Gamma-Divergence. An introduction to new divergence family.

Leonardo E. Riveaud [1,2,*], Diego Mateos [1,3,4], and Pedro W. Lamberti[1,5]

[1]Consejo Nacional de Investigaciones Científicas y T'ecnicas (CONICET), Argentina.
[2]Facultad de ingeniería, Universidad Nacional del Comahue (FAIN, UNComa)
[3]Facultad de Ciencia y Tecnología. Universidad Autónoma de Entre Ríos (UADER). Oro Verde, Entre Ríos, Argentina.
[4]Instituto de Matemática Aplicada del Litoral (IMAL-CONICET-UNL), CCT CONICET, Santa Fé, Argentina.
[5]Facultad de Matemática Astronomía y Física (FaMAF), Universidad Nacional de Córdoba. Córdoba, Argentina.
[*]Corresponding author: Leonardo E. Riveaud, leoriveaud@gmail.com.

Abstract

Divergences have become a very useful tool for measuring similarity (or dissimilarity) between probability distributions. Depending on the field of application, a more appropriate measure may be necessary. In this paper we introduce a family of divergences called γ-Divergences. They are based on the convexity property of the functions that generate them. We demonstrate that these divergences verify all the usually required properties, and we extend them to weighted probability distribution. In addition, we define a generalised entropy closely related to the γ-Divergences. Finally, we apply our findings to the analysis of simulated and real time series.

Keywords— Entropy, Divergence, Information theory, Family divergence, convex function

1 Introduction

There exists a great number of divergences between probability distributions. Remarkably when they are applied to the same statistical problem, in general they do not lead to indistinguishable results. Therefore, it is useful to have a wide set of divergences. In general, the divergences have different origins. Some are purely statistical, others originated in information theory. The Fishers metric is a conspicuous example of the first kind; among the second class are the Kullback-Leibler and the Jensen-Shannon divergences [1]. These measures of similarity (or dissimilarity), between probability distributions have become of great interest in many areas of the science as Physics (classical and quantum), Biology and many other areas of science [2–7].

It is well known that not all distance or divergence are adequate to treat every problem. So, having a variety of divergence could be useful for both theoretical studies and for applications. Sometimes it has been possible to introduce families of divergences, labelling each member with a parameter [8, 9] or by giving a general structure depending on a function of a certain characteristic. An example of this late one is known as Csiszar divergences or f-Divergences. They are defined as:

$$\mathcal{D}_f(P||Q) = \sum_i f\left(\frac{p_i}{q_i}\right) q_i$$

where $f(x)$ is a convex function such that $f(1) = 0$, and p_i ad q_i are discrete probability distributions. This family has been extensively studied in the context of information geometry. A remarkable result is that when p_i and $q_i = p_i + \delta p_i$ are closed probability distributions, the divergence $\mathcal{D}_f(P||Q)$ is proportional to the Fisher (riemannian) metric. The above mentioned of Kullback-Leibler and Jensen-Shannon divergences correspond to $f(t) = t \log t$ and $f(t) = (t+1)\log(\frac{2}{t+1}) + t \log t$, respectively [10].

In this work we present a new family of divergences that we call γ-Divergences, based on the property of convex functions. By mimicking the structure of Euclidean metric, we propose a family of divergences that verify the basic conditions of a "good" divergence [11].

This work is organized as follows. In section 2 we introduce the family of distance based on convex functions. We demonstrate that these distances meet the requirements to be considered as divergences. Later on, we show the characteristics that this γ-Divergence, and also extended our divergence to weighted and N-dimensional distribution. In section 3 we introduce generalized entropy based on the notion of convexity. We show this entropy meets all the requirements that a generalized entropy should meet. In addition, we investigated the relationship between this γ-Entropy and the γ-Divergence defined above. In section 4 we applied the divergence for the detection of dynamics changes in generated sequences and electroencephalographic signals. Finally, in section 5 we discussed the results obtained and future works are proposed.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2020 doi:10.20944/preprints202011.0388.v1

2 The γ-Divergence family

Let $P = \{p_i\}_{i=1}^n$ and $Q = \{q_i\}_{i=1}^n$ two discrete probability distributions for a N-state of a random variable X. The square of the Euclidean metric between these two distribution can be written in the form:

$$\mathcal{E}(P||Q) = \sum_i (q_i - p_i)^2 = 2\ q_i^2 + 2\ p_i^2 - (p_i + q_i)^2 \tag{1}$$

or in an equivalent form:

$$\mathcal{E}(P||Q) = \sum_i 2\left(q_i\ g(q_i) + p_i\ g(p_i) - (p_i + q_i)\ g\left(\frac{p_i + q_i}{2}\right)\right) \tag{2}$$

where $g(x) = x$ is the identity function. The square root of the Euclidean distance is a true metric, in the sense that it verifies the triangle inequality.

Let

$$\mathcal{D}_{JS}(P||Q) = \sum_i \frac{1}{2}p_i\ \log(p_i) + \frac{1}{2}q_i\ \log(q_i) - \frac{1}{2}(p_i + q_i)\ \log\left(\frac{q_i + p_i}{2}\right) \tag{3}$$

be the Jensen Shannon divergence, and let $d_{JS} = \sqrt{\mathcal{D}_{JS}}$ its square root. It is known that d_{JS} is a true metric [1]. Its square can be rewritten in the form:

$$\mathcal{D}_{JS}(P||Q) = \sum_i q_i\ g(q_i) + p_i\ g(p_i) - (q_i + p_i)\ g\left(\frac{q_i + p_i}{2}\right) \tag{4}$$

where $g(x) = \frac{1}{2}\log(x)$. Both, the Euclidean and the JSD have the same structure. Furthermore, the function $x \cdot g(x)$ is convex, both in the case of the Euclidean distance as in the case of the JSD. This simple observation leads us to propose, for each function $g(x)$ such that $x \cdot g(x)$ is convex, a divergence

$$\mathcal{D}_\gamma(P||Q) = \sum_i \gamma_g(p_i, q_i) \tag{5}$$

where

$$\gamma_g(p_i, q_i) = p_i\ g(p_i) + q_i\ g(q_i) - (p_i + q_i)\ g\left(\frac{p_i + q_i}{2}\right) \tag{6}$$

Theorem: Let $P = \{p_i\}_{i=1}^n$ and $Q = \{q_i\}_{i=1}^n$ two probability distributions for a given N -state random variable X. Let $g : \mathcal{R}^+ \to \mathcal{R}$ such that $f(x) := x \cdot g(x)$ a convex function. Then the functional defined as:

1. $\mathcal{D}_\gamma(P||Q) = \mathcal{D}_\gamma(Q||P)$ $\quad$ (*Symmetry*)
2. $\mathcal{D}_\gamma(P||Q) > 0$ $\quad$ for $\quad Q \not\equiv P$ $\quad$ (*Positivity*)
3. $\mathcal{D}_\gamma(P||Q) = 0 \iff P \equiv Q$

