THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 18, Number 4/2017, pp. 293–301

ENTROPY AND DIVERGENCE RATES FOR MARKOV CHAINS: I. THE ALPHA-GAMMA AND BETA-GAMMA CASE

Vlad Stefan BARBU[1], Alex KARAGRIGORIOU[2], Vasile PREDA[3]

[1] LMRS, Université de Rouen, FRANCE; barbu@univ-rouen.fr
[2] University of the Aegean, Department of Mathematics, Greece; alex.karagrigoriou@aegean.gr
[3] University of Bucharest and ISMMA of the Romanian Academy, Romania; vasilepreda0@gmail.com
Corresponding author: Vlad Stefan BARBU, barbu@univ-rouen.fr

Abstract. Divergence measures are of great importance in statistical inference. Equally important are their limiting versions, known as divergence rates. In this work we focus on generalized divergence measures for Markov chains. We consider generalizations of Alpha divergence measure (Amari and Nagaoka [3]) and Beta divergence measures (Basu *et. al* [6]) for Markov chains and investigate their limiting behavior. This work is continued in [4], where we study the corresponding weighted generalized divergence measures and the associated rates and in [5], where we present generalized Cressie and Read power divergence class of measures and numerically investigate some properties of all these generalized divergence measures and rates.

Key words: divergence measures, information measures, Markov chains, entropy, divergence rates.

1. INTRODUCTION

Shannon's 1948 paper provided the foundation for the development of information theory. Shannon introduced, through an axiomatic derivation, the notion of entropy as a measure of information for a probability distribution. The notion was quickly not only generalized to other entropy measures but also extended to measure the mutual information concerning two distributions and a plethora of the so-called divergence measures was introduced. Such measures of divergence are used as indices of similarity or dissimilarity between populations. In other words, they are used to measure the distance or the discrepancy between two distributions. It should be also noted that such measures are used for the construction of model selection criteria (Akaike [1], Cavanaugh [7], Dutta *et al*. [10]).

The original entropy measure introduced by Shannon [24] is given by

$$I^S(X) \equiv I^S(p) = -\int p \log p \mathrm{d}\mu,$$

where X is a random variable with density function $p(x)$ and μ a probability measure on **R.** Shannon derived the discrete version of $I^S(p)$, where p is a probability mass function, and named it *entropy* because of its similarity with thermodynamics entropy. The continuous version was defined by analogy. For a finite number of points, Shannon's entropy measures the expected information of a signal transferred without noise from a source X with density $p(x)$ and is related to Kullback-Leibler (KL) divergence through the expression

$$I^S(p) = I^S(h) - I^{KL}(p,h),$$

where h is the density of the uniform distribution and the Kullback-Leibler divergence between two densities $p(x)$ and $q(x)$ is given by

$$I^{KL}(p,q) = -\int p \log(p/q) \mathrm{d}\mu. \tag{1}$$

Many generalizations of Shannon's entropy were hereupon introduced. The Rényi's entropy [23] is given by

$$I^{R,\alpha}(p)=\frac{1}{\alpha-1}\log\left(\int p^{\alpha}\mathrm{d}\mu\right),\ \alpha>0,\ \alpha\neq 1 \tag{2}$$

and Liese and Vajda's [20] extension of Rényi's entropy is given by

$$I^{R_{lv},\alpha}(p)=\frac{1}{\alpha(\alpha-1)}\log\left(\int p^{\alpha}\mathrm{d}\mu\right),\ \alpha\neq 0,1.$$

Note that for $\alpha\to 1$ and $\alpha\to 0$ we get

$$\lim_{\alpha\to 1} I^{R_{lv},\alpha}(p)=I^{S}(p)\text{ and }\lim_{\alpha\to 0} I^{R_{lv},\alpha}(p)=\int \log p(x)\mathrm{d}x,$$

where the last one is the Burg's entropy (see, e.g., Kapur [18]). As mentioned earlier, a measure of divergence is used as a way to evaluate the distance (divergence) between any two populations or functions p and q. Among the most popular measures of divergence are the Kullback-Leibler measure of divergence (Kullback and Leibler [19]) given in (1) and the Csiszár's φ-divergence family of measures (Csiszár [9], Ali and Silvey [2]) given by

$$I^{\varphi}(p,q)=\int_0^{\infty} q\varphi\left(\frac{p}{q}\right)\mathrm{d}\mu\,, \tag{3}$$

where $\varphi(x)$ is a continuous, differentiable and convex function for $x\geq 0$.

