}$ and $q^{(n)}$ is given by

$$\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) = \frac{1}{\alpha-1}\log\lambda,$$

where λ is the largest positive real eigenvalue of R.

Proof: By Proposition 2, let λ be the largest positive real eigenvalue of R with associated positive right eigenvector $\mathbf{b} > 0$. Then

$$R^{n-1}\mathbf{b} = \lambda^{n-1}\mathbf{b}. \tag{2}$$

Let $R^{n-1} = (r_{ij}^{(n-1)})$ and $\mathbf{b}^t = (b_1, b_2, \ldots, b_M)$. Also, let $b_L = \min_{1\le i\le M}(b_i)$ and $b_U = \max_{1\le i\le M}(b_i)$. Thus $0 < b_L \le b_i \le b_U$ $\forall i$. Let $R^{n-1}\mathbf{1}^t = \mathbf{y}^t$ where $\mathbf{y} = (y_1,\ldots,y_M)$. Then, by (2)

$$\lambda^{n-1}b_i = \sum_{j=1}^{M} r_{ij}^{(n-1)}b_j \le \sum_{j=1}^{M} r_{ij}^{(n-1)}b_U = b_U y_i.$$

Similarly, it can be shown that $\lambda^{n-1}b_i \ge b_L y_i$, $\forall i = 1,\ldots,M$. Therefore

$$\frac{b_i}{b_U} \le \frac{y_i}{\lambda^{n-1}} \le \frac{b_i}{b_L}, \quad \forall i = 1,\ldots,M. \tag{3}$$

Since $\mathbf{s}R^{n-1}\mathbf{1}^t = \sum_{i=1}^{M} s_i y_i$, it follows directly from (3) that

$$\frac{\sum_i s_i b_i}{b_U} \le \frac{\mathbf{s}R^{n-1}\mathbf{1}^t}{\lambda^{n-1}} \le \frac{\sum_i s_i b_i}{b_L},$$

or

$$\begin{aligned}\frac{1}{n}\log\left(\frac{\sum_i s_i b_i}{b_U}\right) &\le \frac{1}{n}\log\left(\frac{\mathbf{s}R^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right)\\ &\le \frac{1}{n}\log\left(\frac{\sum_i s_i b_i}{b_L}\right).\end{aligned}$$

Note that s_i, b_i, b_U, b_L do not depend on n. Therefore,

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{\mathbf{s}R^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right) = 0,$$

since it is sandwiched between two expressions that approaches 0 as $n\to\infty$. Hence

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\left(\mathbf{s}R^{n-1}\mathbf{1}^t\right) &= \lim_{n\to\infty}\frac{1}{n}\log\lambda^{n-1}\\ &\quad+\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{\mathbf{s}R^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right)\\ &= \log\lambda,\end{aligned}$$

and thus

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) &= \lim_{n\to\infty}\frac{\log\left(\mathbf{s}R^{n-1}\mathbf{1}^t\right)}{n(\alpha-1)}\\ &= \frac{1}{\alpha-1}\log\lambda.\end{aligned}$$

$\square$

Using Lemma 1 and Proposition 1, we obtain the following general result.

Theorem 1 Let R_i, $i = 1,\ldots,g$, be the irreducible matrices along the diagonal of the canonical form of the matrix R as shown in Proposition 1. Write the vector $\mathbf{s}$ as

$$\mathbf{s} = (\tilde{s}_1,\ldots,\tilde{s}_h,\tilde{s}_{h+1},\ldots,\tilde{s}_g,s_{g+1},\ldots,s_l),$$

where the vector $\tilde{s}_i$ corresponds to R_i, $i = 1,\ldots,g$. The scalars $s_{g+1},\ldots,s_l$ correspond to non self-communicating classes.

- Let λ_k be the largest positive real eigenvalue of R_k for which the corresponding vector $\tilde{s}_k$ is different from the zero vector, $k = 1,\ldots,g$. Let $\lambda^\star$ be the maximum over these λ_k's. If $\tilde{s}_k = 0$, $\forall k = 1,\ldots,g$, then let $\lambda^\star = 0$.
- For each inessential class C_i with corresponding vector $\tilde{s}_i \ne 0$, $i = h+1,\ldots,g$ or corresponding scalar $s_i \ne 0$, $i = g+1,\ldots,l$, let λ_j be the largest positive real eigenvalue of R_j if class C_j is reachable from class C_i. Let $\lambda^\dagger$ be the maximum over these λ_j's. If $\tilde{s}_i = 0$ and $s_i = 0$ for every inessential class C_i, then let $\lambda^\dagger = 0$.

Let $\lambda = \max\{\lambda^\star,\lambda^\dagger\}$. Then the Rényi divergence rate is given by

$$\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) = \frac{1}{\alpha-1}\log\lambda.$$

Proof: Cf. [6].

