# Rényi's Divergence and Entropy Rates for Finite Alphabet Markov Sources

Ziad Rached Fady Alajaji L. L. Campbell *

**Abstract**

In this work, we examine the existence and the computation of the Rényi divergence rate, $\lim_{n\to\infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)})$, 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. This yields a generalization of a 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 are the theory of non-negative matrices and Perron-Frobenius theory. We also provide numerical examples and investigate the limits of the Rényi divergence rate as $\alpha \to 1$ and as $\alpha \downarrow 0$. Similarly, we provide a formula for the Rényi entropy rate $\lim_{n\to\infty} \frac{1}{n} H_\alpha(p^{(n)})$ of Markov sources and examine its limits as $\alpha \to 1$ and as $\alpha \downarrow 0$. Finally, we briefly provide an application to source coding.

*Index Terms*: Time-invariant Markov sources, Rényi's divergence and entropy rates, non-negative matrices, Perron-Frobenius theory, Kullback-Leibler divergence rate, Shannon and Hartley entropy rates.

---

*This research was supported in part by the Natural Sciences and Engineering Research Council of Canada. Parts of this work will be presented at the International Symposium on Information Theory and its Applications, Honolulu, Hawaii, November 2000.

The authors 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 Rényi divergence [20] of order $\alpha$ 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 $0 < \alpha < 1$. This definition can be extended to $\alpha > 1$ if all $\hat{q}_i > 0$. The base of the logarithm is arbitrary. Similarly, the Rényi entropy of order $\alpha$ for $\hat{p}$ is defined as

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

where $\alpha > 0$ and $\alpha \neq 1$. As $\alpha \to 1$, the Rényi divergence approaches the Kullback-Leibler divergence (relative entropy) given by

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

and the Rényi entropy approaches the Shannon entropy.

The above generalized information measures and their subsequent variations [23] were originally introduced for the analysis of memoryless sources. One natural direction for further studies is the investigation of the Rényi divergence rate

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

where

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

and of the Rényi entropy rate

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

where

$$H_\alpha(p^{(n)}) = \frac{1}{1-\alpha} \log \left( \sum_{i^n \in \mathcal{X}^n} [p^{(n)}(i^n)]^\alpha \right),$$

for sources with memory, in particular Markov sources. Nemetz addressed these problems in [16], 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 strictly positive (i.e., all $p_i$'s and $q_i$'s are strictly positive).

The Rényi divergence rate has played a significant role in certain hypothesis testing questions [14, 16, 17]. Furthermore, the Rényi entropy and the Rényi entropy rate have revealed several operational characterizations in the problem of fixed-length source coding [7, 6], variable-length source coding [4, 5, 13, 19], error exponent calculations [8], and other areas [1, 2, 3, 18].

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 $\alpha$ as $n \to \infty$ and as $\alpha \downarrow 0$. 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 as $\alpha \to 1$ is valid. We also give an example of non-interchangeability of limits as $n \to \infty$ and as $\alpha \to 1$. We also show that the interchangeability of limits as $n \to \infty$ and $\alpha \downarrow 0$ always holds. We next address the computation and the existence of the Rényi entropy rate $\lim_{n\to\infty} \frac{1}{n} H_\alpha(p^{(n)})$ for a Markov source with distribution $p^{(n)}$ and examine its limits as $\alpha \downarrow 0$ and as $\alpha \to 1$.

The rest of this paper is organized as follows. In the following section, we review some definitions and relevant results from the theory of non-negative matrices and Perron-Frobenius theory. In Section 3, we provide a general formula for the Rényi divergence rate between $p^{(n)}$ and $q^{(n)}$, with no restriction on the initial probabilities $p$ and $q$, and illustrate it numerically. The result is first proved for first-order Markov sources, and is then extended for Markov sources of arbitrary order. In Section 4, we show that if the probability transition matrix $P$ associated with the Markov source under $p^{(n)}$ is irreducible and if both the initial probability $q$ and the probability transition matrix $Q$ associated with the Markov source under $q^{(n)}$ are positive (with strictly positive entries), then the Rényi divergence rate reduces to the Kullback-Leibler divergence rate as $\alpha \to 1$. We also show that the interchangeability of limits as $n \to \infty$ and as $\alpha \downarrow 0$ is always valid. In Section 5, we address similar questions for the Rényi entropy rate and briefly illustrate it with an application to variable-length source coding. Finally, concluding remarks are stated in Section 6.

# 2 Non-negative matrices

We begin with some results about non-negative matrices. Most of what follows may be found in [22] and [9].

Matrices and vectors are *positive* if all their components are positive and *non-negative* if all their components are non-negative. Throughout this section, $A$ 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* (or *transient*). An index which leads to no index at all (this arises when $A$ has a row of zeros) is also called inessential. Also, an index that is not inessential 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 (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*

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

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 sub-matrices which correspond to non self-communicating classes $C_i$, $i = g+1, \dots, l$, are $1 \times 1$ zero matrices. In every row $i = g+1, \dots, l$ any of the matrices $A_{i1}, \dots, A_{ii-1}$ may be zero.

A class $C_j$ is *reachable* from another class $C_i$ if $A_{ij} \neq 0$, or if for some $i_1, \dots, i_c$, $A_{ii_1} \neq 0$,$A_{i_1 i_2} \neq 0$, … ,$A_{i_c,j} \neq 0$, 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 $c = 1, \dots, l-1$, where $A_{ij}^{(c)}$ is the $ij$-th submatrix of $A^c$.

**Proposition 2 (Frobenius)** If $A$ is irreducible, then $A$ has a real positive eigenvalue $\lambda$ 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 $\lambda$, where $\mathbf{a}$ is a row vector and $\mathbf{b}$ is a column vector.

**Proposition 3 [12, p. 508]** If $A$ is irreducible, then the largest positive real eigenvalue has algebraic multiplicity 1.