In the case of the KL measure, the Liese and Vajda's corresponding generalization is given by

$$I^{R,\alpha}(p,q)=\frac{1}{\alpha(\alpha-1)}\log\left(\int p^{\alpha}q^{1-\alpha}\mathrm{d}\mu\right),\ \alpha\neq 0,1. \tag{4}$$

Equally important to the above divergence measures are their limiting versions, known as *divergence rates*. Besides the numerous limiting properties of these special divergences, such rates can be used in statistical inference in exactly the same manner as the typical (non-limiting) divergence measures. For a comprehensive review of various properties of rates we refer to Gray [13]. Formally, the divergence rate of a general divergence measure, say I^{D}, between two distributions p and q is defined by

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

Results for the rates of the standard KL and the Rényi measures for Markov chains have been provided by Rached *et. al* [21, 22]. The authors also provide the connection to the Shannon entropy rate. Results on the entropy of a more general class of stochastic processes, the processes with complete connections, were obtained by Iosifescu [15] and Iosifescu and Theodorescu [16].

The present work is concerned with measures of divergence for Markov chains. We consider generalizations of the measures presented above and we investigate their limiting behavior.

2. ALPHA AND BETA DIVERGENCE RATES FOR MARKOV CHAINS

In this section we discuss two broad classes of divergence measures, namely the Alpha and Beta measures and provide the basic results on their rates. Let (A,Ω,μ) be a measurable space and μ_p and μ_q some finite measures (not necessarily probability measures) defined on this space, with densities p and q with respect to a certain measure μ. The Alpha measure between μ_p and μ_q or, equivalently, between p and q, (Chernoff [8], Amari and Nagaoka [3]) is given by

$$D_A(p,q)=\frac{1}{\alpha(\alpha-1)}\left(\int p^{\alpha}(x)q^{1-\alpha}(x)-\alpha p(x)+(\alpha-1)q(x)\right)\mathrm{d}\mu\,(x), \tag{5}$$

where, for $\alpha=0$ and 1, it is defined by continuity. The same prolongation by continuity will be used for all the divergence measures considered in the rest of the paper. Through the transformation

$$c_0\int p^{c_1}(x)q^{c_2}(x)\mathrm{d}\mu(x)\to\log\left(\int p^{c_1}(x)q^{c_2}(x)\mathrm{d}\mu(x)\right)^{c_0} \tag{6}$$

the Alpha measure takes the form

$$D_{AG}(p,q)=\frac{1}{\alpha(\alpha-1)}\log\left(\frac{\int p^{\alpha}(x)q^{1-\alpha}(x)\mathrm{d}\mu(x)}{\left(\int p(x)\mathrm{d}\mu(x)\right)^{\alpha}\left(\int q(x)\mathrm{d}\mu(x)\right)^{1-\alpha}}\right). \tag{7}$$

It is easy to see that, by setting $\tilde{p}(x)=\frac{p(x)}{\int p(x)\mathrm{d}\mu(\mathrm{x})}$ and $\tilde{q}(x)=\frac{q(x)}{\int q(x)\mathrm{d}\mu(\mathrm{x})}$, the measure (7) becomes the so called *normalized Rényi measure*

$$D_{AG}(p,q)=\frac{1}{\alpha(\alpha-1)}\log\left(\int \tilde{p}^{\alpha}(x)\tilde{q}^{1-\alpha}(x)\mathrm{d}\mu(x)\right). \tag{8}$$

The Beta divergence between p and q (multiplied by the constant $1/(1+\alpha)$) introduced by Basu et. al [6] is given by

$$D_B(p,q)=\frac{1}{\alpha+1}\left(\int q^{1+\alpha}(x)-\left(1+\frac{1}{\alpha}\right)p(x)q^{\alpha}(x)+\frac{1}{\alpha}p^{1+\alpha}(x)\right)\mathrm{d}\,\mu(x), \tag{9}$$

which under the same transformation takes the form

$$D_{BG}(p,q)=-\frac{1}{\alpha}\log\left(\int \tilde{p}(x)\tilde{q}^{\alpha}(x)\right)\mathrm{d}\,\mu(x), \tag{10}$$

where, in this case,

$$\tilde{p}(x)=\frac{p(x)}{\left(\int p^{1+\alpha}(x)\mathrm{d}\,\mu(x)\right)^{1/(1+\alpha)}} \text{ and } \tilde{q}(x)=\frac{q(x)}{\left(\int q^{1+\alpha}(x)\mathrm{d}\,\mu(x)\right)^{1/(1+\alpha)}}.$$