Remark: In [5], Nemetz showed that the Rényi divergence rate between two time-invariant Markov sources with *strictly positive* initial distributions is given by

$\frac{1}{\alpha-1}\log\lambda$ where λ is the largest positive real eigenvalue of R. Nemetz also pointed out that this assumption could be replaced by other conditions, although he did not provide them. Note that by Theorem 1, the Rényi divergence rate between two-time invariant Markov sources with *arbitrary* initial distributions is not necessarily equal to $\frac{1}{\alpha-1}\log\lambda$, where λ is the largest positive real eigenvalue of R. However, if the initial distributions are strictly positive, which implies directly that $\mathbf{s} > 0$, then Theorem 1 reduces to the Nemetz result. This follows directly from the fact that, in this case, $\lambda^\star = \max\{\lambda_k\}$, $k = 1, \ldots, g$, and the fact that the determinant of a block lower triangular matrix is equal to the product of the determinants of the sub-matrices along the diagonal.

We also have the following results about the interchangeability of limits as $\alpha \to 1$ and as $\alpha \to 0$.

Theorem 2 [6] Let P and Q be the probability transition matrices on $\mathcal{X}$ associated with $p^{(n)}$ and $q^{(n)}$ respectively. If the matrix P is irreducible, the matrix Q is positive, and the initial distribution q under $q^{(n)}$ is positive, then

$$\lim_{\alpha\to 1}\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) = \lim_{n\to\infty}\lim_{\alpha\to 1}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}),$$

and therefore, the Rényi divergence rate reduces to the Kullback-Leibler divergence rate as $\alpha \to 1$.

In the following example, we show that the interchangeability of limits does not necessarily hold if the conditions of the theorem are not satisfied.

Example: Let P and Q be the following:

$$P = \begin{pmatrix} 1/4 & 3/4 & 0 \\ 3/4 & 1/4 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad Q = \begin{pmatrix} 3/4 & 1/4 & 0 \\ 1/4 & 3/4 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Suppose that $p^{(n)}$ is stationary with stationary distribution $(b/2, b/2, 1-b)$, where $0 < b < 1$ is arbitrary. Also, suppose that the initial distribution under $q^{(n)}$ is positive. A simple computation [2, p. 40] yields that the Kullback-Leibler divergence rate is given by $(b\log 3)/2$.

The eigenvalues of R are: $\lambda_1 = (3^\alpha + 3^{1-\alpha})/4$, $\lambda_2 = (3^{1-\alpha} - 3^\alpha)/4$, and $\lambda_3 = 1$. Note that $\mathbf{s} > 0$ and that, since $0 < \alpha < 1$, $\max_{1\le i\le 3}\{\lambda_i\} = 1$. By Theorem 1, the Rényi divergence rate is 0.

Therefore, the interchangeability of limits is not valid, i.e.,

$$\lim_{\alpha\to 1}\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) \neq \lim_{n\to\infty}\lim_{\alpha\to 1}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}).$$

The reason behind this inequality is that $\max_{1\le i\le 3}\{\lambda_i\}$ is not differentiable at $\alpha = 1$ [4, p. 371] because, at $\alpha = 1$, $\lambda = 1$ is a double eigenvalue.

Theorem 3 [6] The interchangeability of limits as $n \to \infty$ and as $\alpha \to 0$ is always valid; i.e.,

$$\lim_{\alpha\to 0}\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) = \lim_{n\to\infty}\lim_{\alpha\to 0}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$$

4. Concluding remarks

In this work, we derived a formula for the Rényi divergence rate between two time-invariant finite alphabet Markov sources of arbitrary order and arbitrary initial distributions. We also investigated the limits of the Rényi divergence rate as $\alpha \to 1$ and as $\alpha \to 0$. Numerical examples were presented. Finally, we would like to point out that if $q^{(n)}$ is stationary memoryless with uniform marginal distribution then for any $\alpha > 0$, $\alpha \neq 1$,

$$D_\alpha(p^{(n)}\|q^{(n)}) = n\log M - H_\alpha(p^{(n)}).$$

Hence, the existence and the computation of the Rényi entropy rate follows directly from Theorem 1. An important application of this result is the extension of the variable-length source coding theorem in [1] and [7] to time-invariant Markov sources.

References

[1] L. L. Campbell, "A coding theorem and Rényi's entropy," *Information and Control*, vol. 8, pp. 423–429, 1965.

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

[3] L. H. Koopmans, "Asymptotic rate of discrimination for Markov processes," *Ann. Math. Stat.*, vol. 31, pp. 982–994, 1960.

[4] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, 2nd edition, Academic, Toronto, 1985.

[5] T. Nemetz, "On the α-divergence rate for Markov-dependent hypotheses," *Problems of Control and Information Theory,* vol. 3 (2), pp. 147–155, 1974.

[6] Z. Rached, F. Alajaji, and L. L. Campbell, "Rényi's Divergence and Entropy Rates for Finite Alphabet Markov Sources," submitted to *IEEE Transactions on Information Theory*, March 2000.

[7] Z. Rached F. Alajaji and L. L. Campbell, "Rényi's entropy rate for discrete Markov sources," *Proc. CISS'99*, March 17–19, Baltimore, MD, 1999.

[8] A. Rényi, "On measures of entropy and information," *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1**, 547–561, Univ. of California Press, Berkeley, 1961.

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