ISSN 0032-9460, Problems of Information Transmission, 2007, Vol. 43, No. 3, pp. 167–189. © Pleiades Publishing, Inc., 2007.
Original Russian Text © J.-F. Coeurjolly, R. Drouilhet, J.-F. Robineau, 2007, published in Problemy Peredachi Informatsii, 2007, Vol. 43, No. 3, pp. 3–27.

INFORMATION THEORY

Normalized Information-Based Divergences

J.-F. Coeurjolly, R. Drouilhet, and J.-F. Robineau

Université Pierre Mendès-France, Grenoble, France
Jean-Francois.Coeurjolly@upmf-grenoble.fr
Remy.Drouilhet@upmf-grenoble.fr

Received April 11, 2006; in final form, May 16, 2007

Abstract—This paper is devoted to the mathematical study of some divergences based on mutual information which are well suited to categorical random vectors. These divergences are generalizations of the "entropy distance" and "information distance." Their main characteristic is that they combine a complexity term and the mutual information. We then introduce the notion of (normalized) information-based divergence, propose several examples, and discuss their mathematical properties, in particular, in some prediction framework.

DOI: 10.1134/S0032946007030015

1. INTRODUCTION

Shannon's information theory, usually just called *information theory*, was introduced in 1948 [1]. The theory is aimed at providing means for measuring information. More precisely, the amount of information in an object may be measured by its *entropy* and may be interpreted as the length of the description of the object in some encoding way. In the Shannon approach, the objects to be encoded are assumed to be outcomes of a known source. Shannon's theory also provides the notion of *mutual information* (related to two objects), which plays a central role in many applications, from lossy compression to machine learning methods.

Several authors noted that it would be useful to modify the mutual information such that the resulting quantity becomes a metric in a strict sense. As a first example, [2, 3] introduced the *entropy distance* defined as the sum of conditional entropies. Other interesting measures are the *information distance* [4] and its normalized version, the *similarity metric*, introduced in [5] in the context of the Kolmogorov complexity theory. More precisely, the information distance is defined as the maximum of the conditional Kolmogorov complexities. The similarity metric is universal in the sense defined by the authors and is not computable since it is based on an uncomputable notion of the Kolmogorov complexity.

Recent papers have demonstrated that applications of suitable versions of the similarity metric are of use in areas as diverse as genomics, virology, languages, literature, music, handwritten digits, and astronomy [6]. To apply the metric to real data, the authors have to replace the use of the non-computable Kolmogorov complexity with an approximation obtained by using standard real-world compressors: GenCompress for genomics [7], the *Normalized Compression Distance* (NCD) for music clustering [8], and the *Normalized Google Distance* (NGD) for automatic meaning discovery [9] are examples of effective compressors. To include the information distance and similarity metric in a framework based on information theory concepts, we make use of the principle that *the expected Kolmogorov complexity equals the Shannon entropy*; an interested reader is referred to [10–12] for more details.

Consequently, both the entropy distance and information distance are expressed in terms of conditional entropies: the first one as their sum and the second as their maximum. In [13] there is given a proof of the triangle inequality for these distances and their respective normalized versions.

In the supervised learning framework, the use of some method of selecting covariables among a large number is required when it is assumed that the data size is too small with respect to the number of available covariables (in order to apply any existing discriminant analysis method). Such a problem has been widely treated (see, e.g., [14]). The approach undertaken in [15] is mainly based on three kinds of methodological tools. The first one is a supervised quantization method consisting in the simplification of too complex covariables (in particular, with a too large number of possible values). Indeed, our main belief is that in order to predict the class variable generally representing a small number of categories of data, each possibly predictive covariable must be not too complex. The second one is a more usual step-by-step selection method combining the simplified covariables together in order to detect a cluster of data of the same class. The last one is aimed at detecting redundancy among the set of covariables. These three tasks can be realized using the entropy or information distances (or their normalized versions). Let us emphasize some properties allowing one to understand the usefulness of these criteria in such a context. The entropy and information distances, D^E and D^I, can be rewritten as the difference between some term (respectively, the joint entropy and the maximum of the marginal entropies) and the mutual information. The first term may be interpreted as a complexity term. Moreover, both are independence measures with the particular property to be minimal (in fact equal to 0) when random vectors share exactly the same information. In [15] it was then proposed to extend the definition of the entropy and information distances by introducing the notion of information-based divergence $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ between two categorical random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, defined as the difference of some complexity term $C_{\boldsymbol{X},\boldsymbol{Y}}$ and the mutual information $I_{\boldsymbol{X},\boldsymbol{Y}}$ and such that $C_{\boldsymbol{X},\boldsymbol{Y}}$ is an upper bound for $I_{\boldsymbol{X},\boldsymbol{Y}}$ reached when $\boldsymbol{X}$ and $\boldsymbol{Y}$ share exactly the same information. The notion of the normalized information-based divergence $\delta_{\boldsymbol{X},\boldsymbol{Y}}$ is directly derived by dividing the associated information-based divergence $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ by the complexity term $C_{\boldsymbol{X},\boldsymbol{Y}}$. The normalized versions d^E and d^I of D^E and D^I are particular examples. Other examples are given in [15]. Among them, one is of particular interest since its complexity term C^S is the mean of marginal entropies. The associated (unnormalized) information-based divergence Δ^S is not so different from D^E since it corresponds to its half. Nevertheless, the expression for its complexity term C^S really differs from the complexity term C^E of D^E (i.e., the joint entropy). For practical purposes, we may argue that D^I, D^E, and Δ^S are not well suited in the prediction framework since a small value of these distances means that both the explained and explicative variables have a good knowledge of each other. This is due to the fact that both conditional entropies have at least the same weight.

In this paper, this drawback is weakened by introducing a natural extension $C^{S,\alpha}$ of the complexity term C^S defined as a weighted mean (by α and $1-\alpha$ for some $0 < \alpha \leq 1$) of the minimum and maximum of marginal entropies. This kind of complexity term leads to an expected I-divergence $\Delta^{S,\alpha}$, which is the weighted mean of the minimum and maximum of conditional entropies.

The paper is organized as follows. In Section 2 we recall the definitions and the main properties of the entropy and information distances (and their normalized versions). Similarly to [16], we extract the main characteristics to define some general concept of information divergence which could theoretically be applied in a more general setting (continuous, discrete, etc.). In Section 3 we concentrate on categorical data (and in particular discrete) random vectors, since this is usually the case in most applications that use entropy or information distance. We give the definition of the (normalized) information-based divergence and propose several examples. We study their mathematical properties in the general context and propose some sufficient conditions for these divergences to verify some triangle-type inequality. Finally, in Section 4 we exhibit some properties

of information-based divergences in the special prediction framework. In particular, we show that these divergences are useful to detect redundancy.

2. NORMALIZED ENTROPY DISTANCE AND NORMALIZED INFORMATION DISTANCE

Let us denote by Γ the set of categorical random vectors, that is, discrete-valued random vectors with finite entropy. In the following, $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ are three elements of Γ.

2.1. Notation

We denote by $H_{\boldsymbol{X}}$ (when it exists) the Shannon entropy of $\boldsymbol{X}$ given by

$$H_{\boldsymbol{X}} = -\sum_{\boldsymbol{x}\in\Omega_{\boldsymbol{X}}} p_{\boldsymbol{X}}(\boldsymbol{x})\log(p_{\boldsymbol{X}}(\boldsymbol{x})), \quad \text{with} \quad p_{\boldsymbol{X}}(\boldsymbol{x}) = \mathbf{P}(\boldsymbol{X}=\boldsymbol{x}).$$

In the same way, one can define the joint entropy of $\boldsymbol{X}$ and $\boldsymbol{Y}$, denoted by $H_{\boldsymbol{X},\boldsymbol{Y}}$, and the conditional entropy of $\boldsymbol{X}$ by $\boldsymbol{Y}$ (respectively, $\boldsymbol{Y}$ by $\boldsymbol{X}$), denoted by $H_{\boldsymbol{X}|\boldsymbol{Y}}$ (respectively, $H_{\boldsymbol{Y}|\boldsymbol{X}}$). Finally, we denote by $I_{\boldsymbol{X},\boldsymbol{Y}}$ the mutual information between the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. When these different quantities exist, there are the following relations (see, e.g., [17]):

$$H_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X}} + H_{\boldsymbol{Y}|\boldsymbol{X}} = H_{\boldsymbol{Y}} + H_{\boldsymbol{X}|\boldsymbol{Y}}, \tag{1}$$

$$I_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X}} - H_{\boldsymbol{X}|\boldsymbol{Y}} = H_{\boldsymbol{Y}} - H_{\boldsymbol{Y}|\boldsymbol{X}} = H_{\boldsymbol{X}} + H_{\boldsymbol{Y}} - H_{\boldsymbol{X},\boldsymbol{Y}}. \tag{2}$$

2.2. Definitions and Some Basic Properties

We now present some measures allowing us to overcome some drawbacks of the mutual information. As a first generalization, several authors noted that it would be useful to modify the mutual information such that the resulting quantity becomes a metric in a strict sense. Two such measures exist and are well known in the literature. The first one, called the "entropy distance," is derived from the domain of information theory. The second one, called "information distance," originates in works around the Kolmogorov complexity. These measures are defined (when they exist) for two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ as follows:

- Entropy distance
$$D^E_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X}|\boldsymbol{Y}} + H_{\boldsymbol{Y}|\boldsymbol{X}}; \tag{3}$$
- Information distance
$$D^I_{\boldsymbol{X},\boldsymbol{Y}} = \max(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}}). \tag{4}$$

Both measures are indeed some modifications of the mutual information since from (1) and (2) we have

$$D^E_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X},\boldsymbol{Y}} - I_{\boldsymbol{X},\boldsymbol{Y}} \qquad \text{and} \qquad D^I_{\boldsymbol{X},\boldsymbol{Y}} = \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}) - I_{\boldsymbol{X},\boldsymbol{Y}}. \tag{5}$$

The quantities $H_{\boldsymbol{X},\boldsymbol{Y}}$ and $\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ are upper bounds for the mutual information $I_{\boldsymbol{X},\boldsymbol{Y}}$ and are reached when $\boldsymbol{X}$ and $\boldsymbol{Y}$ share exactly the same information. In other words, these two measures are nonnegative and vanish if and only if $H_{\boldsymbol{Y}|\boldsymbol{X}} = H_{\boldsymbol{X}|\boldsymbol{Y}} = 0$, expressing the fact that $\boldsymbol{X}$ (respectively, $\boldsymbol{Y}$) predicts $\boldsymbol{Y}$ (respectively, $\boldsymbol{X}$) with probability 1.

These measures satisfy

$$D^E_{\boldsymbol{X},\boldsymbol{Y}} \le H_{\boldsymbol{X},\boldsymbol{Y}} \qquad \text{and} \qquad D^I_{\boldsymbol{X},\boldsymbol{Y}} \le \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}), \tag{6}$$

where the equality holds if the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. In [18, 19] it was noted that in bioinformatics an unnormalized distance may not be a proper evolutionary distance measure. To overcome this problem within the algorithmic framework, they form two normalized versions of distances, D^E and D^I. Their Shannon versions were proposed and studied in [13].