Note that in all the previous definitions of divergences, one can consider either the case where the finite measures μ_p and μ_q, defined on a measurable space (A, Ω, μ) are absolutely continuous with respect to a certain measure μ (Lebesgue measure, for instance) and have associated densities p and q, or consider the discrete case, and p and q be the associated mass functions. As in the discrete case we can write the divergences in an integral form, taking μ as the counting measure on A, we expressed up to here the divergences in integral forms (i.e., not in terms of a sum). Nonetheless note also that, when we will be interested in divergences for Markov chains, we will use the notations with sums, not with integrals, because we consider only finite Markov chains.

Let us now focus on divergence measures for Markov chains. Let $(X_n)_{n\in N}$ be an ergodic time-homogeneous Markov chain with finite state space $\chi=\{1,...,M\}$. For this Markov chain, we consider two

different probability laws. Under the first law, let $p_i = P(X_1 = i), i \in \chi,$ denote the initial distribution of the chain and $p_{ij} = P(X_{k+1} = j \mid X_k = i), i, j \in \chi,$ the associated transition probabilities. Let also $\mathbf{p}_n$ denote the joint probability distribution of $(X_1, X_2, ..., X_n),$ i.e.,

$$\mathbf{p}_n(i_{1:n}) = P(X_1 = i_1, ...X_n = i_n) = p_{i_1} p_{i_1 i_2} ... p_{i_{n-1} i_n}, i_1, ..., i_n \in \chi,$$

were we denoted by $i_{1:n}$ the n-tuple $(i_1, ..., i_n) \in \chi^n$. Similarly we define under the second law q_i, q_{ij}, $\mathbf{q}_n(i_{1:n})$ and $\mathbf{q}_n$. Under this setting of finite state space Markov chains, the Alpha-Gamma measure between the two models is defined as the Alpha-Gamma measure between the two joint probability distributions $\mathbf{p}_n$ and $\mathbf{q}_n$, that is

$$D_{AG}(\mathbf{p}_n, \mathbf{q}_n) = \frac{1}{\alpha(\alpha-1)} \log\left(\sum_{i_{1:n} \in \chi^n} \frac{\mathbf{p}_n^{\alpha}(i_{1:n})}{\left(\sum_{i_{1:n} \in \chi^n} \mathbf{p}_n(i_{1:n})\right)^{\alpha}} \times \frac{\mathbf{q}_n^{1-\alpha}(i_{1:n})}{\left(\sum_{i_{1:n} \in \chi^n} \mathbf{q}_n(i_{1:n})\right)^{1-\alpha}} \right). \tag{11}$$

This can be written under the normalized form

$$D_{AG}(\mathbf{p}_n, \mathbf{q}_n) = \frac{1}{\alpha(\alpha-1)} \log\left(\sum_{i_{1:n} \in \chi^n} \tilde{\mathbf{p}}_n^{\alpha}(i_{1:n}) \tilde{\mathbf{q}}_n^{1-\alpha}(i_{1:n}) \right), \tag{12}$$

where

$$\tilde{\mathbf{p}}_n = \tilde{p}_{i_1} \tilde{p}_{i_1 i_2} ... \tilde{p}_{i_{n-1} i_n}, \tilde{\mathbf{q}}_n = \tilde{q}_{i_1} \tilde{q}_{i_1 i_2} ... \tilde{q}_{i_{n-1} i_n},$$

with $\tilde{p}_i$, $\tilde{p}_{ij}$, $\tilde{q}_i$ and $\tilde{q}_{ij}$, $i, j \in \chi,$ defined by

$$\tilde{p}_i = \frac{p_i}{\left(\sum_{i_{1:n} \in \chi^n} \mathbf{p}_n(i_{1:n})\right)^{1/n}}, \tilde{p}_{ij} = \frac{p_{ij}}{\left(\sum_{i_{1:n} \in \chi^n} \mathbf{p}_n(i_{1:n})\right)^{1/n}}, \tag{13}$$

$$\tilde{q}_i = \frac{q_i}{\left(\sum_{i_{1:n} \in \chi^n} \mathbf{q}_n(i_{1:n})\right)^{1/n}}, \tilde{q}_{ij} = \frac{q_{ij}}{\left(\sum_{i_{1:n} \in \chi^n} \mathbf{q}_n(i_{1:n})\right)^{1/n}}. \tag{14}$$