**Proposition 4 [12, p. 494]** If $A$ has a positive eigenvector $x$, then for all $m = 1, 2, \dots$ and for all $i = 1, \dots, M$ we have

$$\sum_{j=1}^{M} a_{ij}^{(m)} \leq \left[ \frac{\max_{1 \leq k \leq M} x_k}{\min_{1 \leq k \leq M} x_k} \right] \rho^m(A),$$

where $A^m = (a_{ij}^{(m)})$ and $\rho(A) \triangleq \max\{|\lambda| : \lambda \text{ eigenvalue of } A\}$ is the spectral radius of $A$.

The following corollary follows directly from the previous proposition by observing that, $a_{ij}^{(m)} \leq \sum_{j=1}^{M} a_{ij}^{(m)}$ $\forall$ $i = 1, \ldots, M$ and $j = 1, \ldots, M$.

**Corollary 1** If $A$ is irreducible, then $A^m \leq \lambda^m C$ (i.e., $a_{ij}^{(m)} \leq \lambda^m c_{ij}$), for all $m = 1, 2, \ldots$, where $\lambda$ is the largest positive real eigenvalue of $A$ and $C = (\frac{\max_{1\leq k\leq M} x_k}{\min_{1\leq k\leq M} x_k})$ is a matrix with identical entries that are independent of $m$.

**Proposition 5 [15, p. 371]** The eigenvalues of a matrix are continuous functions of the entries of the matrix.

**Proposition 6 [15, p. 396]** Let $A(\alpha)$ be an $M \times M$ matrix whose entries are all analytic functions of $\alpha$ in some neighborhood of $\alpha_0$. Let $\lambda$ be an eigenvalue of $A(\alpha_0)$ of algebraic multiplicity 1. Then $A(\alpha)$ has an eigenvalue $\lambda(\alpha)$ which is an analytic function in the neighborhood of $\alpha_0$ and for which $\lambda(\alpha_0) = \lambda$.

# 3 The Rényi divergence rate

## 3.1 First-order Markov sources

We assume first that the Markov source $\{X_1, X_2, \ldots\}$ is of order one. Later, we generalize the results for an arbitrary order $k$. With the same notation, as presented at the beginning of the introduction, the joint distributions of the random variables $(X_1, \ldots, X_n)$ under $p^{(n)}$ and $q^{(n)}$ are given respectively by

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

and

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

Let

$$V(n, \alpha) = \sum_{i^n \in \mathcal{X}^n} [p^{(n)}(i^n)]^{\alpha} [q^{(n)}(i^n)]^{1-\alpha}.$$

Then

$$V(n, \alpha) = \sum p_{i_1}^{\alpha} q_{i_1}^{1-\alpha} p_{i_1 i_2}^{\alpha} q_{i_1 i_2}^{1-\alpha} \cdots p_{i_{n-1} i_n}^{\alpha} q_{i_{n-1} i_n}^{1-\alpha},$$

where the sum is over $i_1, \ldots, i_n \in \mathcal{X}$. 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 (i.e., if $s_i = 0 \ \forall i$), $D_\alpha(p^{(n)} \| q^{(n)})$ is infinite. We also assume that $0 < \alpha < 1$; we can allow the case of $\alpha > 1$ if $q > 0$ and $Q > 0$. Before stating our first main theorem, we prove 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 $\lambda$ is the largest positive real eigenvalue of $R$, and $0 < \alpha < 1$. Furthermore, the same result holds for $\alpha > 1$ if $q > 0$ and $Q > 0$.

**Proof:** By Proposition 2, let $\lambda$ 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)

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

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

$$\frac{1}{n}\log\left(\frac{\sum_i s_i b_i}{b_U}\right) \le \frac{1}{n}\log\left(\frac{\mathbf{s}R^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right) \le \frac{1}{n}\log\left(\frac{\sum_i s_i b_i}{b_L}\right). \tag{4}$$

Note that $s_i, b_i, b_U, b_L$ do not depend on $n$. Therefore, by (4),

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

since it is upper and lower bounded by two quantities that approach 0 as $n \to \infty$. Hence

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

and thus

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

□

We next use Lemma 1 and the canonical form of $R$ to prove the following general result.

**Theorem 1** Let $R_i$, $i = 1, \dots, g$, 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

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

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

- Let $\lambda_k$ be the largest positive real eigenvalue of $R_k$ for which the corresponding vector $\tilde{s}_k$ is different from the zero vector, $k = 1, \dots, g$. Let $\lambda^\star$ be the maximum over these $\lambda_k$'s. If $\tilde{s}_k = 0$, $\forall k = 1, \dots, g$, then let $\lambda^\star = 0$.

- For each inessential class $C_i$ with corresponding vector $\tilde{s}_i \neq 0$, $i = h+1, \dots, g$ or corresponding scalar $s_i \neq 0$, $i = g+1, \dots, l$, let $\lambda_j$ 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 $\lambda_j$'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

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

where $0 < \alpha < 1$. Furthermore, the same result holds for $\alpha > 1$ if $q > 0$ and $Q > 0$.

**Proof**: By Proposition 2, let $\lambda_i$ be the largest positive real eigenvalue of $R_i$ with associated positive right eigenvector $\tilde{b}_i > 0$, $i = 1, \dots, g$. Let

$$\mathbf{b}^t = (\tilde{b}_1, \dots, \tilde{b}_h, \tilde{b}_{h+1}, \dots, \tilde{b}_g, 0, \dots, 0),$$

where the zeros correspond to non self-communicating classes. By Proposition 1 we have that

$$R^{n-1} = \begin{bmatrix} R_1^{n-1} & \dots & 0 & 0 & \dots & 0 & \dots & \dots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & R_h^{n-1} & 0 & \dots & 0 & \dots & \dots & 0 \\ R_{h+11}^{(n-1)} & \dots & R_{h+1h}^{(n-1)} & R_{h+1}^{n-1} & \dots & 0 & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ R_{g1}^{(n-1)} & \dots & R_{gh}^{(n-1)} & R_{gh+1}^{(n-1)} & \dots & R_g^{n-1} & \dots & \dots & 0 \\ R_{g+11}^{(n-1)} & \dots & R_{g+1h}^{(n-1)} & R_{g+1h+1}^{(n-1)} & \dots & R_{g+1g}^{(n-1)} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ R_{l1}^{(n-1)} & \dots & R_{lh}^{(n-1)} & R_{lh+1}^{(n-1)} & \dots & R_{lg}^{(n-1)} & R_{lg+1}^{(n-1)} & \dots & 0 \end{bmatrix}.$$