Definition 1. When they exist, one defines the following two measures:

- Normalized entropy distance
$$d^E_{\boldsymbol{X},\boldsymbol{Y}} = \frac{H_{\boldsymbol{X}|\boldsymbol{Y}} + H_{\boldsymbol{Y}|\boldsymbol{X}}}{H_{\boldsymbol{X},\boldsymbol{Y}}};$$
- Normalized information distance
$$d^I_{\boldsymbol{X},\boldsymbol{Y}} = \frac{\max(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}})}{\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})}.$$

Since $H_{\boldsymbol{X},\boldsymbol{Y}} = 0 \Leftrightarrow H_{\boldsymbol{X}} = H_{\boldsymbol{Y}} = 0 \Leftrightarrow \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}) = 0$, we set by convention $d^E_{\boldsymbol{X},\boldsymbol{Y}} = 0$ (respectively, $d^I_{\boldsymbol{X},\boldsymbol{Y}} = 0$) when $H_{\boldsymbol{X}} = H_{\boldsymbol{Y}} = 0$.

We are encouraged to define the following equivalence class: the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be equivalent if $\boldsymbol{X}$ (respectively, $\boldsymbol{Y}$) predicts $\boldsymbol{Y}$ (respectively, $\boldsymbol{X}$) with probability 1; we denote

$$\boldsymbol{X} \sim \boldsymbol{Y} \quad \Longleftrightarrow \quad H_{\boldsymbol{Y}|\boldsymbol{X}} = H_{\boldsymbol{X}|\boldsymbol{Y}} = 0 \quad \Longleftrightarrow \quad I_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X}} = H_{\boldsymbol{Y}}. \tag{7}$$

Due to the previous convention,

$$d^E_{\boldsymbol{X},\boldsymbol{Y}} = 0 \quad \Longleftrightarrow \quad d^I_{\boldsymbol{X},\boldsymbol{Y}} = 0 \quad \Longleftrightarrow \quad \boldsymbol{X} \sim \boldsymbol{Y}.$$

From (1) and (2), one can obtain the following expressions for these two measures, allowing some new interpretations.

Proposition 1. *We have the following expressions for* $d^E_{\boldsymbol{X},\boldsymbol{Y}}$ *and* $d^I_{\boldsymbol{X},\boldsymbol{Y}}$:

$$d^E_{\boldsymbol{X},\boldsymbol{Y}} = 1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{H_{\boldsymbol{X},\boldsymbol{Y}}}, \tag{8}$$

$$d^I_{\boldsymbol{X},\boldsymbol{Y}} = 1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})} \tag{9}$$

$$= \max\left(\frac{H_{\boldsymbol{X}|\boldsymbol{Y}}}{H_{\boldsymbol{X}}}, \frac{H_{\boldsymbol{Y}|\boldsymbol{X}}}{H_{\boldsymbol{Y}}}\right). \tag{10}$$

Proposition 2. *The measures* d^E *and* d^I *constitute two distances bounded by* 1.

To our knowledge, these results were proved in [13]. The proofs are very similar to those of [20], where algorithmic version of these distances were considered. The proof is therefore omitted, but in Section 3.3 we propose a result extending this one in the sense that we give conditions for measures that can be written as (8) and (9) to constitute a metric.

2.3. Concept of Information Divergence

From the previous study related to D^I, D^E, d^I, and d^E, we can exhibit some characteristics useful for an attempt to define the concept of information divergence Δ in a more general setting. Let us first consider a similarity measure $\mathcal{I}_{\boldsymbol{X},\boldsymbol{Y}}$ (not necessarily the mutual information) which is minimal (in fact equal to 0) when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent and maximal (in fact equal to $\mathcal{I}_{\boldsymbol{X},\boldsymbol{X}} = \mathcal{I}_{\boldsymbol{Y},\boldsymbol{Y}}$) when the distributions of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ and $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ are trivial. An information divergence $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ could satisfy the following properties:

[P1] symmetry: $\Delta_{\boldsymbol{X},\boldsymbol{Y}} = \Delta_{\boldsymbol{Y},\boldsymbol{X}}$;
[P2] nonnegativeness: $\Delta_{\boldsymbol{X},\boldsymbol{Y}} \geq 0$;
[P3] $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ is minimal (i.e., $\Delta_{\boldsymbol{X},\boldsymbol{Y}} = 0$) if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ share exactly the same information (i.e., $\mathcal{I}_{\boldsymbol{X},\boldsymbol{Y}}$ is maximal);
[P4] $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ is maximal if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent (i.e., $\mathcal{I}_{\boldsymbol{X},\boldsymbol{Y}} = 0$).

Other supplementary properties could be that $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$:

[P5] is normalized: $\Delta_{\boldsymbol{X},\boldsymbol{Y}} \in [0,1]$ and $\Delta_{\boldsymbol{X},\boldsymbol{Y}} = 1$ when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent;
[P6] satisfies the triangle inequality: $\Delta_{\boldsymbol{X},\boldsymbol{Y}} \leq \Delta_{\boldsymbol{X},\boldsymbol{Z}} + \Delta_{\boldsymbol{Z},\boldsymbol{Y}}$;
[P7] is invariant under continuous and strictly increasing transformations $\boldsymbol{\varphi}(\cdot)$ and $\boldsymbol{\psi}(\cdot)$ of the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ whenever they are quantitative random vectors.

There exists a vast literature discussing criteria that satisfy the stated properties. We may cite [21] or a recent work [16], where it is proposed to detect the dependence between two possibly nonlinear processes through the Bhattacharya–Matusita–Hellinger measure of dependence, given by

$$S_\rho = \frac{1}{2}\iint \left(\sqrt{f_1(\boldsymbol{x},\boldsymbol{y})} - \sqrt{f_2(\boldsymbol{x},\boldsymbol{y})}\right)^2 d\boldsymbol{x}\, d\boldsymbol{y},$$

where f_1 (respectively, f_2) is the joint density (respectively, product of marginal densities) of $\boldsymbol{X}$ and $\boldsymbol{Y}$. This measure, which has another advantage to be applicable to both continuous or discrete variables, satisfies properties **[P1]**–**[P7]** (in fact, let us precise that **[P7]** is only valid if $\boldsymbol{\varphi}(\cdot) = \boldsymbol{\psi}(\cdot)$).

In some framework where the purpose is to predict some reference variable, one may find interesting to work with a divergence $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ which combines the minimization of a nonnegative complexity term denoted by $\mathcal{C}_{\boldsymbol{X},\boldsymbol{Y}}$ and the maximization of a nonnegative information term $\mathcal{I}_{\boldsymbol{X},\boldsymbol{Y}}$. The quantity $\mathcal{C}_{\boldsymbol{X},\boldsymbol{Y}}$ is called a complexity term since it is assumed to be expressed as a function of $\mathcal{H}_{\boldsymbol{X}}$, $\mathcal{H}_{\boldsymbol{Y}}$, and $\mathcal{H}_{\boldsymbol{X},\boldsymbol{Y}}$ measuring in some way the complexity of the vectors $\boldsymbol{X}$, $\boldsymbol{Y}$, and $(\boldsymbol{X},\boldsymbol{Y})$, respectively. In other words, we may expect that an information divergence $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ could also satisfy the following properties:

[P8] When $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the same complexity (in the sense that $\mathcal{C}_{\boldsymbol{Y},\boldsymbol{X}_1} = \mathcal{C}_{\boldsymbol{Y},\boldsymbol{X}_2}$), we have $\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} < \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}$ whenever $\boldsymbol{X}_1$ has a better knowledge about $\boldsymbol{Y}$ than $\boldsymbol{X}_2$ (i.e., $\mathcal{I}_{\boldsymbol{Y},\boldsymbol{X}_1} > \mathcal{I}_{\boldsymbol{Y},\boldsymbol{X}_2}$);
[P9] When $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the same knowledge about $\boldsymbol{Y}$ (i.e., $\mathcal{I}_{\boldsymbol{Y},\boldsymbol{X}_1} = \mathcal{I}_{\boldsymbol{Y},\boldsymbol{X}_2}$), we have $\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} < \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}$ whenever $\boldsymbol{X}_1$ is simpler than $\boldsymbol{X}_2$ in the sense that $\mathcal{C}_{\boldsymbol{Y},\boldsymbol{X}_1} < \mathcal{C}_{\boldsymbol{Y},\boldsymbol{X}_2}$. Moreover, in this particular situation
[P10] $\mathcal{C}_{\boldsymbol{Y},\boldsymbol{X}_1} \leq \mathcal{C}_{\boldsymbol{Y},\boldsymbol{X}_2}$ must be equivalent to $\mathcal{H}_{\boldsymbol{X}_1} \leq \mathcal{H}_{\boldsymbol{X}_2}$;
[P11] When $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ share almost exactly the same information (i.e., $\mathcal{I}_{\boldsymbol{X}_1,\boldsymbol{X}_2}$ is almost maximal and $\Delta_{\boldsymbol{X}_1,\boldsymbol{X}_2} \simeq 0$), the difference between the divergences $\Delta_{\boldsymbol{Y},\boldsymbol{X}_1}$ and $\Delta_{\boldsymbol{Y},\boldsymbol{X}_2}$ is almost zero (i.e., $\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} \simeq \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}$).

A class of candidates that satisfy **[P8]** and **[P9]** could be of the form

$$\Delta_{\boldsymbol{X},\boldsymbol{Y}} = \frac{\mathcal{C}_{\boldsymbol{X},\boldsymbol{Y}} - \mathcal{I}_{\boldsymbol{X},\boldsymbol{Y}}}{\mathcal{W}_{\boldsymbol{X},\boldsymbol{Y}}}, \tag{11}$$

where $\mathcal{W}_{\boldsymbol{X},\boldsymbol{Y}}$ is a positive term. When $\mathcal{W}_{\boldsymbol{X},\boldsymbol{Y}} = \mathcal{C}_{\boldsymbol{X},\boldsymbol{Y}}$, we obtain a normalized information divergence. Properties **[P2]** and **[P3]** and relation (11) imply that $\mathcal{C}_{\boldsymbol{X},\boldsymbol{Y}}$ is an upper bound for $\mathcal{I}_{\boldsymbol{X},\boldsymbol{Y}}$, reached when $\boldsymbol{X}$ and $\boldsymbol{Y}$ share exactly the same information.

In the rest of this paper we concentrate on divergences of the form (11) that are in addition well suited to categorical random variables (and in particular discrete random variables). In such a framework, we shall only describe some entropic-based criteria (i.e., $\mathcal{H}_{\boldsymbol{X}} = H_{\boldsymbol{X}}$), and so the information term will be set to the mutual information $I_{\boldsymbol{X},\boldsymbol{Y}}$.