Proof: The divergence D_γ is a sum of N terms of $\gamma_g(p_i, q_i)$ Therefore, if each of these terms are symmetric, positive and null if and only if $Q \equiv P$ these properties are inherited by D_γ.
From now on we will use the following notation,

$$m_i := \frac{p_i + q_i}{2} \tag{7}$$

i) Symmetry
It is direct to check that 6 we can see that $\gamma_g(p_i, q_i)$ satisfies

$$\gamma_g(p_i, q_i) = \gamma_g(q_i, p_i),\ \forall i \tag{8}$$

$\diamondsuit$
To prove $\mathcal{D}_\gamma(P||Q) \geq 0$ we need to demonstrate that

$$\gamma_g(q_i, p_i) \geq 0, \quad \forall i. \tag{9}$$

Under the hypotheses that f is a convex function, for all $t \in [0, 1]$ and $p, q \in [0, 1]$ the Jensen inequality leads to

$$t\ f(q) + (1 - t)\ f(p) \geq f(t\ q + (1 - t)\ p) \tag{10}$$

if we replace $f(x) = x\ g(x)$ we have

$$t\ (p_i\ g(p_i)) + (1 - t)\ (q_i\ g(q_i)) - (t\ p_i + (1 - t)\ p_i)\ \ g\,(t\ p_i + (1 - t)\ q_i) \geq 0 \tag{11}$$

choosing $t = 1/2$ we obtain

$$\frac{p_i}{2}\ g(p_i) + \frac{q}{2}\ g(q_i) - \frac{(p_i + q_i)}{2}\ g\left(\frac{p_i + q_i}{2}\right) \geq 0 \tag{12}$$

using definition 6 we have

$$\frac{\gamma_g(p_i, q_i)}{2} \geq 0 \Longrightarrow \gamma_g(p_i, q_i) \geq 0, \quad \forall i \tag{13}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2020 doi:10.20944/preprints202011.0388.v1

therefore, the sum of positive amounts is positive

$$D_\gamma(P||Q) = \sum_i \gamma_g(p_i, q_i) \geq 0 \tag{14}$$

♢

$D_\gamma(P||Q) = 0 \iff P \equiv Q$:

⇐)

Replacing $P \equiv Q$ in the definition of $\gamma_g(p_i, q_i)$ given in 6 we have

$$\gamma_g(q_i, q_i) = q_i\ g(q_i) + q_i\ g(q_i) - (q_i + q_i)\ g\left(\frac{q_i + q_i}{2}\right) = 0 \quad \forall i \tag{15}$$

then $\mathcal{D}_\gamma(P||Q) = 0$.

⇒)

Previously, it has been shown $\gamma_g(p_i, q_i)$ is positive $\forall i$, therefore,

$$\mathcal{D}_\gamma(P||Q) = 0 \iff \gamma_g(p_i, q_i) = 0, \qquad \forall i \tag{16}$$

with simple algebraic steps

$$m_i\ g(m_i) = \frac{p_i\ g(p_i) + q_i\ g(q_i)}{2}, \tag{17}$$

which allows to write $f(x)$,

$$f(m_i) = \frac{f(p_i) + f(q_i)}{2} \tag{18}$$

$f(m_i)$ is convex and different from identity, therefore equality is satisfied if and only if $p_i = m_i$ and $q_i = m_i$, implying $p_i = q_i$.

♢ •

2.1 Subset of functions $g(x)$

Due to the previous result, we can find a subset of functions $g(x)$ that satisfy the hypothesis of the theorem 2. We know that if $f(x)$ is two times differentiable, it is strictly convex if and only if

$$\frac{d^2 f(x)}{dx^2} = \frac{d^2(x \cdot g(x))}{dx^2} > 0 \tag{19}$$

Then, the inequality

$$x\frac{d^2 g(x)}{dx^2} + 2\frac{dg(x)}{dx} > 0 \qquad \forall x \geq 0 \tag{20}$$

gives a subset of functions $g(x)$ that satisfies the theorem 2, building a subfamily of γ-Divergence.

2.2 Linearity

Next, we will show that a family of γ-Divergences can be generated using the function $g(x)$ as linear sum of functions $g_k(x)$ which meet the conditions presented in 2.1.

Let $g(x) = \sum_{k=1}^m \alpha_k\ g_k(x)$, with $\alpha_k > 0$ for all $k : 1, ..., m$, where the functions $g_k(x)$ satisfy the hypotheses of the theorem 2. Replacing in equation 6 we obtain:

$$\gamma_g(p_i, q_i) = \sum_k \alpha_k\ \gamma_{g_k}(p_i, q_i) \tag{21}$$

where

$$\gamma_{g_k}(p_i, q_i) := p_i\ g_k(p_i) + q_i\ g_k(q_i) - (p_i + q_i)\ g_k\left(\frac{p_i + q_i}{2}\right), \tag{22}$$

then γ-Divergence takes the following form

$$\mathcal{D}_\gamma(P||Q) = \sum_k \alpha_k \mathcal{D}_{\gamma_k}(P||Q) \tag{23}$$

where

$$\mathcal{D}_{\gamma_k}(P||Q) := \sum_i \gamma_{g_k}(p_i, q_i) \tag{24}$$

Now we will show that γ_{g_k} comply the properties of the theorem 2.

Every term $g_k(x)$ satisfies the theorem 2 hypotheses for $\forall k$ implying that $\mathcal{D}_{\gamma_k} \geq 0$. In consequence if $\alpha_k > 0$ then

$$\mathcal{D}_\gamma(Q||P) \geq 0. \tag{25}$$

On the other hand, considering $\mathcal{D}_{\gamma_k}(P||Q) = \mathcal{D}_{\gamma_k}(P||Q)$ for all k, we have

$$\mathcal{D}_\gamma(P||Q) = \mathcal{D}_\gamma(Q||P). \tag{26}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2020 doi:10.20944/preprints202011.0388.v1

Finally, due to $\mathcal{D}_{\gamma_{g_k}}(Q||P) = 0 \iff$ if $P \equiv Q$ for all k by hypothesis, if $\alpha_k > 0\ \forall k$, we have

$$P \equiv Q \Longrightarrow \mathcal{D}_\gamma(Q||P) = \sum_k \alpha_k \mathcal{D}_{\gamma_k}(P||Q) = 0 \tag{27}$$

If $\mathcal{D}_{\gamma_g}(P||Q) = 0$ implies each terms of the sum must be equal to zero. Due to $\alpha_k > 0$ for all k, then $\mathcal{D}_{\gamma_k}(P||Q) = 0$. By hypothesis each of the $\mathcal{D}_{\gamma_k}$ satisfy the theorem 2, this means

$$\mathcal{D}_{\gamma_k}(P||Q) = 0 \Longrightarrow P \equiv Q \qquad \forall k \tag{28}$$

therefore $\mathcal{D}_\gamma(P||Q) = 0 \Longrightarrow Q \equiv P$.