Then

$$\begin{aligned} \mathbf{s}R^{n-1}\mathbf{b} &= \sum_{i=1}^{g} \tilde{s}_i R_i^{n-1} \tilde{b}_i + \sum_{i=h+1}^{g} \tilde{s}_i \left( R_{i1}^{(n-1)} \tilde{b}_1 + \dots + R_{ii-1}^{(n-1)} \tilde{b}_{i-1} \right) \\ &+ \sum_{i=g+1}^{l} s_i \left( R_{i1}^{(n-1)} \tilde{b}_1 + \dots + R_{ig}^{(n-1)} \tilde{b}_g \right). \end{aligned}$$

Rewrite the vector $\mathbf{1}$ as

$$\mathbf{1} = (\tilde{1}_1, \dots, \tilde{1}_h, \tilde{1}_{h+1}, \dots, \tilde{1}_g, 1, \dots, 1),$$

where $\tilde{1}_i$, $i = 1, \dots, g$ correspond to essential and inessential self-communicating classes and the 1's correspond to non self-communicating classes. Let $R^{n-1}\mathbf{1}^t = \mathbf{y}^t$ where

$$\mathbf{y} = (\tilde{y}_1, \dots, \tilde{y}_h, \tilde{z}_{h+1} + \tilde{y}_{h+1}, \dots, \tilde{z}_g + \tilde{y}_g, \tilde{z}_{g+1}, \dots, \tilde{z}_l),$$

and

$$\begin{aligned}
\tilde{y}_i &= R_i^{n-1}\tilde{1}_i^t, & i &= 1,\ldots,g,\\
\tilde{z}_i &= \textstyle\sum_{j=1}^{i-1} R_{ij}^{(n-1)}\tilde{1}_j^t, & i &= h+1,\ldots,g,\\
\tilde{z}_i &= \textstyle\sum_{j=1}^{g} R_{ij}^{(n-1)}\tilde{1}_j^t + \sum_{j=g+1}^{i-1} R_{ij}^{(n-1)}, & i &= g+1,\ldots,l.
\end{aligned}$$

Therefore

$$\mathbf{s}R^{n-1}\mathbf{1}^t = \sum_{i=1}^{g} \tilde{s}_i\tilde{y}_i + \sum_{i=h+1}^{g} \tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l} s_i\tilde{z}_i. \tag{5}$$

As in the proof of Lemma 1, since $R_i\tilde{b}_i = \lambda_i\tilde{b}_i$, we can write

$$R_i^{n-1}\tilde{b}_i = \lambda_i^{n-1}\tilde{b}_i \le b_U\tilde{y}_i, \quad i = 1,\ldots,g,$$

where $b_U = \max_{1\le i\le g}(b_{U_i})$ and $b_{U_i}$ is the largest component of $\tilde{b}_i$, $i = 1,\ldots,g$. Similarly,

$$R_i^{n-1}\tilde{b}_i = \lambda_i^{n-1}\tilde{b}_i \ge b_L\tilde{y}_i, \quad i = 1,\ldots,g,$$

where $b_L = \min_{1\le i\le g}(b_{L_i})$ and $b_{L_i}$ is the smallest component of $\tilde{b}_i$, $i = 1,\ldots,g$. Therefore

$$\frac{\lambda_i^{n-1}\tilde{b}_i}{b_U} \le \tilde{y}_i \le \frac{\lambda_i^{n-1}\tilde{b}_i}{b_L}, \quad i = 1,\ldots,g.$$

Hence

$$\frac{1}{b_U}\sum_{i=1}^{g} \tilde{s}_i\lambda_i^{n-1}\tilde{b}_i \le \sum_{i=1}^{g} \tilde{s}_i\tilde{y}_i \le \frac{1}{b_L}\sum_{i=1}^{g} \tilde{s}_i\lambda_i^{n-1}\tilde{b}_i.$$

Therefore, by (5)

$$\begin{aligned}
\frac{1}{b_U}\sum_{i=1}^{g} \tilde{s}_i\lambda_i^{n-1}\tilde{b}_i + \sum_{i=h+1}^{g} \tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l} s_i\tilde{z}_i &\le \mathbf{s}R^{n-1}\mathbf{1}^t\\
&\le \frac{1}{b_L}\sum_{i=1}^{g} \tilde{s}_i\lambda_i^{n-1}\tilde{b}_i + \sum_{i=h+1}^{g} \tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l} s_i\tilde{z}_i,
\end{aligned}$$

or

$$\begin{aligned}
&\frac{1}{n}\log\left(\frac{1}{b_U}\sum_{i=1}^{g} \tilde{s}_i\left(\frac{\lambda_i}{\lambda}\right)^{n-1}\tilde{b}_i + \frac{1}{\lambda^{n-1}}\left(\sum_{i=h+1}^{g} \tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l} s_i\tilde{z}_i\right)\right)\\
\le\ &\frac{1}{n}\log\left(\frac{\mathbf{s}R^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right)\\
\le\ &\frac{1}{n}\log\left(\frac{1}{b_L}\sum_{i=1}^{g} \tilde{s}_i\left(\frac{\lambda_i}{\lambda}\right)^{n-1}\tilde{b}_i + \frac{1}{\lambda^{n-1}}\left(\sum_{i=h+1}^{g} \tilde{s}_i\tilde{z}_i + \sum_{i=g+1}^{l} s_i\tilde{z}_i\right)\right),
\end{aligned}$$

where $\lambda$ is as defined in the statement of the theorem. To show that $\frac{1}{n}\log\left(\frac{\mathbf{s}R^{n-1}\mathbf{1}^t}{\lambda^{n-1}}\right)$ converges to 0 as $n\to\infty$, it is sufficient to prove that the lower and upper bounds converge to 0. Since the lower and upper bounds are within a constant scaling of each other, it is enough to show that the lower bound converges to 0 as $n\to\infty$. Note that

$$\sum_{i=h+1}^{g}\tilde{s}_i\tilde{z}_i+\sum_{i=g+1}^{l}s_i\tilde{z}_i=\sum_{i=h+1}^{g}\sum_{j=1}^{i-1}\tilde{s}_iR_{ij}^{(n-1)}\tilde{1}_j^t+\sum_{i=g+1}^{l}\sum_{j=1}^{g}s_iR_{ij}^{(n-1)}\tilde{1}_j^t+\sum_{i=g+1}^{l}\sum_{j=g+1}^{i-1}s_iR_{ij}^{(n-1)}.$$