3. INFORMATION-BASED DIVERGENCES AND THEIR NORMALIZED VERSIONS

3.1. Definition and Examples

Definition 2. Two criteria Δ and δ are called, respectively, an information-based divergence and a normalized information-based divergence (for short, *I*-divergence and *NI*-divergence) if they

can be written as

$$\Delta_{\boldsymbol{X},\boldsymbol{Y}} = C_{\boldsymbol{X},\boldsymbol{Y}} - I_{\boldsymbol{X},\boldsymbol{Y}}, \tag{12}$$

$$\delta_{\boldsymbol{X},\boldsymbol{Y}} = \frac{C_{\boldsymbol{X},\boldsymbol{Y}} - I_{\boldsymbol{X},\boldsymbol{Y}}}{C_{\boldsymbol{X},\boldsymbol{Y}}} = 1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{C_{\boldsymbol{X},\boldsymbol{Y}}}, \tag{13}$$

where the term $C_{\boldsymbol{X},\boldsymbol{Y}}$ constitutes a complexity term satisfying

(i) $C_{\boldsymbol{X},\boldsymbol{Y}} = C_{\boldsymbol{Y},\boldsymbol{X}}$;
(ii) $I_{\boldsymbol{X},\boldsymbol{Y}} \leq C_{\boldsymbol{X},\boldsymbol{Y}}$, and this bound is achieved if and only if the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are equivalent, i.e., if and only if $\boldsymbol{X} \sim \boldsymbol{Y}$.

We set by convention $\delta_{\boldsymbol{X},\boldsymbol{Y}} = 0$ when $C_{\boldsymbol{X},\boldsymbol{Y}} = I_{\boldsymbol{X},\boldsymbol{Y}} = 0$.

This definition implies automatically that an I-divergence $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ (respectively, NI-divergence $\delta_{\boldsymbol{X},\boldsymbol{Y}}$) satisfies properties **[P1]**–**[P4]** (respectively, **[P1]**–**[P5]**). In the rest of the paper, the term $C_{\boldsymbol{X},\boldsymbol{Y}}$ is expressed as

$$C_{\boldsymbol{X},\boldsymbol{Y}} = f_C(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}}, I_{\boldsymbol{X},\boldsymbol{Y}}), \tag{14}$$

where $f_C(\cdot,\cdot,\cdot)$ is a nonnegative function. Under this expression for $C_{\boldsymbol{X},\boldsymbol{Y}}$, property **[P7]** is ensured since the conditional entropies and mutual information depend only on the joint probability distribution of the categorical random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$.

From now on, we propose a series of examples where we adopt the following convention: an I-divergence (respectively, NI-divergence) satisfying the triangle inequality is denoted by D (respectively, d) rather than Δ (respectively, δ). Moreover, each example will be particularized by some discriminating additional letter in the same manner as D^E and D^I (respectively, d^E and d^I), which clearly constitute I-divergences (respectively, NI-divergences).

In [15], we investigate two new entropic criteria naturally expressed by

$$\delta^D_{\boldsymbol{X},\boldsymbol{Y}} = \frac{1}{2}\left(\frac{H_{\boldsymbol{X}|\boldsymbol{Y}}}{H_{\boldsymbol{X}}} + \frac{H_{\boldsymbol{Y}|\boldsymbol{X}}}{H_{\boldsymbol{Y}}}\right) \qquad \text{and} \qquad \delta^S_{\boldsymbol{X},\boldsymbol{Y}} = \frac{H_{\boldsymbol{X}|\boldsymbol{Y}} + H_{\boldsymbol{Y}|\boldsymbol{X}}}{H_{\boldsymbol{X}} + H_{\boldsymbol{Y}}},$$

which can be rewritten as NI-divergences:

$$\delta^D_{\boldsymbol{X},\boldsymbol{Y}} = 1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{C^D_{\boldsymbol{X},\boldsymbol{Y}}} \quad \text{with} \quad C^D_{\boldsymbol{X},\boldsymbol{Y}} = \left(\frac{1}{2}\left(\frac{1}{H_{\boldsymbol{X}}} + \frac{1}{H_{\boldsymbol{Y}}}\right)\right)^{-1}, \tag{15}$$

$$\delta^S_{\boldsymbol{X},\boldsymbol{Y}} = 1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{C^S_{\boldsymbol{X},\boldsymbol{Y}}} \quad \text{with} \quad C^S_{\boldsymbol{X},\boldsymbol{Y}} = \frac{1}{2}(H_{\boldsymbol{X}} + H_{\boldsymbol{Y}}). \tag{16}$$

Their unnormalized versions are expressed as $\Delta^D_{\boldsymbol{X},\boldsymbol{Y}} = C^D_{\boldsymbol{X},\boldsymbol{Y}} - I_{\boldsymbol{X},\boldsymbol{Y}}$ and $D^S_{\boldsymbol{X},\boldsymbol{Y}} = C^S_{\boldsymbol{X},\boldsymbol{Y}} - I_{\boldsymbol{X},\boldsymbol{Y}}$.

In this paper, we are interested in a large family of I-divergences or NI-divergences with complexity terms of the form

$$C^\alpha_{\boldsymbol{X},\boldsymbol{Y}} = g^{-1}(\alpha g(m_{\boldsymbol{X},\boldsymbol{Y}}) + (1-\alpha) g(M_{\boldsymbol{X},\boldsymbol{Y}})), \tag{17}$$

with $m_{\boldsymbol{X},\boldsymbol{Y}} = \min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ and $M_{\boldsymbol{X},\boldsymbol{Y}} = \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ and where $0 \leq \alpha < 1$ and $g(\cdot)$ is any monotone function on $\mathbb{R}^+$. When this is not ambiguous, we set $m = m_{\boldsymbol{X},\boldsymbol{Y}}$ and $M = M_{\boldsymbol{X},\boldsymbol{Y}}$. To be convinced that I-divergences and NI-divergences with complexity terms of the form (17) satisfy Definition 2(ii), let us note that

$$I_{\boldsymbol{X},\boldsymbol{Y}} = g^{-1}(\alpha g(I_{\boldsymbol{X},\boldsymbol{Y}}) + (1-\alpha) g(I_{\boldsymbol{X},\boldsymbol{Y}})) \leq g^{-1}(\alpha g(m) + (1-\alpha) g(M)).$$

When $\alpha = 0$, the complexity term C^α corresponds to C^I. When $\alpha = 1$, the complexity term defined as $\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ and denoted by $C^{\min}_{\boldsymbol{X},\boldsymbol{Y}}$ does not satisfy Definition 2(ii) and therefore **[P3]**. The associated $\Delta^{\min}$ (respectively, $\delta^{\min}$) is not an I-divergence (respectively, NI-divergence).

Now we pay particular attention to the complexity terms $C^{D,\alpha}$, $C^{S,\alpha}$, $C^{R,\alpha}$, and $C^{P,\alpha}$ of the form (17) with, respectively, $g^D(\cdot) = 1/\cdot$, $g^S(\cdot) = \cdot$, $g^R(\cdot) = \sqrt{\cdot}$, and $g^P(\cdot) = \log(\cdot)$:

$$C^{D,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} = \left(\alpha\frac{1}{\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})} + (1-\alpha)\frac{1}{\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})}\right)^{-1}, \tag{18}$$

$$C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} = \alpha\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}) + (1-\alpha)\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}), \tag{19}$$

$$C^{R,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} = \left(\alpha\sqrt{\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})} + (1-\alpha)\sqrt{\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})}\right)^2, \tag{20}$$

$$C^{P,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} = \min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})^{\alpha}\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})^{1-\alpha}. \tag{21}$$

The previous measures Δ^S, δ^S, Δ^D, and δ^D are particular examples of such a family since the value of $\alpha = \frac{1}{2}$ leads to $C^{1/2}_{\boldsymbol{X},\boldsymbol{Y}} = g^{-1}\big(\frac{1}{2}g(H_{\boldsymbol{X}}) + \frac{1}{2}g(H_{\boldsymbol{Y}})\big)$. When $\alpha = \frac{1}{2}$, $\Delta^{\bullet,\alpha}$ and $\delta^{\bullet,\alpha}$ will simply be denoted by $\Delta^{\bullet}$ and $\delta^{\bullet}$, where $\bullet$ stands for S, R, P, and D.

Let us first comment on the particular expressions of the divergences $\Delta^{S,\alpha}$ and $\delta^{D,\alpha}$ associated to $C^{D,\alpha}$ and $C^{S,\alpha}$ given by

$$\begin{aligned}
\Delta^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} &= \alpha\min(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}}) + (1-\alpha)\max(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}}) \\
&= \alpha\Delta^{\min}_{\boldsymbol{X},\boldsymbol{Y}} + (1-\alpha)D^I_{\boldsymbol{X},\boldsymbol{Y}}, \\
\delta^{D,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} &= \alpha\min\left(\frac{H_{\boldsymbol{X}|\boldsymbol{Y}}}{H_{\boldsymbol{X}}}, \frac{H_{\boldsymbol{Y}|\boldsymbol{X}}}{H_{\boldsymbol{Y}}}\right) + (1-\alpha)\max\left(\frac{H_{\boldsymbol{X}|\boldsymbol{Y}}}{H_{\boldsymbol{X}}}, \frac{H_{\boldsymbol{Y}|\boldsymbol{X}}}{H_{\boldsymbol{Y}}}\right) \\
&= \alpha\delta^{\min}_{\boldsymbol{X},\boldsymbol{Y}} + (1-\alpha)d^I_{\boldsymbol{X},\boldsymbol{Y}}.
\end{aligned}$$

Clearly, the previous representation of $\Delta^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}$ (respectively, $\delta^{D,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}$) as a convex combination of $\Delta^{\min}_{\boldsymbol{X},\boldsymbol{Y}}$ and $D^I_{\boldsymbol{X},\boldsymbol{Y}}$ (respectively, $\delta^{\min}_{\boldsymbol{X},\boldsymbol{Y}}$ and $d^I_{\boldsymbol{X},\boldsymbol{Y}}$) introduces a degree of freedom that could be useful for practical purposes in the prediction framework, where $\boldsymbol{Y}$ could represent some class variable. According to the parameter α, one may choose between one or two prediction terms, $H_{\boldsymbol{X}|\boldsymbol{Y}}$ and $H_{\boldsymbol{Y}|\boldsymbol{X}}$ (respectively, $\frac{H_{\boldsymbol{X}|\boldsymbol{Y}}}{H_{\boldsymbol{X}}}$ and $\frac{H_{\boldsymbol{Y}|\boldsymbol{X}}}{H_{\boldsymbol{Y}}}$). This possibility to introduce a nonuniform mixing of the entropic contributions in the expression of the complexity terms seems to be not feasible by a direct adaptation of $C^I_{\boldsymbol{X},\boldsymbol{Y}}$.