For the Beta-Gamma measure $D_{BG}(p_n, q_n)$ the corresponding form is

$$D_{BG}(\mathbf{p}_n, \mathbf{q}_n) =$$
$$= -\frac{1}{\alpha} \log\left(\sum_{i_{1:n} \in \chi^n} \frac{\mathbf{p}_n(i_{1:n})}{\left(\sum_{i_{1:n} \in \chi^n} \mathbf{p}_n^{1+\alpha}(i_{1:n})\right)^{1/(1+\alpha)}} \cdot \left(\frac{\mathbf{q}_n(i_{1:n})}{\left(sum_{i_{1:n} \in \chi^n} \mathbf{q}_n^{1+\alpha}(i_{1:n})\right)^{1/(1+\alpha)}} \right)^{\alpha} \right).$$

As previously, this can be written under the normalized form

$$D_{BG}(\mathbf{p}_n, \mathbf{q}_n) = -\frac{1}{\alpha} \log\left(\sum_{i_{1:n} \in \chi^n} \tilde{\mathbf{p}}_n(i_{1:n}) \tilde{\mathbf{q}}_n^{\alpha}(i_{1:n}) \right), \tag{15}$$

where

$$\tilde{\mathbf{p}}_n = \tilde{p}_{i_1} \tilde{p}_{i_1 i_2} ... \tilde{p}_{i_{n-1} i_n}, \tilde{\mathbf{q}}_n = \tilde{q}_{i_1} \tilde{q}_{i_1 i_2} ... \tilde{q}_{i_{n-1} i_n},$$

with $\tilde{p}_i$, $\tilde{p}_{ij}$, $\tilde{q}_i$ and $\tilde{q}_{ij}$, $i, j \in \chi$, defined by

$$\tilde{p}_i = \frac{p_i}{\left(\sum_{i_{1:n}\in\chi^n} \mathbf{p}_n^{1+\alpha}(i_{1:n})\right)^{1/n(1+\alpha)}}, \quad \tilde{p}_{ij} = \frac{p_{ij}}{\left(\sum_{i_{1:n}\in\chi^n} \mathbf{p}_n^{1+\alpha}(i_{1:n})\right)^{1/n(1+\alpha)}}, \tag{16}$$

$$\tilde{q}_i = \frac{q_i}{\left(\sum_{i_{1:n}\in\chi^n} \mathbf{q}_n^{1+\alpha}(i_{1:n})\right)^{1/n(1+\alpha)}}, \quad \tilde{q}_{ij} = \frac{q_{ij}}{\left(\sum_{i_{1:n}\in\chi^n} \mathbf{q}_n^{1+\alpha}(i_{1:n})\right)^{1/n(1+\alpha)}}. \tag{17}$$

The following theorems provide the divergence rates of Alpha-Gamma and Beta-Gamma measures. Since the $\tilde{R}$ matrices used in the proofs of Theorems 1 and 2, although they differ, they have the same structure, we will treat the two matrices in a similar way and we will proceed with a detailed proof of Theorem 1, only.

THEOREM 1. *Under the setting of the present section, we have*

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

where $\lambda(\alpha) := \lim_{n\to\infty} \lambda_n(\alpha)$ *(assumed to exist), where* $\lambda_n(\alpha)$ *is the largest positive eigenvalue of* $\tilde{R}(n) = (\tilde{r}_{ij}(\alpha))_{i,j\in\chi}$, *where*

$$\tilde{r}_{ij}(\alpha) = \tilde{p}_{ij}\tilde{q}_{ij}^{\alpha} = \frac{p_{ij}}{\left(\sum_{i_{1:n}\in\chi^n} \mathbf{p}_n^{1+\alpha}(i_{1:n})\right)^{1/n(1+\alpha)}} \left(\frac{q_{ij}}{\left(\sum_{i_{1:n}\in\chi^n} fq_n^{1+\alpha}(i_{1:n})\right)^{1/n(1+\alpha)}} \right)^{\alpha},$$

with $\tilde{p}_{ij}$ *and* $\tilde{q}_{ij}$ *defined in Equations (16) and (17), respectively.*