If $R_{ij}^{(n-1)}\neq 0$ for some $n$, then class $C_j$ is reachable from class $C_i$ (it is enough to check for $n=2,\ldots,l$, since the number of classes is $l$). From the block form of $R$, if $R_{ij}^{(n-1)}\neq 0$, then it is a weighted sum involving products of powers of $R_i$ and $R_j$ (which are irreducible) and possibly some other sub-matrices (which are irreducible) along the diagonal [1] of $R$. By applying Corollary 1 to each of these irreducible sub-matrices if $\tilde{s}_i\neq 0$ or $s_i\neq 0$ (since $R_{ij}^{(n-1)}$ is multiplied by $\tilde{s}_i$ or $s_i$), the above expression is upper bounded by linear combinations of powers of the largest eigenvalues of the sub-matrices along the diagonal of $R$ for which $\tilde{s}_i\neq 0$, $i=h+1,\ldots,g$, or for which the corresponding class is reachable from class $C_i$, $i=g+1,\ldots,l$.

For example, in the case of the $R$ as given in the footnote, $R_{21}^{(n-1)}\leq(n-1)\lambda^{n-2}C$, where $C>0$ and its entries are independent of $n$. Hence

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{1}{b_U}\sum_{i=1}^{g}\tilde{s}_i\left(\frac{\lambda_i}{\lambda}\right)^{n-1}\tilde{b}_i+\frac{1}{\lambda^{n-1}}\left(\sum_{i=h+1}^{g}\tilde{s}_i\tilde{z}_i+\sum_{i=g+1}^{l}s_i\tilde{z}_i\right)\right)=0.$$

This follows from the fact that for large $n$, the argument of the logarithm is a polynomial expression of first degree in $n$, and $\lim_{n\to\infty}\frac{1}{n}\log(an+b)=0$.

If $R$ has three sub-matrices along the diagonal, then from the block form of $R$, the matrix $R_{31}^{(n-1)}$ depends recursively on a weighted sum involving $R_{21}$ and $R_{32}$. Therefore it is bounded by $\lambda^{n-2}p(n)D$, where $p(n)$ is a polynomial of second degree in $n$, and $D>0$ with entries independent of $n$.

In general, for large $n$, the argument of the logarithm is a polynomial expression in the variable $n$ of degree at most $l-1$ ($l$ is the number of classes), and hence it follows that

$$\lim_{n\to\infty}\frac{1}{n}\log\left(\frac{1}{b_U}\sum_{i=1}^{g}\tilde{s}_i\left(\frac{\lambda_i}{\lambda}\right)^{n-1}\tilde{b}_i+\frac{1}{\lambda^{n-1}}\left(\sum_{i=h+1}^{g}\tilde{s}_i\tilde{z}_i+\sum_{i=g+1}^{l}s_i\tilde{z}_i\right)\right)=0.$$

Therefore

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

[1] For example, if $R=\begin{bmatrix}R_1 & 0\\ R_{21} & R_2\end{bmatrix}$, then $R_{21}^{(n-1)}=\sum_{i=0}^{n-2}R_2^iR_{21}R_1^{n-i-2}$.

and thus

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

□

**Remark:** In [16], Nemetz showed that the Rényi divergence rate between two time-invariant Markov sources with *strictly positive* initial distributions is given by $\frac{1}{\alpha-1}\log\tilde{\lambda}$ where $\tilde{\lambda}$ 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 $\frac{1}{\alpha-1}\log\tilde{\lambda}$, where $\tilde{\lambda}$ 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, $\lambda = \lambda^\star = \max\{\lambda_k\}$, $k = 1, \ldots, g$, 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 (thus the largest eigenvalue of this matrix is given by $\max\{\lambda_k\}$).

## 3.2 Numerical examples

In this section, we use the natural logarithm. Let $P$ and $Q$ be two possible probability transition matrices for $\{X_1, X_2, \ldots\}$ defined as follows:

$$P = \begin{pmatrix} 1/4 & 3/4 & 0 & 0 & 0 \\ 1/3 & 2/3 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/5 & 4/5 & 0 \\ 0 & 1/6 & 1/2 & 0 & 1/3 \end{pmatrix}, \quad Q = \begin{pmatrix} 1/5 & 4/5 & 0 & 0 & 0 \\ 1/6 & 5/6 & 0 & 0 & 0 \\ 0 & 0 & 1/4 & 3/4 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/3 & 0 & 1/6 \end{pmatrix}.$$

Let the parameter $\alpha = 1/3$. The largest eigenvalues of the three sub-matrices along the diagonal of $R$ are respectively: $\lambda_1 = 0.98676$, $\lambda_2 = 0.95937$, and $\lambda_3 = 0.20998$. Let $p = (0, 0, 3/4, 1/4, 0)$ and $q = (0, 0, 1/3, 2/3, 0)$ be two possible initial distributions under $p^{(n)}$ and $q^{(n)}$ respectively. For these given initial distributions, we get by Theorem 1 that $\lambda^\star = \lambda_2$ and $\lambda^\dagger = 0$. Therefore, the Rényi divergence rate is $\ln(\lambda_2)/(\alpha - 1) = 0.06221$. Note that $\lambda_2$ is *not* the largest eigenvalue of $R$. We also obtain the following.

| $n$ | $\frac{1}{n}D_\alpha(p^{(n)} \| q^{(n)})$ |
|---|---|
| 1000 | 0.06227 |
| 2000 | 0.06224 |
| 3000 | 0.06223 |

Clearly, as $n$ gets large $\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$ is closer to the Rényi divergence rate. Note however that, in general, the function $\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$ is not monotonic in $n$.