2.3 Similarity

Let $q_i = p_i + \delta p_i$ for each i, with $\sum_{i=1}^N \delta p_i = 0$, and let $g(x)$ a differential function, then the γ-divergence can be written as:

$$\begin{aligned} \mathcal{D}_\gamma(P + \delta P||P) &= \sum_i \gamma_g(p_i + \delta p_i, p_i) \\ &= \sum_i (p_i + \delta p_i)\left(g\left(p_i + \delta p_i\right) - g\left(p_i + \frac{\delta p_i}{2}\right)\right) \\ &+ \ p_i \ \left(g(p_i) - g\left(p_i + \frac{\delta p_i}{2}\right)\right) \end{aligned} \tag{29}$$

taking the Taylor's expansion linear term of the function $g(x)$ valued in $y = p_i + \delta p_i$, we obtain

$$g(p_i + \delta p_i) \simeq g(p_i) + \dot{g}(p_i)\delta p_i \tag{30}$$

where $\dot{g}(x) := \frac{dg}{dx}$.
Then γ-Divergence approximation is:

$$D_\gamma(P + \delta P||P) \simeq (p_i + \delta p_i)\left(\dot{g}(p_i)\delta p_i + g(p_i) - \dot{g}(p_i)\frac{\delta p_i}{2} - g(p_i)\right) + p_i\left(g(p_i) - \dot{g}(p_i)\frac{\delta p_i}{2} - g(p_i)\right) \tag{31}$$

with some algebraic step we have

$$D_\gamma(P + \delta P||P) \simeq (p_i + \delta p_i)\ \dot{g}(p_i)\ \frac{\delta p_i}{2} - p_i\ \dot{g}(p_i)\frac{\delta p_i}{2} \tag{32}$$

which can be rearranged in the form

$$\mathcal{D}_\gamma(P + \delta P||P) \simeq \sum_i \dot{g}(p_i)\frac{(\delta p_i)^2}{2} \tag{33}$$

2.4 Weighted γ-Divergences

In several contexts, it is be useful to assign distinctive relevance to a different probability distribution, for example in Bayesian inference. Here, we propose a way to define a generalised γ-Divergence between weighted probability distribution.

Let $\pi_P,\ \pi_Q \geq 0$ with $\pi_P + \pi_Q = 1$ be arbitrary weights for the probability distributions P and Q. We can see in equation 11 a natural assignment of weights are $t = \pi_Q$ and $(1 - t) = \pi_P$ obtaining:

$$\mathcal{D}_\gamma^{\pi_P,\pi_Q}(P||Q) = \sum_i \pi_P\ p_i\ g(p_i) + \pi_Q\ q_i\ g(q_i) - m_i\ g(m_i) = \tag{34}$$

$$= \sum_i \pi_P\ p_i\big(g(p_i) - g(m_i)\big) + \pi_Q\ q_i\big(g(q_i) - g(m_i)\big) \tag{35}$$

where $m_i = \pi_P\ p_i + \pi_Q\ q_i$. This assignment assures that $\mathcal{D}_\gamma^{\pi_P \pi_Q}(P||Q) \geq 0$, since

$$\gamma_g^{\pi_P\ \pi_Q}(p_i, q_i) = \pi_P\ p_i\ g(p_i) + \pi_Q\ q_i\ g(q_i) - m_i\ g(m_i) \geq 0 \qquad \forall\ i : 1, \ldots, N \tag{36}$$

2.5 Generalization for more than two distributions

It is also possible to generalize the γ-Divergence for more than two probability distributions. Let $f(\mathbf{x})$, $\mathbb{R}^N \to \mathbb{R}$ is a convex function and let $\mathbf{x} = \{x_1, \ldots, x_N\}$ are the values in the domain of f. Now, using Jensen's inequality (see theorem 2.6.2 [12]) we have

$$f\left(\sum_{k=1}^N \pi_k\ x^k\right) \leq \sum_{k=1}^N\ \pi_k\ f(x^k) \tag{37}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2020 doi:10.20944/preprints202011.0388.v1

where $\pi_k \in [0,1]$ and satisfying $\sum_{k=1}^{N} \pi_k = 1$. Then using the definition of $f(x) := x \cdot g(x)$, we have

$$\sum_{k=1}^{N} \pi_k \; x^k \; g(x^k) - \left(\sum_{k=1}^{N} \pi_k \; x^k \right) \; g\left(\sum_{k=1}^{N} \pi_k \; x^k \right) \geq 0 \tag{38}$$

We can interpret the left side is just the extension of $\gamma_g^{\pi_Q \pi_P}(q_i, p_i)$ for N probability distribution. In the same way, we can assign the $\{\pi_1, \ldots, \pi_N\}$ as weights of the distributions, and defining as $\gamma_g^{\pi_1,\ldots,\pi_N}(p^1, \ldots, p^N)$. Then,

$$\mathcal{D}_\gamma^{\pi_1,\ldots,\pi_N}(P^1, \ldots, P^N) := \sum_{i=1}^{N} \gamma_g^{\pi_1,\ldots,\pi_N}(p_i^1, \ldots, p_i^N) = \sum_{i}^{N} \left[\left(\sum_{k=1}^{N} \pi_k \; p_i^k \; g(p_i^k) \right) - m_i \; g(m_i) \right] \geq 0 \tag{39}$$

where $m_i := \sum_{k=1}^{N} \pi_k \; p^k$. If we take $g(x) = \ln x$ we get the generalized $\mathcal{D}_{JS}$ [13]

$$\mathcal{D}_{JS}^{\pi_1,\ldots,\pi_N}(P^1, \ldots, P^N) = H_S \left(\sum_{k=1}^{W} \pi_k P^k \right) - \sum_{k=1}^{W} \pi_k H_S(P^k). \tag{40}$$

3 Introduction of the new generalized entropy

Entropy can be viewed as the significant amount in the information known of a system. It is possible to generalize the concept of entropy proposed by Shannon [14], giving the fundamental characteristics that an entropy should have in general. There are many ways of introducing a generalized entropy. In the cases of the Havrda-Charvat-Tsallis (HCT) and Renyì entropies [15–17] we sought to generalize the concept of entropy through a parameter. For Salicru's entropy [18], the intention was to introduce a set of entropies through two functions h and ϕ. In our case we look for an entropy related to γ-Divergence described before.