Remark 1. By choosing $g(\cdot) = (\cdot)^{\gamma}$ for some $\gamma > 0$, the complexity term is given by $C^{\gamma,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} = \big\|\big(\alpha^{\frac{1}{\gamma}}m, (1-\alpha)^{\frac{1}{\gamma}}M\big)\big\|_{\gamma}$, where $\|\boldsymbol{x}\|_{\gamma} = \Big(\sum\limits_{i=1}^{2}|x_i|^{\gamma}\Big)^{1/\gamma}$ denotes the norm of some vector $\boldsymbol{x}$ of length 2. Note that for any $0 \le \alpha \le 1$, we have

$$(\alpha^{\wedge})^{\frac{1}{\gamma}}\|(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})\|_{\gamma} \le C^{\gamma,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \le (\alpha^{\vee})^{\frac{1}{\gamma}}\|(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})\|_{\gamma},$$

with $\alpha^{\wedge} = \min(\alpha, 1-\alpha)$ and $\alpha^{\vee} = \max(\alpha, 1-\alpha)$. When γ goes to infinity, $C^{\gamma,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}$ converges towards $C^I_{\boldsymbol{X},\boldsymbol{Y}}$.

Remark 2. The complexity term C^{α} is invariant under linear transformations of g. In particular, g and $-g$ provide the same complexity term. Consequently, without loss of generality we could restrict g to be an increasing function.

Let us now propose a result to arrange these different examples considered in this paper. First, some preliminary result is given.

Lemma 1. *Let $C^{(1)}$ and $C^{(2)}$ be two complexity terms of the form* (17) *with functions g_1 and g_2. Assume that either the function $g_1 \circ g_2^{-1}$ is concave or $g_2 \circ g_1^{-1}$ is convex. Then $C^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} \le C^{(2)}_{\boldsymbol{X},\boldsymbol{Y}}$.*

Proof. By rewriting $g_1 = (g_1 \circ g_2^{-1}) \circ g_2$ when $g_1 \circ g_2^{-1}$ is concave and $g_1^{-1} = g_2^{-1} \circ (g_2 \circ g_1^{-1})$ when $(g_2 \circ g_1^{-1})$ is convex, one gets

$$g_1^{-1}(\alpha g_1(m) + (1-\alpha) g_1(M)) \leq \begin{cases} g_2^{-1}(\alpha (g_2 \circ g_1^{-1}) \circ g_1(m) + (1-\alpha)(g_2 \circ g_1^{-1}) \circ g_1(M)) \\ g_1^{-1}(g_1 \circ g_2^{-1}(\alpha g_2(m) + (1-\alpha) g_2(M))) \end{cases}$$
$$\leq g_2^{-1}(\alpha g_2(m) + (1-\alpha) g_2(M)),$$

where $m = \min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ and $M = \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$. $\triangle$

Proposition 3. *For any I-divergences $\Delta^{(1)}$ or $\Delta^{(2)}$ or any NI-divergences $\delta^{(1)}$ or $\delta^{(2)}$ with respective complexity terms $C^{(1)}$ and $C^{(2)}$, we have the following equivalence*:

$$\Delta^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} \leq \Delta^{(2)}_{\boldsymbol{X},\boldsymbol{Y}} \iff \delta^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} \leq \delta^{(2)}_{\boldsymbol{X},\boldsymbol{Y}} \iff C^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} \leq C^{(2)}_{\boldsymbol{X},\boldsymbol{Y}}. \tag{22}$$

Since for any $0 \leq \alpha \leq \alpha' \leq 1$ we have

$$C^{\alpha'}_{\boldsymbol{X},\boldsymbol{Y}} \leq C^{\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \leq C^{I}_{\boldsymbol{X},\boldsymbol{Y}}, \tag{23}$$

the associated I-divergences and NI-divergences are ordered according to equation (22). *Furthermore, a similar result holds for the main examples of this paper since*

$$C^{D,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \leq C^{P,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \leq C^{R,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \leq C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \leq C^{I}_{\boldsymbol{X},\boldsymbol{Y}} \leq C^{E}_{\boldsymbol{X},\boldsymbol{Y}}. \tag{24}$$

Proof. Equation (22) is direct. The left-hand side of (23) comes from

$$C^{\alpha}_{\boldsymbol{X},\boldsymbol{Y}} = g^{-1}(\alpha g(\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})) + (1-\alpha) g(\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})))$$
$$\leq g^{-1}(g(\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}))) = C^{I}_{\boldsymbol{X},\boldsymbol{Y}},$$

and the right-hand side is direct. Since $g^P \circ (g^D)^{-1}(\cdot) = -\log(\cdot)$, $g^R \circ (g^P)^{-1}(\cdot) = \exp\left(\frac{1}{2}\cdot\right)$, and $g^S \circ (g^R)^{-1}(\cdot) = (\cdot)^2$ are convex functions, (24) is a direct consequence of Lemma 1. $\triangle$

Remark 3. By assuming that either $g(\cdot)$ is a concave function or $g^{-1}(\cdot)$ is a convex function, we have the inequality

$$C^{\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \leq \alpha m + (1-\alpha) M = C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}},$$

which means that any Δ^{α} (respectively, δ^{α}) (satisfying the previous assumption) is upper bounded by $\Delta^{S,\alpha}$ (respectively, $\delta^{S,\alpha}$).

The following proposition gives a larger class of examples of I-divergences and NI-divergences.

Proposition 4. *Let $(\alpha^{(j)})_{j=1,\ldots,J}$ be some vector of probability weights for some $J \geq 1$.*

(i) *Let $\delta^{(1)}, \ldots, \delta^{(J)}$ be NI-divergences. Then the measure defined by*

$$\delta_{\boldsymbol{X},\boldsymbol{Y}} = \sum_{j=1}^{J} \alpha^{(j)} \delta^{(j)}_{\boldsymbol{X},\boldsymbol{Y}} \tag{25}$$

is an NI-divergence with complexity term given by

$$C_{\boldsymbol{X},\boldsymbol{Y}} = \left(\sum_{j=1}^{J} \frac{\alpha^{(j)}}{C^{(j)}_{\boldsymbol{X},\boldsymbol{Y}}} \right)^{-1}. \tag{26}$$

(ii) *Let $\Delta^{(1)}, \ldots, \Delta^{(j)}$ be I-divergences and $\delta^{(1)}, \ldots, \delta^{(j)}$ be NI-divergences with complexity terms $C^{(1)}_{\boldsymbol{X},\boldsymbol{Y}}, \ldots, C^{(J)}_{\boldsymbol{X},\boldsymbol{Y}}$. Then the measures defined by*

$$\Delta_{\boldsymbol{X},\boldsymbol{Y}} = C_{\boldsymbol{X},\boldsymbol{Y}} - I_{\boldsymbol{X},\boldsymbol{Y}} \qquad \text{and} \qquad \delta_{\boldsymbol{X},\boldsymbol{Y}} = 1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{C_{\boldsymbol{X},\boldsymbol{Y}}}, \quad \text{with} \quad C_{\boldsymbol{X},\boldsymbol{Y}} = \sum_{j=1}^{J} \alpha^{(j)} C^{(j)}_{\boldsymbol{X},\boldsymbol{Y}}, \tag{27}$$

are also, respectively, an I-divergence and NI-divergence.

The proof is immediate.

3.2. Around Property [**P3**]

The fact that an I-divergence Δ (respectively, NI-divergence δ) satisfies property [**P3**] may be expressed as follows: $\Delta_{\boldsymbol{X},\boldsymbol{Y}} = 0 \Leftrightarrow D^I_{\boldsymbol{X},\boldsymbol{Y}} = 0$ (respectively, $\delta_{\boldsymbol{X},\boldsymbol{Y}} = 0 \Leftrightarrow d^I_{\boldsymbol{X},\boldsymbol{Y}} = 0$). In fact, [**P3**] should be extended to a more useful assumption: $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ (or $\delta_{\boldsymbol{X},\boldsymbol{Y}}$) is near from the minimum 0 if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ share almost the same information. This may be translated by the following implications related to an I-divergence Δ (respectively, NI-divergence δ):

- For all $\gamma > 0$ there exists $\varepsilon > 0$ such that for all $(\boldsymbol{X}, \boldsymbol{Y}) \in \Upsilon$

$$\Delta_{\boldsymbol{X},\boldsymbol{Y}} \le \varepsilon \implies D^I_{\boldsymbol{X},\boldsymbol{Y}} \le \gamma \qquad (\text{respectively,} \quad \delta_{\boldsymbol{X},\boldsymbol{Y}} \le \varepsilon \implies d^I_{\boldsymbol{X},\boldsymbol{Y}} \le \gamma);$$

- For all $\varepsilon > 0$ there exists $\gamma > 0$ such that for all $(\boldsymbol{X}, \boldsymbol{Y}) \in \Upsilon$

$$D^I_{\boldsymbol{X},\boldsymbol{Y}} \le \gamma \implies \Delta_{\boldsymbol{X},\boldsymbol{Y}} \le \varepsilon \qquad (\text{respectively,} \quad d^I_{\boldsymbol{X},\boldsymbol{Y}} \le \gamma \implies \delta_{\boldsymbol{X},\boldsymbol{Y}} \le \varepsilon).$$

An I-divergence Δ (respectively, NI-divergence δ) inherits the previous property if it satisfies:

[P3′(Υ, k_1, k_2)] There exist some positive constants k_1 and k_2 ($k_1 \le k_2$) such that for all pairs $(\boldsymbol{X}, \boldsymbol{Y}) \in \Upsilon \subset \Gamma^2$ we have

$$k_1 D^I_{\boldsymbol{X},\boldsymbol{Y}} \le \Delta_{\boldsymbol{X},\boldsymbol{Y}} \le k_2 D^I_{\boldsymbol{X},\boldsymbol{Y}} \qquad (\text{respectively,} \quad k_1 d^I_{\boldsymbol{X},\boldsymbol{Y}} \le \delta_{\boldsymbol{X},\boldsymbol{Y}} \le k_2 d^I_{\boldsymbol{X},\boldsymbol{Y}}). \tag{28}$$

Among our examples, we assert that D^E and d^E both satisfy [**P3′(Γ², 1, 2)**], that is,

$$D^I_{\boldsymbol{X},\boldsymbol{Y}} \le D^E_{\boldsymbol{X},\boldsymbol{Y}} \le 2D^I_{\boldsymbol{X},\boldsymbol{Y}} \qquad (\text{respectively,} \quad d^I_{\boldsymbol{X},\boldsymbol{Y}} \le d^E_{\boldsymbol{X},\boldsymbol{Y}} \le 2d^I_{\boldsymbol{X},\boldsymbol{Y}}).$$

Most of complexity terms considered in this paper are of the particular form (17) where the function $g(\cdot)$ is a monotone function on $\mathbb{R}^+$. From (23), we can point out that for such complexity terms (expressed in terms of Δ or δ), the constant k_2 is equal to 1. Moreover, we assert that if Δ satisfies [**P3′(Υ, k_1, 1)**], then the associated δ also satisfies [**P3′(Υ, k_1, 1)**] since

$$k_1 d^I_{\boldsymbol{X},\boldsymbol{Y}} = \frac{k_1 D^I_{\boldsymbol{X},\boldsymbol{Y}}}{C^I_{\boldsymbol{X},\boldsymbol{Y}}} \le \frac{\Delta_{\boldsymbol{X},\boldsymbol{Y}}}{C_{\boldsymbol{X},\boldsymbol{Y}}} = \delta_{\boldsymbol{X},\boldsymbol{Y}}.$$

Therefore, the results presented hereafter in the rest of this section will be expressed for I-divergences only.