Suppose that $\mathbf{s}$ has zero components on the first two classes. For example, let $p = (0, 1/4, 1/4, 0, 1/2)$ and $q = (1/4, 0, 0, 1/4, 1/2)$. In this case, $\lambda^\star = \lambda_3$, and $\lambda^\dagger = \max\{\lambda_1, \lambda_2\}$ (the first and second classes are reachable from the third). Therefore, the Rényi divergence rate is $\ln(\lambda_1)/(\alpha - 1) = 0.01999$. We also get the following.

| $n$ | $\frac{1}{n}D_\alpha(p^{(n)} \| q^{(n)})$ |
|---|---|
| 1000 | 0.02223 |
| 2000 | 0.02111 |
| 3000 | 0.02074 |

Clearly, as $n$ gets large $\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$ is closer to the Rényi divergence rate.

Suppose now that $\mathbf{s}$ has strictly positive components (as required in the Nemetz result). For example, let $p = (1/8, 1/4, 1/8, 1/4, 1/4)$ and $q = (1/10, 3/10, 2/10, 2/10, 2/10)$. In this case, $\lambda^\star = \lambda^\dagger = \max\{\lambda_1, \lambda_2, \lambda_3\} = \lambda_1$. Therefore, the Rényi divergence rate is $\ln(\lambda_1)/(\alpha - 1) = 0.01999$. Note that $\lambda_1$ is the largest eigenvalue of $R$ which is expected since the components of $\mathbf{s}$ are strictly positive. We also get the following.

| $n$ | $\frac{1}{n}D_\alpha(p^{(n)} \| q^{(n)})$ |
|---|---|
| 1000 | 0.02105 |
| 2000 | 0.02052 |
| 3000 | 0.02034 |

Clearly, as $n$ gets large $\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})$ is closer to the Rényi divergence rate.

## 3.3 $k$-th order Markov sources

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

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

Then

$$Pr(W_n = w_n | W_{n-1} = w_{n-1}, \ldots, W_1 = w_1) = Pr(W_n = w_n | W_{n-1} = w_{n-1}),$$

and $\{W_n\}$ is a first order Markov source with $M^k$ states. Let $p_{w_{n-1}w_n} \triangleq Pr(W_n = w_n|W_{n-1} = w_{n-1})$. We next write the joint distributions of $\{X_n\}$ in terms of the conditional probabilities of $\{W_n\}$. For $n \geq k$, $V(n,\alpha)$, as defined before, is given by

$$V(n,\alpha) = \sum p_{w_1}^{\alpha} q_{w_1}^{1-\alpha} p_{w_1 w_2}^{\alpha} q_{w_1 w_2}^{1-\alpha} \cdots p_{w_{n-k} w_{n-k+1}}^{\alpha} q_{w_{n-k} w_{n-k+1}}^{1-\alpha},$$

where the sum is over $w_1, w_2, \ldots, w_{n-k+1} \in \mathcal{X}^k$. For simplicity of notation, let $(p_1, \ldots, p_{M^k})$ and $(q_1, \ldots, q_{M^k})$ 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, $i,j = 1, \ldots, M^k$. Define a new matrix $R = (r_{ij})$ by

$$r_{ij} = p_{ij}^{\alpha} q_{ij}^{1-\alpha}, \quad i,j = 1, \ldots, M^k. \tag{6}$$

Also, define two new $1 \times M^k$ vectors $\mathbf{s} = (s_1, \ldots, s_{M^k})$ 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

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

where $\mathbf{1}^t$ denotes the transpose of the vector $\mathbf{1}$. It follows directly that with the new matrix $R$ as defined in (6), all the previous results also hold for a Markov source of arbitrary order.

# 4 Interchangeability of limits

## 4.1 Limit as $\alpha \to 1$

We herein show that although the Rényi divergence reduces to the Kullback-Leibler divergence as $\alpha \to 1$, the Rényi divergence rate does not necessarily reduce to the Kullback-Leibler divergence rate. Without loss of generality, we will herein 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. Let us first note the following result about the computation of the Kullback-Leibler divergence rate between two time-invariant Markov sources which follows based on [10, p. 68] and [12, Theorem 8.6.1].

**Proposition 7** Let $\{X_1, X_2, \ldots\}$ be a time-invariant Markov source with finite alphabet $\mathcal{X}$. Let $p^{(n)}$ and $q^{(n)}$ be two $n$-dimensional probability distributions on $\mathcal{X}^n$. Let $P$ and $Q$ be the probability transition matrices associated with $p^{(n)}$ and $q^{(n)}$ respectively. Let $q$ be the initial distribution with respect to $q^{(n)}$. If $Q > 0$, $q > 0$, and $P$ is irreducible, then the Kullback-Leibler divergence rate between $p^{(n)}$ and $q^{(n)}$ is given by

$$\lim_{n\to\infty} \frac{1}{n} D(p^{(n)} \| q^{(n)}) = -H_p(\mathcal{X}) - \sum_{i,j} \pi_i p_{ij} \log q_{ij},$$

where $H_p(\mathcal{X}) \triangleq \lim_{n\to\infty} \frac{1}{n} H(p^{(n)}) = -\sum_{i,j} \pi_i p_{ij} \log p_{ij}$ denotes the Shannon entropy rate of the source with respect to $p^{(n)}$, and $\pi$ denotes the stationary distribution associated with $p^{(n)}$.

We first show the following lemma; a direct consequence of this lemma generalizes a result of [21, p. 21] for ergodic Markov sources to irreducible Markov sources.

**Lemma 2** Let $A = (a_{ij})$ be an $n \times n$ matrix of rank $n-1$ with the property that $\sum_j a_{ij} = 0$ for each $i$. Define $c_i$ to be the cofactor of $a_{ii}$; i.e., the determinant of the matrix obtained from $A$ by deleting the $i$-th row and and the $i$-th column and let $c = (c_1, c_2, \ldots, c_n)$. Then $c$ is a non-zero vector and satisfies $cA = 0$.

**Proof:** See Appendix.

We next prove the following theorem.