Let $P = \{p_i\}_{i=1}^{n}$ a probability distribution function and $H_G[P]$ a functional of P. We can say H_G is a *generalized entropy*, if complies the following properties:

- to be continuous for each p_i
- to be non-negative
- Verify the identity $H_G[P, 0] = H_G[P]$
- to be equal to zero in the deterministic case, i.e., $H_G[P] = 0$ for $p_i = 1$ and $p_j = 0, \quad \forall j \neq i$
- reach the maximum when the distribution is uniform $P = U$, i.e., when $p_i = \frac{1}{N} \; \forall i$
- to be concave respect to its argument

The last properties allows to define a "Jensen-like" divergence in the following way:

$$D_G(P||Q) = H_G \left(\frac{P+Q}{2} \right) - \frac{1}{2} H_G[P] - \frac{1}{2} H_G[Q] \tag{41}$$

Let $R = \{r_i\}_{i=1}^{N}$ and $P = \{p_i\}_{i=1}^{N}$ two different probability distribution. We said $R \succ P$ (R majorises P) a if

$$\sum_{i=1}^{N-1} r_i \geq \sum_{i=1}^{N-1} p_i \tag{42}$$

with

$$\sum_{i=1}^{N} r_i = \sum_{i=1}^{N} p_i. \tag{43}$$

From the Karamata theorem we have that for any convex function f,

$$\sum_{i=1}^{n} f(r_i) \geq \sum_{i=1}^{N} f(p_i) \tag{44}$$

Now if $f(x) = x \cdot g(x)$, R is defined as $R = \{1, 0, .., 0\}$ and P is any distribution, using equation 44 we have

$$g(1) \geq \sum_{i=1}^{N} p_i \; g(p_i) \tag{45}$$

On the other hand, let $U = \{1/N, ..., 1/N\}$ the uniform distribution, using the Jensen's inequality

$$f\left(\frac{1}{N} \sum_{i=1}^{N} p_i \right) \leq \frac{1}{N} \sum_{i=1}^{N} \; f(p_i) \tag{46}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2020 doi:10.20944/preprints202011.0388.v1

$$\frac{1}{N}g(1/N) \leq \frac{1}{N}\sum_{i=1}^{N} p_i\ g(p_i) \tag{47}$$

$$n\frac{1}{N}g(1/N) \leq \sum_{i=1}^{N} p_i\ g(p_i) \tag{48}$$

$$\sum_{i=1}^{N}\frac{1}{N}g(1/N) \leq \sum_{i=1}^{N} p_i\ g(p_i) \tag{49}$$

the function $\sum_{i=1}^{N} p_i\ g(p_i)$ is minimum when the distribution is uniform. This result shows that the function

$$H_g[P] = g(1) - \sum_{i=1}^{N} p_i\ g(p_i) \tag{50}$$

has a maximum when the distribution is uniform $U = \{1/N, ..., 1/N\}$ and is equal to zero when is $R = \{1, 0, ..., 0\}$. Moreover, it can be checked the function $H_g[P]$ satisfies

$$H_g[P, 0] = H_g[P] \tag{51}$$

Finally, since $f(x)$ is convex by definition, H_g has the property to be concave.
Let $H_{h,\phi}[P] = h\left(\sum_i \phi(p_i)\right)$ be the entropy defined by Salicrú [18]. Let us particularize by taking h as the identity and ϕ as

$$\phi(p_i) = \frac{g(1)}{N} - p_i\ g(p_i) \tag{52}$$

hen it is direct to show that $H_g[P]$. This tells us that $H_g[P]$ is an example of Salicrú entropy.
Using the definition 41 and replacing H_G with H_g we obtain

$$D_g(P||Q) = g(1) - \sum_i \frac{p_i + q_i}{2}\ g\left(\frac{p_i + q_i}{2}\right) - \frac{1}{2}g(1) + \sum_i \frac{p_i}{2}\ g(p_i) - \frac{1}{2}g(1) + \sum_i \frac{q_i}{2}\ g(q_i) \tag{53}$$

with some algebraic step we obtain

$$D_g(P||Q) = \frac{1}{2}\left(\sum_i p_i\ g(p_i) + q_i\ g(q_i) - (p_i + q_i)\ g\left(\frac{p_i + q_i}{2}\right)\right) \tag{54}$$

This is equal to

$$D_g(P||Q) = \frac{1}{2}D_\gamma(P||Q) \tag{55}$$

This result shows that the generalized entropy H_g is closely related to the γ-Divergence defined in section 2.

4 Applications

In this section we use γ-Divergence as a tool to detect dynamic changes in artificially generated sequences and real electrophysiological signals.

4.1 Dynamics changes detection

We used Monte Carlo simulation to study the efficiency of γ-Divergence to detect dynamical changes in binary sequences artificially generated. For this purpose, we generated M sequences composed by two sub-sequences $s^i = s_1^i + s_2^i$ for $i = 1, ..., M$ with length $L = L_{s_1} + L_{s_2}$. Each sub-sequence had a probability distribution $P_{s_1} = [p_{s_1}\ 1 - p_{s_1}]$ and $P_{s_2} = [p_{s_2}\ 1 - p_{s_2}]$ (Fig. 1A). We defined a pointer ρ which moves step by step across the sequences. In each step, we took two sequences with the same length (L_{win}), one on the right (s_r) and other on left (s_l) of ρ (Fig. 1B). We estimated the probability distribution for both sequences P_{sl} and P_{sr}, and calculated the γ-Divergence between then $D_\gamma = D_\gamma(P_{sl}||P_{sr}))$. This procedure was repeated for each $\rho \in [L_{win+1}\ L - L_{win}]$ (Fig. 1C). The point (ρ^*) where the Divergence reaches the maximum value $D_\gamma(\rho^*) = D_{\gamma_{max}}$ was the transition point between the sequence s_1 to s_2. We did this analysis for the M generated sequences and then calculated their mean value and standard deviation. Beyond that this was developed for binary sequences, can be applied to all type of discreet or continuous (with previous quantification) sequences.

Figure 2 shows the γ-Divergence for four specific $g(x)$ applied over a combined binary sequence. The first $s_1 = 3000$ points was generated with a probability $P = [0.5\ 0.5]$ and the following $s_2 = 3000$ points with $Q = [0.4\ 0.6]$. The analysis was made using the function ($g(x) = e^x$, $log(x)$,$\sqrt{x}$, $sinh(x)$) all of which satisfies the formal condition given in the theorem 2. We used a windows length of $L_{win} = 1000$ data point. The fill line represents the D_γ mean value, and the shadows are the standard deviation over the $M = 1000$ realization. Vertical dash line determines the point where sequence 1 and 2 are joined. For a better visualization of the results, the domain of the function goes from the domain [1001 5000] of the original sequence since the first and last 1000 points the D_γ are zero. We can see for all function $g(x)$ that the maximum

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2020 doi:10.20944/preprints202011.0388.v1