Furthermore, we now consider only complexity terms of the form (17) defined through a function $g(\cdot)$ continuously differentiable on some set $\mathcal{D}^g \subset \mathbb{R}^+$. Let us first introduce the following two subsets of $\mathcal{D}^g$:

$$\mathcal{E}_1^g = \{\Theta \subset \mathcal{D}^g : \ 0 < \varkappa^g_{\inf,\Theta} < \varkappa^g_{\sup,\Theta} < +\infty\} \qquad \text{and} \qquad \mathcal{E}_2^{g,\alpha} = \left\{\Theta \subset \mathcal{E}_1^g : \ \frac{\alpha \varkappa^g_{\sup,\Theta}}{\varkappa^g_{\inf,\Theta}} < 1\right\},$$

with $\varkappa^g_{\inf,\Theta} = \inf\limits_{x\in\Theta} |g'(x)|$ and $\varkappa^g_{\sup,\Theta} = \sup\limits_{x\in\Theta} |g'(x)|$. Denote also $\alpha^\wedge = \min(\alpha, 1-\alpha)$.

In the following, two results ensuring that an I-divergence Δ^α of the form (17) satisfies property [**P3′(Υ, k_1, k_2)**] are proposed. The difference relies upon the framework: the constants k_1 and k_2 differ whenever the set Υ differs.

Proposition 5. *For any $\Theta \in \mathcal{E}_1^g$, the I-divergence Δ^α satisfies* [**P3′(Υ_Θ, $\alpha^\wedge \frac{\varkappa^g_{\inf,\Theta}}{\varkappa^g_{\sup,\Theta}}$, 1)**] *with* $\Upsilon_\Theta = \{(\boldsymbol{X}, \boldsymbol{Y}) \in \Gamma^2 : \ H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}, I_{\boldsymbol{X},\boldsymbol{Y}} \in \Theta\}$.

Proof. Denote $x = \min(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}})$, $y = \max(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}})$, and $z = I_{\boldsymbol{X},\boldsymbol{Y}}$. There exist c_1, c_2, and c_3 such that

$$\begin{aligned} g^{-1}(\alpha g(x+z) + (1-\alpha)g(y+z)) - z & \\ &= (\alpha(g(x+z) - g(z)) + (1-\alpha)(g(y+z) - g(z)))(g^{-1})'(c_1) \\ &= \alpha|g'(c_2)||(g^{-1})'(c_1)|x + (1-\alpha)|g'(c_3)||(g^{-1})'(c_1)|y, \end{aligned}$$

with $c_1 \in [\min(g(z), \alpha g(x+z)+(1-\alpha)g(y+z)), \max(g(z), \alpha g(x+z)+(1-\alpha)g(y+z))]$, $c_2 \in [z, x+z]$, and $c_3 \in [z, y+z]$. Then for all x, y, and z we obtain

$$g^{-1}(\alpha g(x+z) + (1-\alpha)g(y+z)) - z \geq \alpha^{\wedge} \frac{\varkappa^g_{\mathrm{inf},\Theta}}{\varkappa^g_{\mathrm{sup},\Theta}} \max(x, y),$$

which means that $\alpha^{\wedge} \frac{\varkappa^g_{\mathrm{inf},\Theta}}{\varkappa^g_{\mathrm{sup},\Theta}} D^I_{\boldsymbol{X},\boldsymbol{Y}} \leq \Delta^{\alpha}_{\boldsymbol{X},\boldsymbol{Y}}$. $\triangle$

Proposition 6. *For any $\Theta \in \mathcal{E}_2^g$, the I-divergence Δ^{α} satisfies* $\mathbf{[P3'(\Gamma^2_\Theta, 1 - \alpha \frac{\varkappa^g_{\mathrm{sup},\Theta}}{\varkappa^g_{\mathrm{inf},\Theta}}, 1)]}$ *with* $\Gamma_\Theta = \{\boldsymbol{Z} \in \Gamma : H_{\boldsymbol{Z}} \in \Theta\}$.

Proof. We have

$$\begin{aligned} D^I_{\boldsymbol{X},\boldsymbol{Y}} - \Delta^{\alpha}_{\boldsymbol{X},\boldsymbol{Y}} = C^I_{\boldsymbol{X},\boldsymbol{Y}} - C_{\boldsymbol{X},\boldsymbol{Y}} &= \alpha(g^{-1})'(c_1)(g(\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})) - g(\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}))) \\ &= \alpha|(g^{-1})'(c_1)||g'(c_2)||H_{\boldsymbol{X}} - H_{\boldsymbol{Y}}|, \end{aligned}$$

with $c_1 \in [g(\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})), g(\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}))]$ and $c_2 \in [\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}), \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})]$. Then we obtain

$$D^I_{\boldsymbol{X},\boldsymbol{Y}} - \Delta^{\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \leq \alpha \frac{\varkappa^g_{\mathrm{sup},\Theta}}{\varkappa^g_{\mathrm{inf},\Theta}} D^I_{\boldsymbol{X},\boldsymbol{Y}},$$

which leads to the result. $\triangle$

For the sake of simplicity, we use the notation $\varkappa^{\bullet}_{\mathrm{inf},\Theta}$ and $\varkappa^{\bullet}_{\mathrm{sup},\Theta}$ instead of $\varkappa^{g^{\bullet}}_{\mathrm{inf},\Theta}$ and $\varkappa^{g^{\bullet}}_{\mathrm{sup},\Theta}$.

The following result is devoted to our different examples. We apply the previous two propositions and present a new result obtained by taking into account the specific form of each example.

Proposition 7. *The I-divergence $\Delta^{\bullet,\alpha}$ satisfies properties* $\mathbf{[P3'(\Upsilon_\Theta, k_1^{a,\bullet}, 1)]}$ (*from Proposition* 5), $\mathbf{[P3'(\Gamma^2_\Theta, k_1^{b,\bullet}, 1)]}$ (*from Proposition* 6), *and* $\mathbf{[P3'(\Gamma^2_\Theta, k_1^{c,\bullet}, 1)]}$, *where $\bullet$ stands for S, R, P, and D, and where*

$\bullet$	Θ	$\varkappa^{\bullet}_{\mathrm{inf},\Theta}$	$\varkappa^{\bullet}_{\mathrm{sup},\Theta}$	$k_1^{a,\bullet} = \alpha^{\wedge} \frac{\varkappa_{\mathrm{inf},\Theta}}{\varkappa^{\bullet}_{\mathrm{sup},\Theta}}$	$k_1^{b,\bullet} = 1 - \alpha \frac{\varkappa_{\mathrm{sup},\Theta}}{\varkappa^{\bullet}_{\mathrm{inf},\Theta}}$	$k_1^{c,\bullet}$
S	$\mathbb{R}^+$	1	1	$\alpha^{\wedge}$	$1-\alpha$	
R	$[c_1, c_2]$	$\frac{1}{2\sqrt{c_2}}$	$\frac{1}{2\sqrt{c_1}}$	$\frac{\alpha^{\wedge}}{\sqrt{\rho}}$	$1 - \alpha\sqrt{\rho}$ $(if\ \rho < \frac{1}{\alpha^2})$	$(1-\alpha)\left(1 - \frac{\alpha}{\left(1+\frac{1}{\sqrt{\rho}}\right)^2}\right)$
R	$\mathbb{R}^+$					$(1-\alpha)^2$
P	$[c_1, c_2]$	$\frac{1}{c_2}$	$\frac{1}{c_1}$	$\frac{\alpha^{\wedge}}{\rho}$	$1 - \alpha\rho$ $(if\ \rho < \frac{1}{\alpha})$	$\frac{\rho^{1-\alpha} - 1}{\rho - 1}$
D	$[c_1, c_2]$	$\frac{1}{c_2^2}$	$\frac{1}{c_1^2}$	$\frac{\alpha^{\wedge}}{\rho^2}$	$1 - \alpha\rho^2$ $(if\ \rho < \frac{1}{\sqrt{\alpha}})$	$\frac{1}{1 + \frac{\alpha}{1-\alpha}\rho}$

with $0 < c_1 \leq c_2 < +\infty$ *and* $\rho = \frac{c_2}{c_1}$.

Proof. The computation of $k_1^{a,\bullet}$ and $k_1^{b,\bullet}$ follows from Propositions 5 and 6. Hence, let us only concentrate on $k_1^{c,\bullet}$ for the complexity terms $C^{R,\alpha}$, $C^{P,\alpha}$, and $C^{D,\alpha}$. We denote $m = \min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ and $M = \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$.

- Complexity term $C^{R,\alpha}$:

$$\begin{aligned} D^I_{\boldsymbol{X},\boldsymbol{Y}} - \Delta^{R,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} &= \alpha(1-\alpha)(\sqrt{M}-\sqrt{m})^2 + \alpha(M-m) \\ &= \alpha(1-\alpha)\frac{(M-m)^2}{(\sqrt{M}+\sqrt{m})^2} + \alpha(M-m) \\ &\le \alpha(1-\alpha)\frac{(D^I_{\boldsymbol{X},\boldsymbol{Y}})^2}{(\sqrt{M}+\sqrt{m})^2} + \alpha D^I_{\boldsymbol{X},\boldsymbol{Y}}. \end{aligned}$$

  Thus,

$$\Delta^{R,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \ge (1-\alpha) D^I_{\boldsymbol{X},\boldsymbol{Y}} \left(1 - \alpha \frac{D^I_{\boldsymbol{X},\boldsymbol{Y}}}{(\sqrt{M}+\sqrt{m})^2}\right).$$

  The result is obtained by noting that

$$\frac{D^I_{\boldsymbol{X},\boldsymbol{Y}}}{(\sqrt{m}+\sqrt{M})^2} \le \frac{M}{(\sqrt{m}+\sqrt{M})^2} = \frac{1}{\left(1+\sqrt{\frac{m}{M}}\right)^2} \le \frac{1}{\left(1+\sqrt{\frac{c_1}{c_2}}\right)^2} \le 1.$$