**Theorem 2** Given that $\alpha \in (0,1) \cup (1,\infty)$, consider a time-invariant Markov source $\{X_1, X_2, \ldots\}$ with finite alphabet $\mathcal{X}$ and two possible distributions $p^{(n)}$ and $q^{(n)}$ on $\mathcal{X}^n$. 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 $q$ with respect to $q^{(n)}$ is positive then

$$\begin{aligned}\lim_{\alpha\to 1} \lim_{n\to\infty} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) &= \lim_{n\to\infty} \lim_{\alpha\to 1} \frac{1}{n} D_\alpha(p^{(n)} \| q^{(n)}) \\ &= -H_p(\mathcal{X}) - \sum_{i,j} \pi_i p_{ij} \log q_{ij} = \sum_{i,j} \pi_i p_{ij} \log(p_{ij}/q_{ij}),\end{aligned}$$

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

**Proof:** Since $P$ is irreducible and $Q$ is positive, then the matrix $R$ (as defined in Section 3) is irreducible. For convenience of notation denote the largest positive real eigenvalue of $R$ by $\lambda(\alpha, R)$. We know by Proposition 5 that each eigenvalue of $R$ is a continuous function of elements of $R$. Note that since $Q > 0$, $R \to P$ as $\alpha \to 1$, and the largest eigenvalue of the stochastic matrix $P$ is 1. Hence

$$\lim_{\alpha \to 1} \lambda(\alpha, R) = 1.$$

Let $a$ denote an arbitrary base of the logarithm. Then, by l'Hôpital's rule, we find that

$$\lim_{\alpha \to 1} \frac{\log \lambda(\alpha, R)}{\alpha - 1} = \frac{1}{\ln a} \lambda'(1, R) \triangleq \frac{1}{\ln a} \left. \frac{\partial \lambda(\alpha, R)}{\partial \alpha} \right|_{\alpha=1} \tag{7}$$

which is well defined by Proposition 6 since the algebraic multiplicity of $\lambda(\alpha, R)$ is 1 ($R$ is irreducible) by Proposition 3. The equation defining the largest positive eigenvalue $\lambda(\alpha, R) = \lambda$ of $R$ is

$$\begin{vmatrix} p_{11}^{\alpha} q_{11}^{1-\alpha} - \lambda & p_{12}^{\alpha} q_{12}^{1-\alpha} & \cdots & p_{1M}^{\alpha} q_{1M}^{1-\alpha} \\ p_{21}^{\alpha} q_{21}^{1-\alpha} & p_{22}^{\alpha} q_{22}^{1-\alpha} - \lambda & \cdots & p_{2M}^{\alpha} q_{2M}^{1-\alpha} \\ \vdots & \vdots & \ddots & \vdots \\ p_{M1}^{\alpha} q_{M1}^{1-\alpha} & p_{M2}^{\alpha} q_{M2}^{1-\alpha} & \cdots & p_{MM}^{\alpha} q_{MM}^{1-\alpha} - \lambda \end{vmatrix} = 0, \tag{8}$$

where $M = |\mathcal{X}|$. By differentiating this equation with respect to $\alpha$, we get [15],[19]

$$D_1 + D_2 + \cdots + D_M = 0, \tag{9}$$

where $D_i$ is the determinant obtained from (8) by replacing the $i$-th row by

$$(p_{i1}^{\alpha} q_{i1}^{1-\alpha} \ln(p_{i1}/q_{i1}), \ldots, p_{ii}^{\alpha} q_{ii}^{1-\alpha} \ln(p_{ii}/q_{ii}) - \lambda'(\alpha), \ldots, p_{iM}^{\alpha} q_{iM}^{1-\alpha} \ln(p_{iM}/q_{iM})).$$

and leaving the other $M-1$ rows unchanged. In this equation, $\lambda'$ denotes the derivative of $\lambda$ with respect to $\alpha$. Note that if we add in $D_i$ all the other columns to the $i$-th column, the value of the determinant remains unchanged. Therefore, for $\alpha = 1$ and hence $\lambda = 1$, $D_i$ is the determinant

$$\begin{vmatrix} p_{11} - 1 & \ldots & 0 & \ldots & p_{1M} \\ p_{21} & \ldots & 0 & \ldots & p_{2M} \\ \vdots & \vdots & 0 & \ldots & \vdots \\ p_{i-1,1} & \ldots & 0 & \ldots & p_{i-1,M} \\ p_{i1} \ln(p_{i1}/q_{i1}) & \ldots & S(X|i) - \lambda' & \ldots & p_{iM} \ln(p_{iM}/q_{iM}) \\ p_{i+1,1} & \ldots & 0 & \ldots & p_{i+1,M} \\ \vdots & \vdots & 0 & \ldots & \vdots \\ p_{M1} & \ldots & 0 & \ldots & p_{MM} - 1 \end{vmatrix},$$

where

$$S(X|i) = \sum_{j=1}^{M} p_{ij} \ln(p_{ij}/q_{ij}).$$

A zero occurs in all the entries of the $i$-th column except for the $i$-th entry, since $\sum_{j=1}^{M} p_{lj} = 1$. We conclude that

$$D_i = (S(X|i) - \lambda'(1))c_i, \tag{10}$$

where $c_i$ is the $M-1 \times M-1$ cofactor of $p_{ii}-1$ in the determinant of (8) for the case $\alpha = 1$, given by

$$c_i = \left| \begin{array}{ccccc} p_{11}-1 & \dots & p_{1,i-1} & \dots & p_{1M} \\ p_{21} & \dots & p_{2,i-1} & \dots & p_{2M} \\ \vdots & \dots & \dots & \dots & \vdots \\ p_{i-1,1} & \dots & p_{i-1,i-1}-1 & \dots & p_{i-1,M} \\ p_{i+1,1} & \dots & p_{i+1,i-1} & \dots & p_{i+1,M} \\ \vdots & \dots & \dots & \dots & \vdots \\ p_{M1} & \dots & p_{M,i-1} & \dots & p_{MM}-1 \end{array} \right|.$$

After substituting (10) in (9) and solving for $\lambda'(1)$, we obtain by (7) that

$$\lim_{\alpha\to 1} \frac{\log \lambda(\alpha, R)}{\alpha - 1} = \frac{1}{\ln a}\lambda'(1,R) = \frac{1}{\ln a}\sum_{i=1}^{M} \pi_i S(X|i), \tag{11}$$

where

$$\pi_i = \frac{c_i}{\sum_j c_j}.$$

As $\alpha \to 1$, $R \to P$; let $A = P - I$. Since the stationary distribution of the irreducible matrix $R$ is unique, the rank of $A$ is $n-1$ because the nullity of $A$ is 1 in this case. Hence, the conditions in Lemma 2 are satisfied. Therefore, $cA = 0$, which is equivalent to $cP = c$. Note that $c$ is the non-normalized stationary distribution of $P$ and (11) is just the Kullback-Leibler divergence rate between $P$ and $Q$ by Proposition 7. $\square$

The following example illustrates that the Rényi divergence rate does not necessarily reduce to the Kullback-Leibler divergence rate if the conditions of the previous theorem are not satisfied.