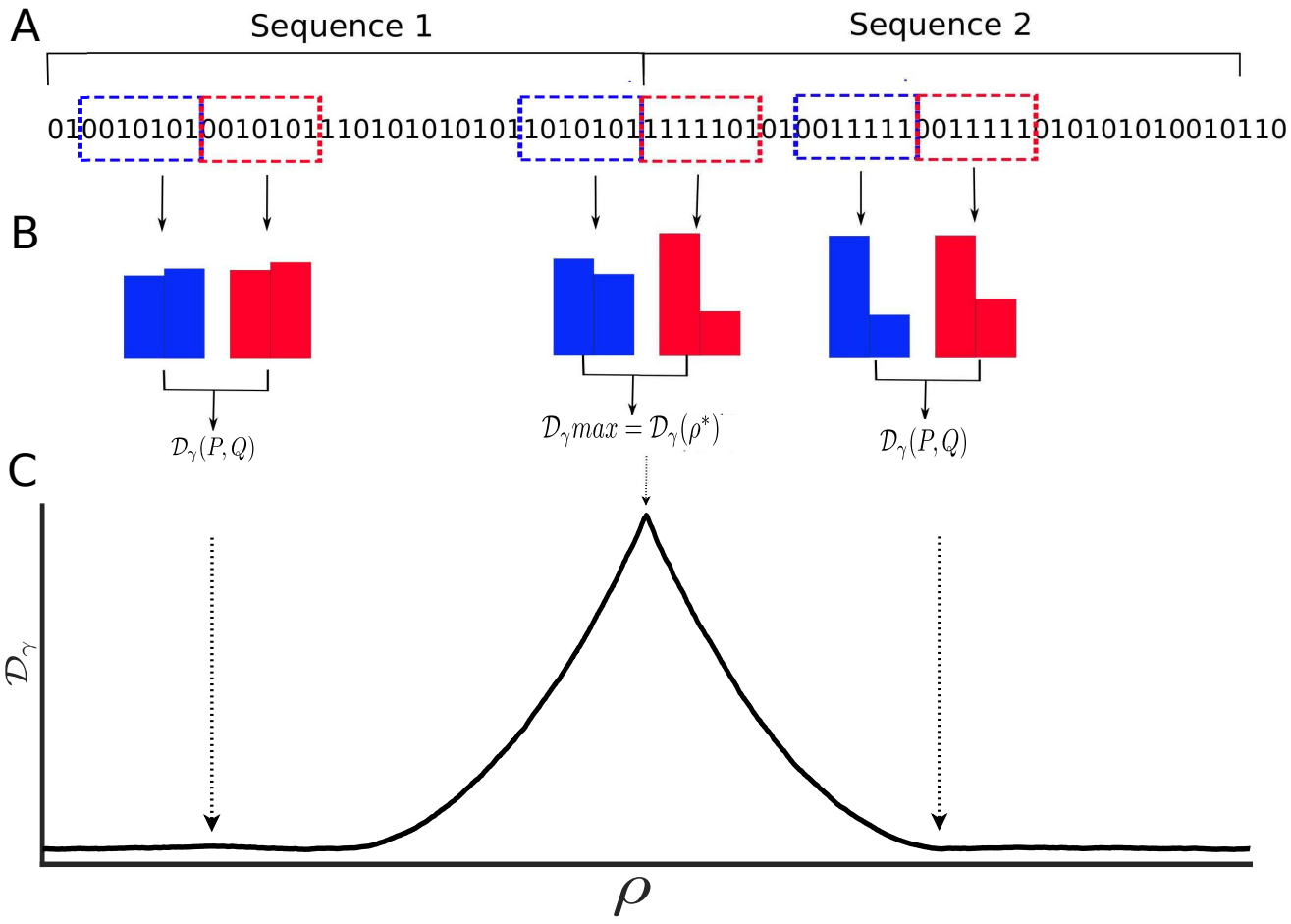

Figure 1: Window method used to detect distribution changes in a sequence. A) Generation of M combined sequences of $s = s_1 + s_2$, with s_1 probability distribution equal to P_{s_1} and s_2 to P_{s_2}. B) The pointer (ρ) moves over the sequence. Two subsequences are taken, on the left s_l (blue) and on the right s_r (red) of the ρ position. Then the probability distribution was calculated for each subsequences and γ-Divergence is calculated. C) The same procedure is performed for all values of $\rho \in [L_{win+1},\ L - L_{win}]$; the point ($\rho^*$) at which $D_\gamma = \mathcal{D}_{\gamma max}$ determines the transition from one sequence to another.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2020 doi:10.20944/preprints202011.0388.v1

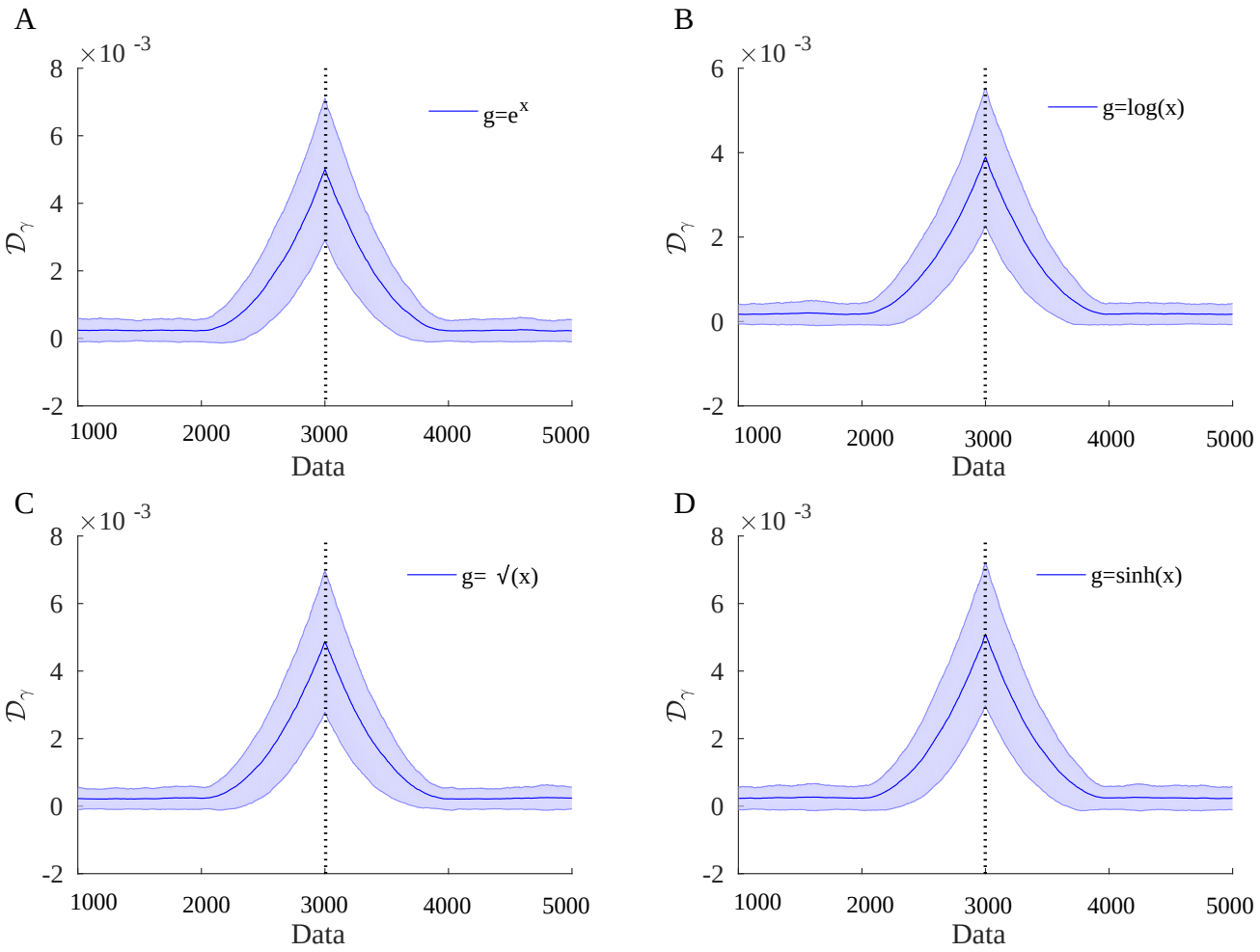

Figure 2: γ-Divergence analysis over a combined binary sequence. The first 3000 points correspond to s_1 with probability distribution $P = [0.5\ 0.5]$ and the latter s_2 with probability distribution $Q = [0.4\ 0.6]$. The windows length are $L_{win} = 1000$. The full line corresponds to the D_γ means that the values and the shadow represent the standard derivation over $M = 1000$ realization. The vertical line represents the exact point where the two sequences join. The same analysis was made for different function $g(x)$:, A) e^x, B) $\log(x)$, C) $\sqrt(x)$ and D) $sinh(x)$. In all cases the $D_{\gamma_{max}}$ is reached at the exact point where the two sequences join.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2020 doi:10.20944/preprints202011.0388.v1