Proof. Define

$$V_{BG}(n,\alpha) = \sum_{i_{1:n}\in\chi^n} \tilde{\mathbf{p}}_n(i_{1:n})\tilde{\mathbf{q}}_n^{\alpha}(i_{1:n}) = \sum_{i_{1:n}\in\chi^n} \tilde{p}_{i_1}\tilde{q}_{i_1}^{\alpha}\tilde{p}_{i_1 i_2}\tilde{q}_{i_1 i_2}^{\alpha}\cdots\tilde{p}_{i_{n-1}i_n}\tilde{q}_{i_{n-1}i_n}^{\alpha}.$$

Define also the column vector $s = (s_1, s_2, \ldots, s_M)'$ by

$$s_i(\alpha) = \tilde{p}_i\tilde{q}_i^{\alpha}, i = 1, \ldots, M.$$

Let $\lambda_n(\alpha)$ be the largest positive real eigenvalue of the matrix $\tilde{R}(n)$ with associated positive right eigenvector $b = (b_1, b_2, \ldots, b_M)'$, so that $\tilde{R}^{n-1}(n)b = \lambda_n^{n-1}(\alpha)b$. Notice that

$$b_L \sum_{j\in\chi} \tilde{r}_{ij}^{(n-1)}(\alpha) \le \lambda_n^{n-1}(\alpha)b_i \le b_U \sum_{j\in\chi} \tilde{r}_{ij}^{(n-1)}(\alpha),$$

where b_L and b_U are the minimum and the maximum element of b and $\tilde{r}_{ij}^{(n-1)}(\alpha)$ is the (i, j) element of $\tilde{R}^{n-1}(n)$. It is now easy to see that

$$\frac{\sum_{i\in\chi} s_i(\alpha)b_i}{b_U} \le \frac{\sum_{i\in\chi}\left(s_i(\alpha)\sum_{j\in\chi}\tilde{r}_{ij}^{(n-1)}(\alpha)\right)}{\lambda_n^{n-1}(\alpha)} \le \frac{\sum_{i\in\chi} s_i(ha)b_i}{b_L}.$$

Notice now that

$$V_{BG}(n,\alpha)=\exp(-\alpha D_{BG}(\mathbf{p}_n,\mathbf{q}_n))=\sum_{i\in\chi}\left(s_i(\alpha)\sum_{j\in\chi}\tilde{r}_{ij}^{(n-1)}(\alpha)\right),$$

so that the above inequality easily becomes

$$-\frac{1}{\alpha n}\log\left(\frac{\sum_{i\in\chi}s_i(\alpha)b_i}{b_L}\right)-\frac{1}{\alpha n}\log(\lambda_n^{n-1}(\alpha))\le\frac{1}{n}D_{BG}(\mathbf{p}_n,\mathbf{q}_n)$$
$$\le-\frac{1}{\alpha n}\log\left(\frac{\sum_{i\in\chi}s_i(\alpha)b_i}{b_U}\right)-\frac{1}{\alpha n}\log(\lambda_n^{n-1}(\alpha)).$$

Taking now the limit as n tends to infinity, the result is immediate, provided that the first term in the lower bound as well as in the upper bound tends to 0. In fact, since b_U and b_L are fixed, it is sufficient to show that $\log(\sum_{i\in\chi}s_i(\alpha)b_i)$ is bounded. By the definition of s_i, $\tilde{p}_i$, and $\tilde{q}_i$ we have

$$\log(\sum_{i\in\chi}s_i(\alpha)b_i)=\log(\sum_{i\in\chi}p_iq_i^{\alpha}b_i)-\frac{1}{(1+\alpha)n}\log\left(\sum_{i_{1:n}\in\chi^n}\mathbf{p}_n^{1+\alpha}(i_{1:n})\right)$$
$$-\frac{\alpha}{(1+\alpha)n}\log\left(\sum_{i_{1:n}\in\chi^n}\mathbf{q}_n^{1+\alpha}(i_{1:n})\right).$$

Observe now that by Schwarz inequality $\sum_{i_{1:n}\in\chi^n}\mathbf{p}_n^{1+\alpha}(i_{1:n})\le n\left(\frac{1}{n}\right)^{1+\alpha}$, which implies $\sum_{i_{1:n}\in\chi^n}\mathbf{p}_n^{1+\alpha}(i_{1:n})=O(1/n^{\alpha})$. The same holds for $\sum_{i_{1:n}\in\chi^n}\mathbf{q}_n^{1+\alpha}(i_{1:n})$ and the result follows.