- Complexity term $C^{P,\alpha}$: by using the Taylor expansion with integral rest, one obtains

$$\begin{aligned} D^I_{\boldsymbol{X},\boldsymbol{Y}} - \Delta^{P,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} &= M^{1-\alpha}(M^\alpha - m^\alpha) \\ &= M^{1-\alpha}(M-m)\int_0^1 \frac{\alpha}{(m+t(M-m))^{1-\alpha}}dt \\ &\le (M-m)\int_0^1 \frac{\alpha}{\left(\frac{1}{\rho}+t\left(1-\frac{1}{\rho}\right)\right)^{1-\alpha}}dt \\ &\le D^I_{\boldsymbol{X},\boldsymbol{Y}} \frac{1}{1-\frac{1}{\rho}} \left[\left(\frac{1}{\rho}+t\left(1-\frac{1}{\rho}\right)\right)^\alpha\right]_0^1 = D^I_{\boldsymbol{X},\boldsymbol{Y}} \frac{1-\left(\frac{1}{\rho}\right)^\alpha}{1-\frac{1}{\rho}}. \end{aligned}$$

  Thus,

$$\Delta^{P,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \ge D^I_{\boldsymbol{X},\boldsymbol{Y}} \left(1 - \frac{1-\left(\frac{1}{\rho}\right)^\alpha}{1-\frac{1}{\rho}}\right) = D^I_{\boldsymbol{X},\boldsymbol{Y}} \frac{\rho^{1-\alpha}-1}{\rho-1}.$$

- Complexity term $C^{D,\alpha}$:

$$D^I_{\boldsymbol{X},\boldsymbol{Y}} - \Delta^{D,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} = M - \frac{mM}{\alpha M + (1-\alpha)m} = \frac{\alpha M}{\alpha M + (1-\alpha)m}(M-m) \le \frac{1}{1+\frac{1-\alpha}{\alpha}\frac{c_1}{c_2}} D^I_{\boldsymbol{X},\boldsymbol{Y}}. \quad \triangle$$

3.3. Around the Triangle Inequality

The question arises now whether an I-divergence or NI-divergence satisfies property **[P6]**, i.e., the triangle inequality. The following proposition establishes sufficient conditions for such measures to constitute a metric.

Lemma 2. *We have*

$$H_{\boldsymbol{X},\boldsymbol{Y}} \leq H_{\boldsymbol{X},\boldsymbol{Z}} + H_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}}, \tag{29}$$

$$I_{\boldsymbol{X},\boldsymbol{Y}} \geq I_{\boldsymbol{X},\boldsymbol{Z}} + I_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}}. \tag{30}$$

Proof. From the general properties on entropy, one can obtain

$$H_{\boldsymbol{X},\boldsymbol{Y}} \leq H_{\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}} = H_{\boldsymbol{X},\boldsymbol{Z}} + H_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}} \leq H_{\boldsymbol{X},\boldsymbol{Z}} + H_{\boldsymbol{Y}|\boldsymbol{Z}} = H_{\boldsymbol{X},\boldsymbol{Z}} + H_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}}. \tag{31}$$

Equation (30) directly follows from (2). $\triangle$

Proposition 8. *Assume that the complexity term defining an I-divergence satisfies the following property*:

$$C_{\boldsymbol{X},\boldsymbol{Y}} \leq C_{\boldsymbol{X},\boldsymbol{Z}} + C_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}}. \tag{32}$$

Then the associated I-divergence satisfies the triangle inequality, that is,

$$\Delta_{\boldsymbol{X},\boldsymbol{Y}} \leq \Delta_{\boldsymbol{X},\boldsymbol{Z}} + \Delta_{\boldsymbol{Y},\boldsymbol{Z}}. \tag{33}$$

In addition, if C satisfies

$$C_{\boldsymbol{X},\boldsymbol{Z}} \geq \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Z}}), \tag{34}$$

then the associated NI-divergence also satisfies the triangle inequality, that is,

$$\delta_{\boldsymbol{X},\boldsymbol{Y}} \leq \delta_{\boldsymbol{X},\boldsymbol{Z}} + \delta_{\boldsymbol{Y},\boldsymbol{Z}}. \tag{35}$$

Proof. Since the quantity

$$A = -(C_{\boldsymbol{X},\boldsymbol{Y}} - I_{\boldsymbol{X},\boldsymbol{Y}}) + (C_{\boldsymbol{X},\boldsymbol{Z}} - I_{\boldsymbol{X},\boldsymbol{Z}}) + (C_{\boldsymbol{Y},\boldsymbol{Z}} - I_{\boldsymbol{Y},\boldsymbol{Z}}),$$

is nonnegative from (30) and (32), equation (33) immediately follows. Moreover, we have the following equation:

$$\delta_{\boldsymbol{X},\boldsymbol{Y}} \leq 1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{C_{\boldsymbol{X},\boldsymbol{Y}} + A}. \tag{36}$$

Now it is also easy to see from (34) that

$$A + C_{\boldsymbol{X},\boldsymbol{Y}} \geq C_{\boldsymbol{X},\boldsymbol{Z}} + C_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} \geq \max(C_{\boldsymbol{X},\boldsymbol{Z}}, C_{\boldsymbol{Y},\boldsymbol{Z}}).$$

From (36) it follows that

$$\delta_{\boldsymbol{X},\boldsymbol{Y}} \leq \frac{C_{\boldsymbol{X},\boldsymbol{Z}} - I_{\boldsymbol{X},\boldsymbol{Z}} + C_{\boldsymbol{Y},\boldsymbol{Z}} - I_{\boldsymbol{Y},\boldsymbol{Z}}}{\max(C_{\boldsymbol{X},\boldsymbol{Z}}, C_{\boldsymbol{Y},\boldsymbol{Z}})} \leq \frac{C_{\boldsymbol{X},\boldsymbol{Z}} - I_{\boldsymbol{X},\boldsymbol{Z}}}{C_{\boldsymbol{X},\boldsymbol{Z}}} + \frac{C_{\boldsymbol{Y},\boldsymbol{Z}} - I_{\boldsymbol{Y},\boldsymbol{Z}}}{C_{\boldsymbol{Y},\boldsymbol{Z}}} = \delta_{\boldsymbol{X},\boldsymbol{Z}} + \delta_{\boldsymbol{Y},\boldsymbol{Z}}. \quad \triangle$$

Remark 4. In Proposition 8 there is no implication between (32) and (34). Indeed, one may check that the NI-divergence δ^S (with $\alpha = 1/2$ for example) satisfies the first inequality but not the second one. Now consider an NI-divergence with complexity term $C_{\boldsymbol{X},\boldsymbol{Y}} = \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}) + H_{\boldsymbol{X}|\boldsymbol{Y}} H_{\boldsymbol{Y}|\boldsymbol{X}}$. By choosing $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ such that $H_{\boldsymbol{X}|\boldsymbol{Z}} = H_{\boldsymbol{Y}|\boldsymbol{Z}} = 0$ and $H_{\boldsymbol{X}|\boldsymbol{Y}} = H_{\boldsymbol{Y}|\boldsymbol{X}} = I_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{Z}}/3 = H_{\boldsymbol{X},\boldsymbol{Y}}/3 > 2$, one sees that (34) is satisfied but (32) is not.

Remark 5. Let us consider an NI-divergence δ with the complexity term

$$C_{\boldsymbol{X},\boldsymbol{Y}} = C'_{\boldsymbol{X},\boldsymbol{Y}} + \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$$

such that $C'_{\boldsymbol{X},\boldsymbol{Y}} \geq 0$ (necessarily $C'_{\boldsymbol{X},\boldsymbol{Y}} = 0$ whenever $\boldsymbol{X} \sim \boldsymbol{Y}$). Then, Δ and δ satisfy the triangle inequality if C' also satisfies the triangle inequality. However, this is not a necessary condition. Indeed, the triangle inequality is not satisfied for the same example of the previous remark with $C'_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X}|\boldsymbol{Y}} H_{\boldsymbol{Y}|\boldsymbol{X}}$, for which $C'_{\boldsymbol{X},\boldsymbol{Z}} = C'_{\boldsymbol{Y},\boldsymbol{Z}} = 0$, whereas $C'_{\boldsymbol{X},\boldsymbol{Y}} > 0$.

Let us now propose some examples and consequences through the following corollary.

Corollary 1. (i) *The measures D^E and D^I satisfy condition* (32) *and so are metrics.*

(ii) *The measures d^E and d^I satisfy conditions* (32) *and* (34) *and so are metrics.*

(iii) *The measure $D^{S,\alpha}$ for $\alpha \le \frac{1}{2}$ satisfies condition* (32) *and so is a metric. Moreover, when $\alpha > \frac{1}{2}$, this measure does not satisfy* (32).

(iv) *Let $(\alpha^{(j)})_{j=1,\dots,J}$ be some vector of probability weights for some $J \ge 1$. Let $\Delta^{(1)}, \dots, \Delta^{(J)}$ be I-divergences (respectively, $\delta^{(1)}, \dots, \delta^{(J)}$ be NI-divergences) with complexity terms $C^{(1)}_{\boldsymbol{X},\boldsymbol{Y}}, \dots, C^{(J)}_{\boldsymbol{X},\boldsymbol{Y}}$ satisfying* (32) (*respectively,* (32) *and* (34)). *Then the measures defined by* (27) *satisfy the triangle inequality.*

Proof. (i), (ii) Equation (29) corresponds exactly to (32) for $C^E_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X},\boldsymbol{Y}}$. Since $H_{\boldsymbol{X},\boldsymbol{Z}} \ge \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Z}})$, we have proved that D^E and d^E are metrics. Concerning D^I and d^I, the complexity term corresponds to $C^I_{\boldsymbol{X},\boldsymbol{Y}} = \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$. Thus, it suffices to prove (32), which is quite obvious. Indeed,

$$\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Z}}) + \max(H_{\boldsymbol{Y}}, H_{\boldsymbol{Z}}) - H_{\boldsymbol{Z}} \ge \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}).$$

(iii) Let $m = \min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ and $M = \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$. We distinguish three cases.

- $H_{\boldsymbol{Z}} < m$:
$$C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} = (2\alpha - 1)H_{\boldsymbol{Z}} + (1-\alpha)(m+M).$$
If $\alpha > \frac{1}{2}$ and $H_{\boldsymbol{X}} = H_{\boldsymbol{Y}}$, the right-hand side of the previous equation equals $(1-2\alpha) \times (C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} - H_{\boldsymbol{Z}}) + C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} < C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}$. Thus, (32) can never be satisfied for $\alpha > \frac{1}{2}$. Now, if $\alpha \le \frac{1}{2}$, we have
$$C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} > (1-\alpha)(m+M) \ge C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}.$$
- $H_{\boldsymbol{Z}} > M$:
$$C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} = (2\alpha - 1)H_{\boldsymbol{Z}} + (1-\alpha)(m+M) \ge \alpha m + (1-\alpha)M = C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}.$$
- $m \le H_{\boldsymbol{Z}} \le M$:
$$C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} = \alpha m + (1-\alpha)M = C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}.$$

(iv) This is trivial. △

We claim that the measures $\Delta^{R,\alpha}$, $\Delta^{P,\alpha}$, and $\Delta^{D,\alpha}$ (and so $\delta^{R,\alpha}$, $\delta^{P,\alpha}$, and $\delta^{D,\alpha}$) do not satisfy condition (32). Consider, for example, $\Delta^{D,\alpha}$. Let us choose $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ such that $H_{\boldsymbol{Z}} > \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ and $H_{\boldsymbol{Z}} = \frac{2-\alpha}{1-\alpha}H_{\boldsymbol{X}} = \frac{2-\alpha}{1-\alpha}H_{\boldsymbol{Y}}$. This leads to

$$C^{D,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + C^{D,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} = H_{\boldsymbol{Z}}\left(\frac{H_{\boldsymbol{X}}}{\alpha H_{\boldsymbol{X}} + (1-\alpha)H_{\boldsymbol{Z}}} + \frac{H_{\boldsymbol{Y}}}{\alpha H_{\boldsymbol{Y}} + (1-\alpha)H_{\boldsymbol{Z}}} - 1\right) = 0 < C^{D,\alpha}_{\boldsymbol{X},\boldsymbol{Y}},$$

which contradicts (32).