divergence ($D\gamma_{max}$) is reached at the exact point where the sequence changes the probability distribution. In all cases, maximum divergence values are much higher than the standard deviations, demonstrating its statistical significance.

Then, we wanted to study the detection limit of the γ-Divergence. In other words, for binary sequences we would like to saw what the smallest probability distribution difference was that can be detected by this family of divergences. To fulfil these aims we generated four combine binary sequences with probability distribution increasingly closer. We analyzed the sequences with the function ($g(x) = e^x$, $log(x)$,$\sqrt{x}$, $sinh(x)$). The windows length used were $L_{win} = 1000$ data point. Fig. 3 shows the analysis for function $g(x) = e^x$.

As the two probability distribution (P and Q) become closer, the maximum divergence value ($D_{\gamma_{max}}$) decreases, being in the limit of detection for $P = [0.5\ 0.5]$ and $Q = [0.45\ 0.55]$ (Fig. 3C) and, making it impossible to distinguish between them when $P = [0.5\ 0.5]$ and $Q = [0.51\ 0.49]$ (Fig. 3D). Similar results we obtained for the function $\log(x)$,$\sqrt{x}$, $sinh(x)$.

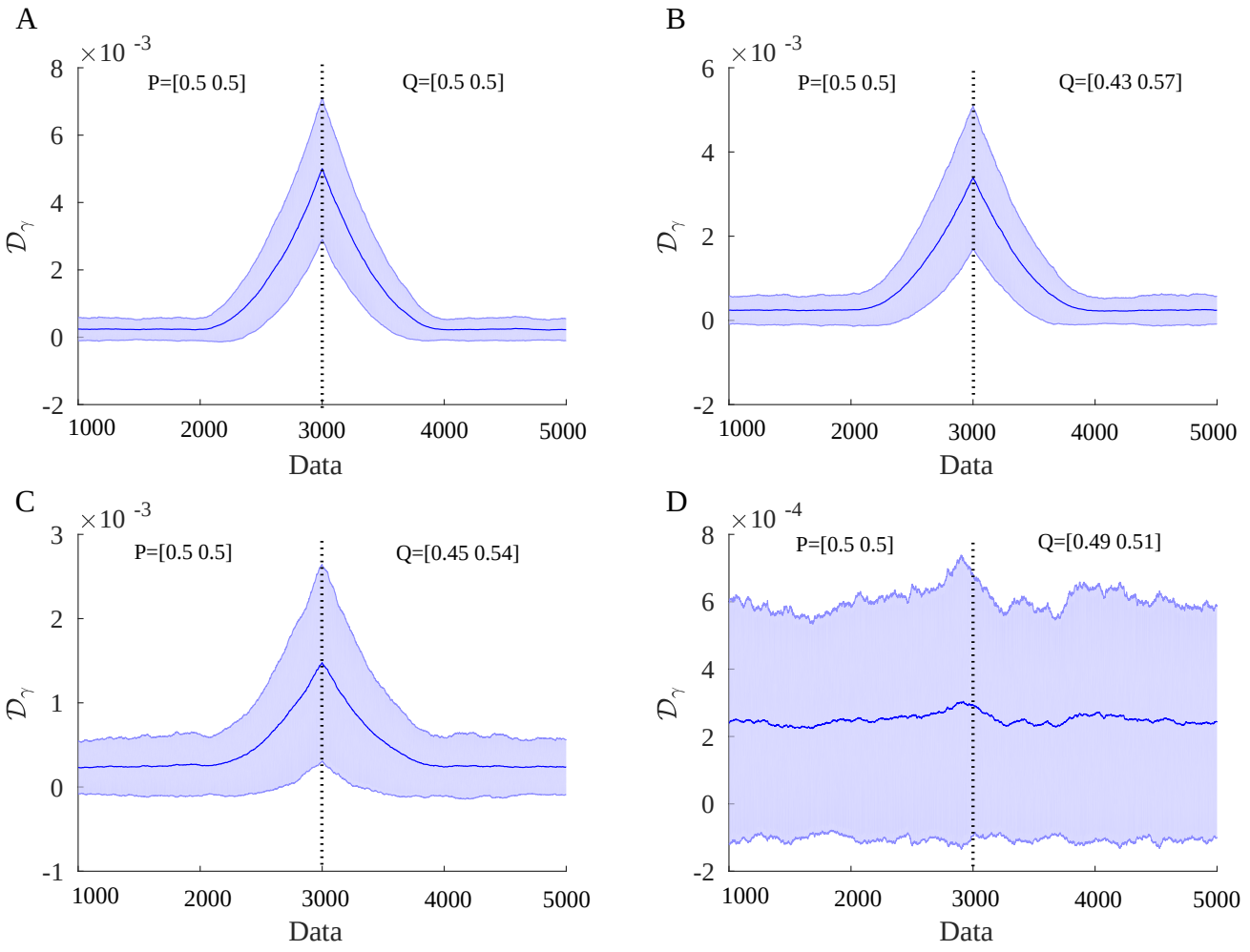

Figure 3: γ-Divergence analysis over a combined binary sequence with diferent P and Q. The first $s_1 = 3000$ points have a P probability distribution and the latter $s_2 = 3000$ have Q probability distribution. The function used for the analysis was $g(x) = e^x$ and the windows length were $L_{win} = 1000$. The vertical line represents the exact point where the two sequences join. A,B As P and Q become more similar, the divergence maximum ($D_{\gamma_{max}}$) becomes smaller; however, detection remains significant. C) For $P = [0.5\ 0.5]$ and $Q = [0.45\ 0.55]$ detection between the two signals is observed, but the statistical significance lies in the limit. D) For $P = [0.5\ 0.5]$ and $Q = [0.49\ 0.51]$ The divergence can't distinguish between the two sequences.

4.2 Transition detection over EEG sleep recording

In the second example, we used the γ-Divergence to identify the transition between sleep states in an electroencephalogram (EEG) signal from a sleep patient.