THEOREM 2. *Under the setting of the present section, we have*

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

where $\lambda(\alpha)$ *is the largest positive eigenvalue of* $\tilde{R}=(\tilde{r}_{ij}(\alpha))_{i,j\in\chi}$, *where*

$$\tilde{r}_{ij}(\alpha)=\tilde{p}_{ij}^{\alpha}\tilde{q}_{ij}^{1-\alpha}=\frac{p_{ij}^{\alpha}}{\left(\sum_{i_{1:n}\in\chi^n}\mathbf{p}_n(i_{1:n})\right)^{\alpha/n}}\cdot\frac{q_{ij}^{1-\alpha}}{\left(\sum_{i_{1:n}\in\diamond i^n}\mathbf{q}_n(i_{1:n})\right)^{(1-\alpha)/n}},$$

with $\tilde{p}_{ij}$ *and* $\tilde{q}_{ij}$ *defined in Equations (13) and (14), respectively.*

Proof. Notice that the form of the vector s_i is given by

$$s_i(\alpha)=p_i^a q_i^{1-\alpha}\frac{1}{\left(\sum_{i_{1:n}\in\chi^n}\mathbf{p}_n(i_{1:n})\right)^{\alpha}\left(\sum_{i_{1:n}\in\chi^n}\mathbf{q}_n(i_{1:n})\right)^{1-\alpha}}.$$

The proof follows exactly the same steps as that of Theorem 1. Note that, in this case,

$$\log(\sum_{i\in\chi}s_i(\alpha)b_i)=\log(\sum_{i\in\chi}p_i^{\alpha}q_i^{1-\alpha}b_i).$$

Remarks:

1. The Alpha-Gamma divergence given in (7) coincides with the Liese & Vajda's measure given in (4) for probability distributions, namely under the assumptions that $\int q(x)d\mu(x) = \int p(x)d\mu(x) = 1$. In that sense, Alpha-Gamma divergence is a natural generalization of the Rényi measure for general distributions.

2. The Alpha divergence is related to the Csiszár's family of measures (Csiszár [9], Ali and Silvey [2]). Indeed, the two families coincide if for Csiszár measure we take

$$\varphi(u) = \frac{1}{\alpha(1-\alpha)}\left[u^{\alpha} - 1 - \alpha(u-1)\right].$$

3. For $\alpha \to 1$ we obtain the Kullback-Leibler measure for the Alpha, the Alpha-Gamma and the Csiszár's measures. For $\alpha \to 0$, the Beta and Beta-Gamma measures tend to the Kullback-Leibler measure, while for $\alpha = 1$ the Beta measure becomes the L_2 distance. Observe that here we are referring to the generalized KL measure in the sense that general distributions (non-necessarily probability) are allowed. Sometimes, the term "Generalized KL" measure is used.

4. The Alpha and Beta measures are related to the Gamma-family of divergences given by

$$D_G(\mathbf{p}_n, \mathbf{q}_n) = \frac{1}{\alpha(\alpha-1)} \frac{\sum_{i_{1:n}\in\chi^n} \mathbf{p}_n^{\alpha}(i_{1:n})\left(\sum_{i_{1:n}\in\chi^n} \mathbf{q}_n^{\alpha}(i_{1:n})\right)^{\alpha-1}}{\left(\sum_{i_{1:n}\in\chi^n} \mathbf{p}_n(i_{1:n})\mathbf{q}_n^{\alpha-1}(i_{1:n})\right)^{\alpha}}.$$

The measure was introduced by Fujisawa and Eguchi [12] and allows *super* robust parameter estimation. This measure is useful especially if outliers are present. The authors have proved that the bias due to outliers becomes significantly small even if heavy contamination is present. The measure is referred to as γ-cross entropy and it is the same as the logarithm of the cross entropy proposed by Jones *et al*. [17] on the basis of Windham [28].

5. Note that the same type of results as those presented in Theorems 1 and 2 hold true also for Markov chains of arbitrary order $k, k > 1$. Indeed, if $(X_n)_{n\in N}$ is a Markov chain of order k with state space χ, then $(Y_n)_{n\in N}$, $Y_n = (X_n, X_{n+1}, ..., X_{n+k-1})$, is a Markov chain of order 1, with state space χ^k and for this Markov chain the mentioned results hold true.