Concerning these divergences and the measures $\Delta^{S,\alpha}$ (for $\alpha > \frac{1}{2}$) and $\delta^{S,\alpha}$, we do not know if they satisfy the triangle inequality. We can only say that our tool cannot be applied to prove it. We suggest to weaken property **[P6]** in the following way in order to obtain more results. Let an *I*-divergence or *NI*-divergence satisfy

[P6′(Υ, c)] There exists $c \ge 1$ such that for all $(\boldsymbol{X},\boldsymbol{Y}), (\boldsymbol{Y},\boldsymbol{Z}), (\boldsymbol{X},\boldsymbol{Z}) \in \Upsilon$

$$\Delta_{\boldsymbol{X},\boldsymbol{Y}} \le c(\Delta_{\boldsymbol{X},\boldsymbol{Z}} + \Delta_{\boldsymbol{Y},\boldsymbol{Z}}).$$

Property [**P6**] is then equivalent to [$\mathbf{P6'(\Gamma^2, 1)}$], and we already know that D^E, d^E, D^I, d^I, and $D^{S,\alpha}$ (for $\alpha \leq \frac{1}{2}$) satisfy [$\mathbf{P6'(\Gamma^2, 1)}$]. When $\Upsilon \subsetneq \Gamma^2$, property [**P6′**] is in a sense local, whereas it is global (as the classical triangle inequality) when $\Upsilon = \Gamma^2$.

Note that if an I-divergence (or NI-divergence) satisfies [$\mathbf{P3'(\Upsilon, k_1, k_2)}$], then [$\mathbf{P6'(\Upsilon, \frac{k_2}{k_1})}$] is satisfied since

$$\Delta_{\boldsymbol{X},\boldsymbol{Y}} \leq k_2 D^I_{\boldsymbol{X},\boldsymbol{Y}} \leq k_2(D^I_{\boldsymbol{X},\boldsymbol{Z}} + D^I_{\boldsymbol{Y},\boldsymbol{Z}}) \leq \frac{k_2}{k_1}(\Delta_{\boldsymbol{X},\boldsymbol{Z}} + \Delta_{\boldsymbol{Y},\boldsymbol{Z}}).$$

We then inherit a lot of results from Proposition 7 related to our examples. In particular, $\Delta^{\bullet,\alpha}$ and $\delta^{\bullet,\alpha}$ (where $\bullet$ stands for S, R, P, and D) both satisfy [$\mathbf{P6'(\Upsilon_\Theta, \frac{1}{k_1^a})}$], [$\mathbf{P6'(\Gamma^2_\Theta, \frac{1}{k_1^b})}$], and [$\mathbf{P6'(\Gamma^2_\Theta, \frac{1}{k_1^c})}$].

In the rest of this section, we attempt to ensure the global property [$\mathbf{P6'(\Gamma^2, c)}$]. From Proposition 7 (with $\Theta = \mathbb{R}^+$) we see that the divergences $\Delta^{S,\alpha}$ (when $\alpha > \frac{1}{2}$) and $\delta^{S,\alpha}$ (respectively, $\Delta^{R,\alpha}$ and $\delta^{R,\alpha}$) satisfy [$\mathbf{P6'(\Gamma^2, \frac{1}{1-\alpha})}$] (respectively, [$\mathbf{P6'(\Gamma^2, \frac{1}{(1-\alpha)^2})}$]).

When $\alpha \leq \frac{1}{2}$, we could improve the previous result on $\Delta^{R,\alpha}$ by proving that it satisfies property [$\mathbf{P6'(\Gamma^2, \frac{1}{\alpha^2 + (1-\alpha)^2})}$], in the same spirit as in the proof leading to [**P3′**]. Indeed,

$$\begin{aligned} D^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} - \Delta^{R,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} &= \alpha(1-\alpha)(m + M - 2\sqrt{mM} \\ &\leq 2\alpha(1-\alpha)(D^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} + I_{\boldsymbol{X},\boldsymbol{Y}} - \sqrt{mM}) \\ &\leq 2\alpha(1-\alpha) D^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}, \end{aligned}$$

which leads to $\Delta^{R,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \geq (\alpha^2 + (1-\alpha)^2) D^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}$. Finally, let us note that

$$\Delta^{R,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \leq D^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \leq D^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + D^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} \leq \frac{1}{\alpha^2 + (1-\alpha)^2}(\Delta^{R,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + \Delta^{R,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}}).$$

Now we give a more general result, allowing us, in particular, to improve [$\mathbf{P6'(\Gamma^2, \frac{1}{1-\alpha})}$] for $\Delta^{S,\alpha}$ when $\alpha > \frac{1}{2}$.

Proposition 9. *Let us consider the following assumptions on a complexity term*: *there exists a constant* $c \geq 1$ *such that*

$$cC_{\boldsymbol{X},\boldsymbol{Z}} + cC_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} - (c-1)(I_{\boldsymbol{X},\boldsymbol{Z}} + I_{\boldsymbol{Y},\boldsymbol{Z}}) \geq C_{\boldsymbol{X},\boldsymbol{Y}}, \tag{37}$$

$$cC_{\boldsymbol{X},\boldsymbol{Z}} + cC_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} - (c-1)(I_{\boldsymbol{X},\boldsymbol{Z}} + I_{\boldsymbol{Y},\boldsymbol{Z}}) \geq \max(C_{\boldsymbol{X},\boldsymbol{Y}}, C_{\boldsymbol{X},\boldsymbol{Z}}, C_{\boldsymbol{Y},\boldsymbol{Z}}). \tag{38}$$

If an I*-divergence satisfies* (37) *or an* NI*-divergence satisfies* (38), *then they satisfy property* [$\mathbf{P6'(\Gamma^2, c)}$].

Proof. Let us introduce

$$A = -(C_{\boldsymbol{X},\boldsymbol{Y}} - I_{\boldsymbol{X},\boldsymbol{Y}}) + c(C_{\boldsymbol{X},\boldsymbol{Z}} - I_{\boldsymbol{X},\boldsymbol{Z}}) + c(C_{\boldsymbol{Y},\boldsymbol{Z}} - I_{\boldsymbol{Y},\boldsymbol{Z}}).$$

From (30) and (37), one may assert that

$$A \geq cC_{\boldsymbol{X},\boldsymbol{Z}} + cC_{\boldsymbol{Y},\boldsymbol{Z}} - C_{\boldsymbol{X},\boldsymbol{Y}} - H_{\boldsymbol{Z}} - (c-1)(I_{\boldsymbol{X},\boldsymbol{Z}} + I_{\boldsymbol{Y},\boldsymbol{Z}}) \geq 0,$$

which implies that the result is valid for Δ. Now from (38) one may write

$$A + C_{\boldsymbol{X},\boldsymbol{Y}} \geq \max(C_{\boldsymbol{X},\boldsymbol{Y}}, C_{\boldsymbol{X},\boldsymbol{Z}}, C_{\boldsymbol{Y},\boldsymbol{Z}}) \geq \max(C_{\boldsymbol{X},\boldsymbol{Z}}, C_{\boldsymbol{Y},\boldsymbol{Z}}),$$

which leads to

$$\delta_{\boldsymbol{X},\boldsymbol{Y}} \le \frac{c(C_{\boldsymbol{X},\boldsymbol{Z}} - I_{\boldsymbol{X},\boldsymbol{Z}}) + c(C_{\boldsymbol{Y},\boldsymbol{Z}} - I_{\boldsymbol{Y},\boldsymbol{Z}})}{\max(C_{\boldsymbol{X},\boldsymbol{Z}}, C_{\boldsymbol{Y},\boldsymbol{Z}})} \le c\delta_{\boldsymbol{X},\boldsymbol{Z}} + c\delta_{\boldsymbol{Y},\boldsymbol{Z}}. \quad \triangle$$

Corollary 2. *The measures* $\Delta^{S,\alpha}$ *for* $\alpha > \frac{1}{2}$ *satisfy* $[\mathbf{P6'(\Gamma^2, \frac{\alpha}{1-\alpha})}]$.

Proof. Let us concentrate on $\Delta^{S,\alpha}$ for $\alpha > \frac{1}{2}$. Let

$$A = cC^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + cC^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} - (c-1)(I_{\boldsymbol{X},\boldsymbol{Z}} + I_{\boldsymbol{Y},\boldsymbol{Z}}).$$

Without loss of generality, we assume that $H_{\boldsymbol{X}} \le H_{\boldsymbol{Y}}$. We distinguish three cases.