Sleeping is a dynamic activity, during which many processes are vital to health and well-being take place. It is essential to help to maintain mood, memory, and cognitive performance [19–21]. A specialist defines two primary sleep stages, REM and non-REM. Non-REM stage is composed by three stages, N1 and N2 that are light sleep, where you drift in and out of sleep and can be awakened easily. Stage N3 has slow brain waves and this is the deepest sleep stage, which resembles a coma state. REM stage (for rapid eyes moves) is an active period of sleep caused by intense brain activity. Brain waves

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2020 doi:10.20944/preprints202011.0388.v1

are fast and desynchronized comparable to those in the waking state, this is also the stage in which most dreams take place. These four stages progress cyclically, from N1 through REM then begin again in N1. It is very important that these cycles are maintained for health. Developing tools that can detect the changes in dynamic sleep stages over EEG signal are highly essential for studying patients with sleep disorders [22]. We used our γ-Divergence to detect changes in the dynamics of the EEG to allow us to distinguish between one state from another.

The data were taken from the *Physionet database: The Sleep-EDF Database [Expanded]* [23, 24], and are freely available at [25]. The EEG were recorded (Fpz-Cz) bipolar channel and the sampling frequency was 100 Hz. Initially, we extracted segments from the original signal belonging to the five different sleep states: Awake, REM, N1, N2, N3 [1]. Each segments had 6000 points (corresponding to 60 sec of recording) and were joined into a single signal [2] Fig. 4A. The signal was pre processed with a band-pass filter between $0.5 - 60$ Hz. We quantify the signal using the permutation vector method approach [26], with parameter $d = 4$ and $\tau = 1$. We applied the γ-Divergence following the method used in the previous section for the binary sequence. The functions used were $(g(x) = e^x,\ log(x), \sqrt{x},\ sinh(x))$.

Figure 4B shows that the divergence detects the transition between different sleep states for all the functions $g(x)$. The $\log(x)$ function shows the highest values and the best differentiation between stages.

The transition between Awake-REM and N2-N3 are more remarkable than in REM-N1 and N1-N2. There is a significant differentiation between N2-N3 because of N3 is the deepest sleep stage. In this stage the body becomes more insensitive to outside stimuli than seeming to coma state. In this state, the EEG signal is mostly composed by slow waves (Delta and Theta) causing the brain dynamics to be sharply different from the other states. Lower γ-Divergences values were found between N1-N2 states showing that both states share similar characteristics in their dynamics. Particularly, N1 is characterized by drowsiness slowing down the brain waves and muscle activity. While N1 is a period of light sleep during which eye movement stops, the two stages (N1-N2) are very similar with the difference that in N1 occasional bursts of rapid waves (12-14 Hz) called *sleep spindles* appear.

Similar result as before can be found between REM and N1. REM and N1 also present lower values of divergence. This is expected considering that the EEG of REM sleep contains frequencies present in the "awake state and in the lighter stages of sleep N1 [27]. Despite the similarities of the REM and N1, there are still enough differences between them such that statistically different values are obtained. The presence of 11-16 Hz activity (sleep spindles) in N1, and more abundant alpha activity (8-13Hz) in REM sleep means that these two stages present activity at an overlapping frequency range, which explains the proximity of the divergence values obtained. Difficulty in detecting N1 and REM sleep has also been found using other measures [28, 29].

5 Discussion

In the first place, we showed the existence of a close relationship between the square of the Euclidean metric and the Jensen-Shannon divergence establishing that both belong to the same family of functionals. Based on this, we introduced a family of divergences called γ-Divergence that depend on the property of convex functions. We demonstrated that this new divergence family satisfies all the requirements to be a generalize divergence. Then, we studied our divergence for small PDF variations, revealing that the behavior of these are quadratic in relation to the variation introduced. Subsequently, we introduced weights to the divergence to give more importance to a specific distribution, based on the necessity of the problem to analyse. Finally, we could generalize this divergence for more than two probability distributions allowing them to be used in N-dimensional distribution problem, for example: multidimensional signal analysis.

Next, we could define a general entropy based on the properties of convex functions. This entropy includes Shannons entropy, and we showed that is a particular case of a larger family of entropies called Salicrú. We proved that this new entropy satisfies the requirements to be a general entropy in the context of information theory and see the relationship between γ-Divergence and the "Jensen-type divergence through this generalized entropy.

Finally, we applied the γ-Divergence in simulated and real sequences. We showed that all the functions $g(x)$ could detect with high significance the exact point where sequences with different probabilities joined. Moreover, we studied the detection threshold -the point where the divergence can no longer distinguish between the two signals-. Later, we analysed EEG signal from sleep patient. We could detect the points where the signal changes its dynamics due to the change in the state of sleep. Showing it can be an alternative tool for detecting different sleep states.

It is also important to mention that when using divergences, we had consider the significance of the value obtained. This significance is what allows us to say if the distance found is really a real and not just a measure of the statistical fluctuations. In this work we used the standard deviation calculated over M realizations as a measure of significance. However, many times we do not have more than one realization being this method useless. Therefore, a theoretical study of the field is necessary. Some studies on this topic have already been addressed by Grosse et al. for the Jensen-Shannon divergence [30]. Similar studies should be carried out for this family of *gamma*-Divergence. However, this exceeds this work but will be a addressed in the near future.

Finally, we know that the definition of metric is much stronger than divergence, becasuse of that we want to know is the γ-Divergence introduce in this works meets the definition of metric. For this aim, it is necessary to demonstrate that divergence complies with the triangular inequality [1]. To prove this property to a family of divergences is not an easy task. For this reason in a future work we will investigate the requirements that the function $g(x)$ must be fulfil to consider

[1] This was done based on the notes provided by the database

[2] This was done for a better visualization of the results, because the time of each sleep state can vary from seconds to several minutes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 November 2020 doi:10.20944/preprints202011.0388.v1