**Example**: Given that $\alpha \in (0,1) \cup (1,\infty)$, let $P$ and $Q$ be the following:

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

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 $q$ is positive. Then following [11, p. 40], a simple computation yields that the Kullback-Leibler divergence rate is given by $\log_2 3 - 2b + (3b/4)\log_2 3$, where the logarithm is to the base 2.

The eigenvalues of $R$ are: $\lambda_1 = 1/(3^{1-\alpha})$, $\lambda_2 = 4^{-\alpha}/(3^{1-\alpha}) + 4^{-\alpha}/(3^{1-2\alpha})$, and $\lambda_3 = 4^{-\alpha}/(3^{1-\alpha}) - 4^{-\alpha}/(3^{1-2\alpha})$. Note that $\mathbf{s} > 0$ and that, if $0 < \alpha < 1$, $\max_{1\le i\le 3}\{\lambda_i\} = \lambda_2$. By Theorem 1, the Rényi divergence rate is $(\alpha-1)^{-1}\log_2 \lambda_2$. By l'Hôpital's rule, we get that $\lim_{\alpha\uparrow 1}(\alpha-1)^{-1}\log_2\lambda_2 = (7/4)\log_2 3 - 2$. Therefore

$$\lim_{\alpha\uparrow 1}\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) = (7/4)\log_2 3 - 2.$$

On the other hand, if $\alpha > 1$, $\max_{1\le i\le 3}\{\lambda_i\} = \lambda_1$. Therefore, the Rényi divergence rate is given by $(\alpha-1)^{-1}\log_2\lambda_1$. Clearly, $\lim_{\alpha\downarrow 1}(\alpha-1)^{-1}\log_2\lambda_1 = \log_2 3$. Hence

$$\lim_{\alpha\downarrow 1}\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) = \log_2 3.$$

Therefore, the interchangeability of limits is not valid since

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

## 4.2 Limit as $\alpha \downarrow 0$

We obtain the following result.

**Theorem 3** Let $\alpha \in (0,1)$. Consider a time-invariant Markov source $\{X_1, X_2, \dots\}$ with finite alphabet $\mathcal{X}$ and two possible distributions $p^{(n)}$ and $q^{(n)}$ on $\mathcal{X}^n$. Let $P$ and $Q$ be the probability transition matrices on $\mathcal{X}$ associated with $p^{(n)}$ and $q^{(n)}$, respectively. Then

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

**Proof:** By Theorem 1, we have

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

By Proposition 5, $\lambda(\alpha, R) \to \lambda(0, R)$ as $\alpha \downarrow 0$. Hence

$$\lim_{\alpha\downarrow 0}\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)}) = -\log\lambda(0, R).$$

On the other hand
$$\lim_{\alpha\downarrow 0}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})=\frac{1}{n}\log \hat{\mathbf{s}}Y\mathbf{1}^t,$$
where $\hat{\mathbf{s}}=\lim_{\alpha\downarrow 0}\mathbf{s}$ and $Y=\lim_{\alpha\downarrow 0}R$. Therefore by again applying Theorem 1 to $Y$ we get
$$\lim_{n\to\infty}\lim_{\alpha\downarrow 0}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})=-\log\lambda(0,R).$$
Hence the interchangeability of limits is always valid between $n$ and $\alpha$ as $n\to\infty$ and as $\alpha\downarrow 0$. □

# 5 Rényi's entropy rate

The existence and the computation of the Rényi entropy rate of an arbitrary time-invariant finite alphabet Markov source can be deduced from the existence and the computation of the Rényi divergence rate. Indeed, if $q^{(n)}$ is stationary memoryless with uniform marginal distribution then for any $\alpha>0$, $\alpha\neq 1$,
$$D_\alpha(p^{(n)}\|q^{(n)})=n\log M-H_\alpha(p^{(n)}).$$
Therefore
$$\lim_{n\to\infty}\frac{1}{n}D_\alpha(p^{(n)}\|q^{(n)})=\log M-\lim_{n\to\infty}\frac{1}{n}H_\alpha(p^{(n)}).$$
Hence, the existence and the computation of the Rényi entropy rate follows directly from Theorem 1. Actually, $\lim_{n\to\infty}\frac{1}{n}H_\alpha(p^{(n)})$ can be computed directly from Theorem 1 by determining $\lambda$ with $R=(p_{ij}^\alpha)$ and $s_i=p_i^\alpha$, and setting $\lim_{n\to\infty}\frac{1}{n}H_\alpha(p^{(n)})=\frac{1}{1-\alpha}\log\lambda$. A formula for the Rényi entropy rate was established earlier in [18] and [19], but only for the particular case of ergodic finite alphabet time-invariant Markov sources.

Although the Rényi entropy reduces to the Shannon entropy, the Rényi entropy rate does not necessarily reduce to the Shannon entropy rate as $\alpha\to 1$. From the results about the interchangeability of limits for the Rényi divergence rate, it follows easily that the Rényi entropy rate always reduces to the Hartley entropy rate as $\alpha\downarrow 0$ ($\lim_{n\to\infty}\frac{1}{n}H_0(p^{(n)})$), and if the Markov source is irreducible, it reduces to the Shannon entropy rate as $\alpha\to 1$.