6. Rényi and Liese & Vajda's entropy measures and their rates are closely related to the Kullback-Leibler measure and the associated rate given by Rached et. al [22]. Consider the Markov chain setting of the previous section. Then, the Kullback-Leibler divergence is given by

$$D_{KL}(\mathbf{p}_n, \mathbf{q}_n) = \sum_{i_{1:n}\in\chi^n} \mathbf{p}_n(i_{1:n}) \log\left(\frac{\mathbf{p}_n(i_{1:n})}{\mathbf{q}_n(i_{1:n})}\right).$$

The rate of $D_{KL}(\mathbf{p}_n, \mathbf{q}_n)$ can be found in Rached at. al (2004) and is given by

$$\lim_{n\to\infty} \frac{1}{n} D_{KL}(\mathbf{p}_n, \mathbf{q}_n) = \sum_{i\in\chi} \pi_i \sum_{j\in\chi} p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right),$$

where $\pi = (\pi_1, ..., \pi_M)'$ is the unique stationary distribution associated with the measure $\mathbf{p}$ (assumed to exist and to be unique). Recall that the discrete version of Shannon's entropy for a Markov chain is given by

$$I^S(\mathbf{p}_n) = -\sum_{i_{1:n}\in\chi^n} \mathbf{p}_n(i_{1:n}) \log \mathbf{p}_n(i_{1:n})$$

and it is related to the KL divergence given in (2) through the expression

$$I^S\left(\mathbf{p}_n\right) = n\log M - D_{KL}(\mathbf{p}_n, \mathbf{u}),$$

where M is the number of states and $\mathbf{u}$ is the uniform distribution over the M values/states. As a result, using (3), the rate of Shannon's entropy is given by

$$\lim_{n\to\infty}\frac{1}{n}I^S\left(\mathbf{p}_n\right) = -\lim_{n\to\infty}\frac{1}{n}D_{KL}(\mathbf{p}_n, \mathbf{u}) + \log M.$$

Observe that since $\mathbf{u}$ is the uniform distribution, the above rate simplifies to

$$\lim_{n\to\infty}\frac{1}{n}I^S\left(\mathbf{p}_n\right) = -\sum_{i\in\chi}\pi_i\sum_{j\in\chi}p_{ij}\log\left(p_{ij}\right).$$

On the other hand, for the case of Liese & Vajda's entropy, relation (4) takes the form

$$I^{R_{lv},\alpha}\left(\mathbf{p}_n\right) = \frac{n\log M}{\alpha} - D_{AG}(\mathbf{p}_n, \mathbf{u}),$$

so that, by Theorem 2, the resulting rate takes the form

$$\lim_{n\to\infty}\frac{1}{n}I^{R_{lv},\alpha}\left(\mathbf{p}_n\right) = \lim_{n\to\infty}D_{AG}(\mathbf{p}_n, \mathbf{u}) = \frac{1}{\alpha(1-\alpha)}\log\lambda(\alpha).$$

Note also that a similar result holds true for Rényi's entropy rate.

ACKNOWLEDGEMENTS

The authors would like to thank their colleague Jean-Baptiste Bardet from Laboratoire de Mathématiques Raphaël Salem, University of Rouen, France, for his suggestions and help on some technical problems related to this paper. The first two authors would also like to express their appreciation to the University of Rouen and University of the Aegean for the opportunity to exchange several visits in both institutions. The research work of Vlad Stefan Barbu was partially supported by the projects *XTerM–Complex Systems, Territorial Intelligence and Mobility* (2014–2018) and *MOUSTIC–Random Models and Statistical, Informatics and Combinatorics Tools* (2016–2019), within the Large Scale Research Networks from the Region of Normandy, France.