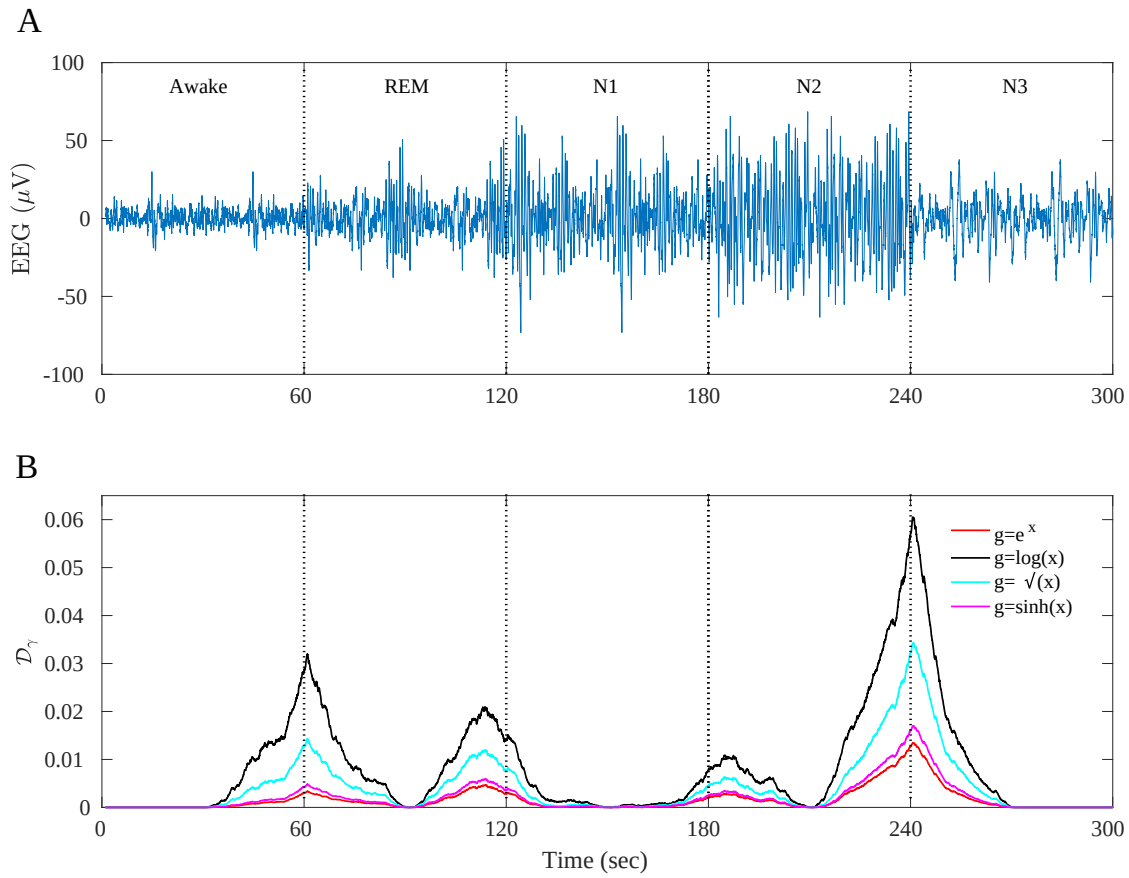

Figure 4: Application the γ-Divergence over a sleep EEG signal using a running windows. A) The EEG signal is composed by 5 sub-signal belonged different sleep stages (Awake, REM, N1, n2, n3). Each state has 6000 point corresponding to 60 sec recording (dashes horizontal lines). B) Using a running windows method the $\mathcal{D}_\gamma$ was applied taken four $g(x)$ compared for the study. The signal was quantified with the permutation vectors with parameter $d = 4$ and $\tau = 1$. For all functions, the maximum values $\mathcal{D}_{\gamma_{max}}$ were reached in the exact point where a transition between sleep states exists.

a *gamma*-Divergence as a metric. In addition, we will study other information measures, such as relative entropy, mutual information and transfer of entropy, based on generalized entropy defined in this paper.

References

[1] T.M. Cover and J.A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[2] M. Parry and E. Fischbach. Probability distribution of distance in a uniform ellipsoid: Theory and applications to physics. *Journal of Mathematical Physics*, 41(4):2417–2433, 2000.

[3] S.N. Ondimu and H. Murase. Effect of probability-distance based markovian texture extraction on discrimination in biological imaging. *Computers and Electronics in Agriculture*, 63(1):2–12, 2008.

[4] A. Kostin. Probability distribution of distance between pairs of nearest stations in wireless network. *Electronics Letters*, 46(18):1299–1300, 2010.

[5] E.M.F. Curado and C. Tsallis. Generalized statistical mechanics: connection with thermodynamics. *Journal of Physics A: Mathematical and General*, 24(2):L69, 1991.

[6] A.P. Majtey, P.W. Lamberti, and D.P. Prato. Jensen-shannon divergence as a measure of distinguishability between mixed quantum states. *Physical Review A*, 72(5):052310, 2005.

[7] M.E. Pereyra, P.W. Lamberti, and O.A. Rosso. Wavelet jensen–shannon divergence as a tool for studying the dynamics of frequency band components in eeg epileptic seizures. *Physica A: Statistical Mechanics and its Applications*, 379(1):122–132, 2007.

[8] T. M Osán, D.G. Bussandri, and P.W. Lamberti. Monoparametric family of metrics derived from classical jensen–shannon divergence. *Physica A: Statistical Mechanics and its Applications*, 495:336–344, 2018.

[9] A.P. Majtey, P.W. Lamberti, and A. Plastino. A monoparametric family of metrics for statistical mechanics. *Physica A: Statistical Mechanics and its Applications*, 344(3-4):547–553, 2004.

[10] S.M. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[11] D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[12] R. Steven. Convex sets and their applications, 1982.

[13] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[14] C.E. Shannon. Communication theory of secrecy systems. *Bell system technical journal*, 28(4):656–715, 1949.

[15] J. Havrda and F. Charvát. Quantification method of classification processes. concept of structural *a*-entropy. *Kybernetika*, 3(1):30–35, 1967.

[16] C. Tsallis. Generalized entropy-based criterion for consistent testing. *Physical Review E*, 58(2):1442, 1998.

[17] A. Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[18] M. Salicru, M.L. Menendez, D. Morales, and L. Pardo. Asymptotic distribution of (h, φ)-entropies. *Communications in Statistics-Theory and Methods*, 22(7):2015–2031, 1993.

[19] R.M. Benca, M. Okawa, M. Uchiyama, S. Ozaki, T. Nakajima, K. Shibui, and W.H Obermeyer. Sleep and mood disorders. *Sleep medicine reviews*, 1(1):45–56, 1997.

[20] R. Stickgold. Sleep-dependent memory consolidation. *Nature*, 437(7063):1272–1278, 2005.

[21] J.J. Pilcher and A.S. Walters. How sleep deprivation affects psychological variables related to college students' cognitive performance. *Journal of American College Health*, 46(3):121–126, 1997.

[22] G. Tononi and C. Cirelli. Sleep function and synaptic homeostasis. *Sleep medicine reviews*, 10(1):49–62, 2006.

[23] A.L. Goldberger, L. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, RG Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H.E. Stanley. Physiobank, physiotoolkit, and physionet: Circulation. *Discovery*, 101(23):1, 1997.

 doi:10.20944/preprints202011.0388.v1

[24] B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C Kamphuisen, and J.J.L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

[25] Physionet databanck. Sleep-EDF Database Expanded. \url{https://physionet.org/content/sleep-edfx/1.0.0/}.

[26] C. Bandt and B. Pompe. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102, 2002.

[27] N. Nicolaou and J. Georgiou. The use of permutation entropy to characterize sleep electroencephalograms. *Clinical EEG and Neuroscience*, 42(1):24–28, 2011.

[28] Q. Noirhomme, M. Boly, V. Bonhomme, P. Boveroux, C. Phillips, P. Peigneux, Soddu, and Others. Bispectral index correlate with regional cerebral blood flow during sleep. *Archives Italiennes de Biologie*, 147(1/2):51–57, 2009.

[29] D.M. Mateos, J. Gómez-Ramírez, and O.A. Rosso. Using time causal quantifiers to characterize sleep stages. *bioRxiv*, page 550152, 2019.

[30] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H.E. Stanley. Analysis of symbolic sequences using the jensen-shannon divergence. *Physical Review E*, 65(4):041905, 2002.