In [19], we established an operational characterization for the Rényi entropy rate by extending the variable-length source coding theorem in [5] for discrete memoryless sources to ergodic Markov sources. Using the above expression for the Rényi entropy rate, this source coding theorem can be easily extended to arbitrary time-invariant Markov sources. We also note that, by the results on the interchangeability of limits, the coding theorem does not necessarily reduce to the Shannon lossless source coding theorem as $\alpha \to 1$. It does reduce to the Shannon coding theorem if for example the Markov source is irreducible.

# 6 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 $\alpha \downarrow 0$. Similarly, we examined the computation and the existence of the Rényi entropy rate for Markov sources and investigated its limits as $\alpha \to 1$ and as $\alpha \downarrow 0$. We also observed that an operational characterization for the Rényi entropy rate can be established by extending a variable-length source coding theorem for memoryless sources to the case of Markov sources.

## Acknowledgment

The authors would like to thank the anonymous referees for their useful suggestions.

# Appendix

**Proof of Lemma 2:**

*Step 1:* First we prove that $c \neq 0$. The first $n-1$ columns of $A$ are linearly independent, because otherwise, the rank of $A$ is less or equal to $n-2$ since the sum of the columns of $A$ is 0. Thus there is at least one non-zero determinant $\Delta$ of size $(n-1) \times (n-1)$ which can be formed by deleting one row and the $n$-th column of $A$ which follows from the fact that the determinant of a matrix is 0 iff the columns are linearly dependent. Let the deleted row be the $k$-th row. If $k = n$, $\Delta = c_n$ and so $c \neq 0$. If $k < n$, add all the columns except the $n$-th column to the $k$-th column; this does not change the value of the determinant $\Delta$. Because $\sum_j a_{ij} = 0$, the elements of the $k$-th column are now $-a_{1n}, -a_{2n}, \ldots, -a_{nn}$. Multiply the elements of this column by $-1$ and move this column to the rightmost position. This yields a new determinant with value $\pm\Delta$ because these operations affect only the sign of the determinant. However, the new determinant is just $c_k$, so that once again, $c \neq 0$. Thus at least one of the cofactors $c_i$ is non-zero. Without loss of generality assume that $c_n \neq 0$. Next we prove that $cA = 0$.

*Step 2:* Consider the $n-1$ equations

$$\sum_{i=1}^{n} a_{ij} x_i = 0 \qquad j \in \{1, 2, \ldots, n-1\}. \tag{12}$$

Note that $\sum_{i=1}^{n} a_{ij} x_i = 0$ is equivalent to $\sum_{i=1}^{n-1} a_{ij} x_i = -a_{nj} x_n$. Since $c_n \neq 0$, we can use Cramer's rule [15, p. 60] to solve these equations for $x_1, \ldots, x_{n-1}$ in terms of $x_n$ as follows:

$$x_k = -x_n \frac{D_k}{c_n}, \tag{13}$$

where

$$D_k = \left| \begin{array}{cccccccc} a_{11} & a_{21} & \cdots & a_{k-1,1} & a_{n1} & a_{k+1,1} & \cdots & a_{n-1,1} \\ a_{12} & a_{22} & \cdots & a_{k-1,2} & a_{n2} & a_{k+1,2} & \cdots & a_{n-1,2} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{1,n-1} & a_{2,n-1} & \cdots & a_{k-1,n-1} & a_{n,n-1} & a_{k+1,n-1} & \cdots & a_{n-1,n-1} \end{array} \right|,$$

where the elements from the $n$-th column have replaced the elements of the $k$-th column. If we add the other rows to the $k$-th row (note that the determinants are transposed here) and use the fact that $\sum_j a_{ij} = 0$ we get a new $k$-th row

$$-a_{1n}, -a_{2n}, \ldots, -a_{k-1,n}, -a_{nn}, -a_{k+1,n}, \ldots, -a_{n-1,n}.$$

After moving the $k$-th row and the $k$-th column to the last row and column position respectively, it follows that $D_k = -c_k$. From (13), if we put $x_n = c_n$, then $x_k = c_k$ for all $k \in \{1, 2, \ldots, n\}$. Because $\sum_j a_{ij} = 0$, any solution of (12) is a solution of the same equation for $j = n$. Thus $c = (c_1, \ldots, c_n)$ satisfies $cA = 0$. $\square$

# References


[1] J. Aczél and Z. Daróczy, *On Measures of Information and Their Characterization*, Academic Press, New York, 1975.

[2] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Transactions on Information Theory,* vol. 42, No. 1, pp. 99–105, January 1996.

[3] M. B. Bassat and J. Raviv, "Rényi's entropy and the probability of error," *IEEE Transactions on Information Theory*, vol. IT-24, no. 3, pp. 324–330, May 1978.

[4] A. C. Blumer and R. J. McEliece, "The Rényi redundancy of generalized Huffman codes," *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 1242–1249, September 1988.

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

[6] P. N. Chen and F. Alajaji, "Csiszár's cutoff rates for arbitrary discrete sources," *IEEE Transactions on Information Theory*, to appear, January 2001.

[7] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Transactions on Information Theory*, vol. 41, pp. 26–34, January 1995.

[8] U. Erez and R. Zamir, "Error exponents of modulo-additive noise channels with side information at the transmitter,"*IEEE Transactions on Information Theory*, to appear.

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

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

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

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

[13] F. Jelinek, "Buffer overflow in variable length coding of fixed rate sources," *IEEE Transactions on Information Theory,* vol. IT-14, pp. 490–501, May 1968.

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

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

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

[17] T. Nemetz, *Information Type Measures and Their Applications to Finite Decision-Problems*, Carleton Mathematical Lecture Notes, no. 17, May 1977.

[18] L. Pronzato, H. P. Wynn and A. A. Zhigljavsky, "Using Rényi entropies to measure uncertainty in search problems," *Lectures in Applied Mathematics*, vol. 33, pp. 253–268, 1997.

[19] 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.

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

[21] V. I. Romanovsky, *Discrete Markov Chains*, Wolters-Noordhoff Publishing, Groningen, 1970.

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

[23] I. J. Taneja and R. M. Capocelli, "On some inequalities and generalized entropies: A unified approach," *Cybernetics and Systems*, vol. 16, 341–376, 1985.