REFERENCES

1. H. AKAIKE, *Information theory and an extension of the maximum likelihood principle*, Proceeding of the Second International Symposium on Information Theory, B.N. Petrov and F. Csaki (eds.), Akademia Kaido, Budapest, 1973, pp. 267–281.
2. S.M. ALI, S.D. SILVEY, *A general class of coefficients of divergence of one distribution from another*, J. Roy. Statist. Soc. B, **28**, pp. 131–142, 1966.
3. S. AMARI, H. NAGAOKA, *Methods of Information Geometry*, Oxford University Press, New York, 2000.
4. V.S. BARBU, A. KARAGRIGORIOU, V. PREDA, *Entropy and divergence rates for Markov chains: II. The weighted case*, submitted, 2017.
5. V.S. BARBU, A. KARAGRIGORIOU, V. PREDA, *Entropy and divergence rates for Markov chains: III. The Cressie and Read case and applications*, submitted, 2017.
6. A. BASU, I.R. HARRIS, N.L. HJORT, M.C. JONES, *Robust and efficient estimation by minimising a density power divergence*, Biometrika, **85**, pp. 549–559, 1998.
7. J.E. CAVANAUGH, *Criteria for linear model selection based on Kullback's symmetric divergence*, Australian and New Zealand Journal of Statistics, **46**, pp. 257–274, 2004.
8. H. CHERNOFF, *A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations*, Ann. Math. Statist., **23**, *4*, pp. 493–507, 1952.
9. I. CSISZÁR, *Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten*, Publication of the Mathematical Institute of the Hungarian Academy of Sciences, **8**, pp. 84–108, 1963.
10. R. DUTTA, M. BOGDAN, J.K. GHOSH, *Model selection and multiple testing - A Bayes and empirical Bayes overview and some new results*, Journal of the Indian Statistical Association, **50**, pp. 105–142, 2012.

11. M.F. FREEMAN, J.W. TUKEY, *Transformations related to the angular and the square-root*, Ann. Math. Statist., **21**, pp. 607–611, 1950.
12. H. FUJISAWA, S. EGUCHI, *Robust parameter estimation with a small bias against heavy contamination*, Multivariate Analysis, **99**, pp. 2053–2081, 2008.
13. R.M. GRAY, *Entropy and Information Theory*, New York, Springer-Verlag, 1990.
14. C. HUBER-CAROL, N. BALAKRISHNAN, M.S. NIKULIN, M. MESBAH, *Goodness-of-fit Tests and Model Validity*, Birkhäuser, Boston, 2002.
15. M. IOSIFESCU, *Sampling entropy for random homogeneous systems with complete connections*, Ann. Math. Statist., **36**, pp. 1433–1436, 1965.
16. M. IOSIFESCU, R. THEODORESCU, *Asupra entropiei lanturilor cu legaturi complete*, Com. Acad. RPR, **11**, pp. 821–824, 1961.
17. M.C. JONES, N.L. HJORT, I.R. HARRIS, A. BASU, *A comparison of related density-based minimum divergence estimators*, Biometrika, **88**, pp. 865–873, 2001.
18. J.N. KAPUR, *Measures of Information and Their Applications*, Wiley, New Delhi, 1994.
19. S. KULLBACK, R. LEIBLER, *On information and sufficiency*, Ann. of Math. Statist., **22**, pp. 79–86, 1951.
20. F. LIESE, I. VAJDA, *Convex Statistical Distances*, Teubner, Leipzig, 1987.
21. Z. RACHED, F. ALAJAJAI, L.L. CAMPBELL, *Rényi's divergence and entropy rates for finite alphabet Markov sources*, IEEE Transc. Inf. Theory, **47**, *4*, pp. 1553–1561, 2001.
22. Z. RACHED, F. ALAJAJAI, L.L. CAMPBELL, *Kullback-Leibler divergence rate between Markov sources*, IEEE Transc. Inf. Theory, **50**, *5*, pp. 917–921, 2004.
23. A. RÉNYI, *On measures of entropy and information*, Proc. of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, **1**, pp. 547–561, 1961.
24. C.E. SHANNON, *A mathematical theory of communication*, Bell System Technical Journal, **27**, pp. 379–423, 1948.
25. B.D. SHARMA, J. MITTER, M. MOHAN, *On measures of "useful" information, Information and Control*, **39**, *3*, pp. 323–336, 1978.
26. A. TOMA, *Minimum Hellinger distance estimators for multivariate distributions from the Johnson system*, J. Statist. Plan. and Infer., **138**, pp. 803–816, 2008.
27. A. TOMA, *Optimal robust M-estimators using divergences*, Statistics and Prob. Letters, **79**, pp. 1–5, 2009.
28. M.P. WINDHAM, *Robustifying model fitting*, J. Roy. Statist. Soc. B, **43**, pp. 599–609, 1995.
29. J. ZHANG, *Powerful goodness-of-fit tests based on likelihood ratio*, J. R. Stat. Soc. Ser. B, **64**, *2*, pp. 281–294, 2002.
30. K. ZOGRAFOS, K. FERENTINOS, T. PAPAIOANNOU, Φ-*divergence statistics: Sampling properties, multinomial goodness of fit and divergence tests*, Comm. in Statist. Theor. Meth., **19**, *5*, pp. 1785–1802, 1990.

Received March 31, 2017