- $H_{\boldsymbol{Z}} \le H_{\boldsymbol{X}} \le H_{\boldsymbol{Y}}$: since $C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} \ge I_{\boldsymbol{Y},\boldsymbol{Z}}$, we have

  $$A \ge c(1-\alpha)H_{\boldsymbol{X}} + (1-\alpha)H_{\boldsymbol{Y}} + (c\alpha + \alpha - 1)H_{\boldsymbol{Z}} - (c-1)I_{\boldsymbol{X},\boldsymbol{Z}}.$$

  Then

  $$A - C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \ge (c(1-\alpha) - \alpha)H_{\boldsymbol{X}} + (c\alpha + \alpha - 1)H_{\boldsymbol{Z}} - (c-1)I_{\boldsymbol{X},\boldsymbol{Z}} \ge (c-1)(H_{\boldsymbol{Z}} - I_{\boldsymbol{X},\boldsymbol{Z}}) \ge 0$$

  as soon as $c \ge \frac{\alpha}{1-\alpha}$.
- $H_{\boldsymbol{X}} \le H_{\boldsymbol{Y}} \le H_{\boldsymbol{Z}}$: we have

  $$A \ge \alpha H_{\boldsymbol{X}} + c\alpha H_{\boldsymbol{Y}} + ((1-\alpha) + c(1-\alpha) - 1)H_{\boldsymbol{Z}} - (c-1)I_{\boldsymbol{Y},\boldsymbol{Z}}.$$

  Then

  $$A - C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \ge (c\alpha - (1-\alpha))H_{\boldsymbol{Y}} + ((1-\alpha) + c(1-\alpha) - 1)H_{\boldsymbol{Z}} - (c-1)I_{\boldsymbol{Y},\boldsymbol{Z}} \ge (c-1)(H_{\boldsymbol{Y}} - I_{\boldsymbol{Y},\boldsymbol{Z}}) \ge 0$$

  as soon as $c \ge \frac{\alpha}{1-\alpha}$.
- $H_{\boldsymbol{X}} < H_{\boldsymbol{Z}} < H_{\boldsymbol{Y}}$: we have

  $$A \ge c\alpha H_{\boldsymbol{X}} + (1-\alpha)H_{\boldsymbol{Y}} + (c(1-\alpha)H_{\boldsymbol{Z}} + \alpha - 1) - (c-1)I_{\boldsymbol{X},\boldsymbol{Z}}.$$

  Then

  $$A - C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \ge (c-1)\alpha H_{\boldsymbol{X}} + (c-1)(1-\alpha)H_{\boldsymbol{Z}} - I_{\boldsymbol{X},\boldsymbol{Z}} \ge 0.$$

Hence, we obtain $A - C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} \ge 0$ for $c = \frac{\alpha}{1-\alpha}$. $\triangle$

Remark 6. The tool presented in Proposition 9 cannot be applied to the I-divergence $\Delta^{D,\alpha}$ and NI-divergence $\delta^{D,\alpha}$. Indeed, let us be given some $c \ge 1$ and consider the quantity

$$A = cC^{D,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + cC^{D,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} - (c-1)(I_{\boldsymbol{X},\boldsymbol{Z}} + I_{\boldsymbol{Y},\boldsymbol{Z}}).$$

In fact, one can always find $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ such that for all $c \ge 1$ the quantity A is negative. Indeed, let us choose $\boldsymbol{Z}$ independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$ and such that $\alpha H_{\boldsymbol{Z}} + (1-\alpha)H_{\boldsymbol{X}} = 3cH_{\boldsymbol{X}}$ and $\alpha H_{\boldsymbol{Z}} + (1-\alpha)H_{\boldsymbol{Y}} = 3cH_{\boldsymbol{Y}}$. It is easy to see that $H_{\boldsymbol{Z}} \ge \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ and therefore $A = H_{\boldsymbol{Z}}\left(\frac{1}{3} + \frac{1}{3} - 1\right) < 0$. In the same manner, the tool is inapplicable to the I-divergence $\Delta^{P,\alpha}$ and NI-divergence $\delta^{P,\alpha}$. Indeed, let us take $\boldsymbol{Z}$ independent of $\boldsymbol{X}$ and $\boldsymbol{Y}$ and such that $H_{\boldsymbol{X}} = H_{\boldsymbol{Y}} = \left(\frac{1}{3c}\right)^{1/\alpha} H_{\boldsymbol{Z}}$; then $H_{\boldsymbol{Z}} \ge \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ and

$$A = cC^{P,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + cC^{P,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} - (c-1)(I_{\boldsymbol{X},\boldsymbol{Z}} + I_{\boldsymbol{Y},\boldsymbol{Z}}) = -\frac{1}{3}H_{\boldsymbol{Z}} < 0.$$

The following result is an extension of Proposition 9, well suited to be applied to $\delta^{D,\alpha}$.

Proposition 10. *Let us assume that there exist two positive integers, I and Jsuch that an NI-divergence $\delta_{\boldsymbol{X},\boldsymbol{Y}}$ can be expressed as*

$$\delta_{\boldsymbol{X},\boldsymbol{Y}} = \sum_{i=1}^{I} \frac{S^{(i)}_{\boldsymbol{X},\boldsymbol{Y}}}{U^{(i)}_{\boldsymbol{X},\boldsymbol{Y}}} = \sum_{j=1}^{J} \alpha^{(j)} \left(1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{C^{(j)}_{\boldsymbol{X},\boldsymbol{Y}}}\right),$$

where $(\alpha^{(j)})_{j=1,\dots,J}$ is some vector of probability weights. Denote $S_{\boldsymbol{X},\boldsymbol{Y}} = \sum\limits_{i=1}^{I} S^{(i)}_{\boldsymbol{X},\boldsymbol{Y}}$ and $U_{\boldsymbol{X},\boldsymbol{Y}} = \max\limits_{i=1,\dots,I} U^{(i)}_{\boldsymbol{X},\boldsymbol{Y}}$. Then, if there exists some real number $c \geq 1$ such that for any $j = 1, \dots, J$ we have

(i) $A^{(j)} = I_{\boldsymbol{X},\boldsymbol{Y}} - C^{(j)}_{\boldsymbol{X},\boldsymbol{Y}} + c(S_{\boldsymbol{X},\boldsymbol{Z}} + S_{\boldsymbol{Z},\boldsymbol{Y}}) \geq 0$,

(ii) $A^{(j)} + C^{(j)}_{\boldsymbol{X},\boldsymbol{Y}} \geq \max(U_{\boldsymbol{X},\boldsymbol{Z}}, U_{\boldsymbol{Z},\boldsymbol{Y}})$,

then δ satisfies $[\mathbf{P6'(\Gamma^2, c)}]$.

Proof. Using assumptions (i) and (ii), one can prove that for all $j = 1, \dots, J$ we have

$$1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{C^{(j)}_{\boldsymbol{X},\boldsymbol{Y}}} \leq 1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{C^{(j)}_{\boldsymbol{X},\boldsymbol{Y}} + A^{(j)}} \leq c \frac{S_{\boldsymbol{X},\boldsymbol{Z}} + S_{\boldsymbol{Y},\boldsymbol{Z}}}{\max(U_{\boldsymbol{X},\boldsymbol{Z}}, U_{\boldsymbol{Z},\boldsymbol{Y}})} \leq c(\delta_{\boldsymbol{X},\boldsymbol{Z}} + \delta_{\boldsymbol{Y},\boldsymbol{Z}}).$$

It follows that

$$\delta_{\boldsymbol{X},\boldsymbol{Y}} = \sum_{j=1}^{J} \alpha^{(j)} \left(1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{C^{(j)}_{\boldsymbol{X},\boldsymbol{Y}}}\right) \leq \sum_{j=1}^{J} \alpha^{(j)} c(\delta_{\boldsymbol{X},\boldsymbol{Z}} + \delta_{\boldsymbol{Y},\boldsymbol{Z}}) = c(\delta_{\boldsymbol{X},\boldsymbol{Z}} + \delta_{\boldsymbol{Y},\boldsymbol{Z}}). \quad \triangle$$

Corollary 3. *The measure $\delta^{D,\alpha}$ satisfies* $[\mathbf{P6'(\Gamma^2, \frac{1}{\alpha^\wedge})}]$.

Proof. We have

$$\begin{aligned}
\delta^{D,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} &= \alpha \min\left(\frac{H_{\boldsymbol{X}|\boldsymbol{Y}}}{H_{\boldsymbol{X}}}, \frac{H_{\boldsymbol{Y}|\boldsymbol{X}}}{H_{\boldsymbol{Y}}}\right) + (1-\alpha) \max\left(\frac{H_{\boldsymbol{X}|\boldsymbol{Y}}}{H_{\boldsymbol{X}}}, \frac{H_{\boldsymbol{Y}|\boldsymbol{X}}}{H_{\boldsymbol{Y}}}\right) \\
&= \alpha \frac{\min(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}})}{\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})} + (1-\alpha) \frac{\max(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}})}{\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})} \\
&= \alpha \left(1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})}\right) + (1-\alpha) \left(1 - \frac{I_{\boldsymbol{X},\boldsymbol{Y}}}{\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})}\right).
\end{aligned}$$

Using the notation introduced in Proposition 10, we have $I = J = 2$, $S^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} = \alpha \min(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}})$, $S^{(2)}_{\boldsymbol{X},\boldsymbol{Y}} = (1-\alpha)\max(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}})$, $U^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} = \min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$, $U^{(2)}_{\boldsymbol{X},\boldsymbol{Y}} = \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$, $C^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} = \min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$, and $C^{(2)}_{\boldsymbol{X},\boldsymbol{Y}} = \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$. Let us fix c to the value $\frac{1}{\alpha^\wedge}$. We have

$$\begin{aligned}
A^{(1)} = I_{\boldsymbol{X},\boldsymbol{Y}} - \min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}) + \frac{1}{\alpha^\wedge} \Big(& \alpha \min(H_{\boldsymbol{X}|\boldsymbol{Z}}, H_{\boldsymbol{Z}|\boldsymbol{X}}) + (1-\alpha)\max(H_{\boldsymbol{X}|\boldsymbol{Z}}, H_{\boldsymbol{Z}|\boldsymbol{X}}) \\
& + \alpha \min(H_{\boldsymbol{Y}|\boldsymbol{Z}}, H_{\boldsymbol{Z}|\boldsymbol{Y}}) + (1-\alpha)\max(H_{\boldsymbol{Y}|\boldsymbol{Z}}, H_{\boldsymbol{Z}|\boldsymbol{Y}}) \Big).
\end{aligned}$$

Clearly, from (29) we have

$$\begin{aligned}
A^{(1)} &\geq \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}) - H_{\boldsymbol{X},\boldsymbol{Y}} + 2H_{\boldsymbol{X},\boldsymbol{Z}} + 2H_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{X}} - H_{\boldsymbol{Y}} - 2H_{\boldsymbol{Z}} \\
&\geq H_{\boldsymbol{X},\boldsymbol{Z}} + H_{\boldsymbol{Y},\boldsymbol{Z}} - \min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}) - H_{\boldsymbol{Z}} \geq 0.
\end{aligned}$$

Furthermore,

$$\min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}) + A^{(1)} \geq H_{\boldsymbol{X},\boldsymbol{Z}} + H_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} \geq \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}, H_{\boldsymbol{Z}}) = \max(U_{\boldsymbol{X},\boldsymbol{Z}}, U_{\boldsymbol{Y},\boldsymbol{Z}}).$$

It follows that $A^{(1)}$ satisfies conditions (i) and (ii) of Proposition 10 with $c = \frac{1}{\alpha^\wedge}$. The proof is strictly similar for $A^{(2)}$. $\triangle$

4. PREDICTION FRAMEWORK

Let us consider properties related to the prediction of some fixed random vector $\boldsymbol{Y}$.

4.1. Prediction Framework

Recall that our purpose is to find a random vector $\boldsymbol{X}$ that minimizes $\Delta_{\boldsymbol{Y},\boldsymbol{X}}$ (respectively, $\delta_{\boldsymbol{Y},\boldsymbol{X}}$), which combines a complexity term $C_{\boldsymbol{X},\boldsymbol{Y}}$ (to minimize) and an information term $I_{\boldsymbol{X},\boldsymbol{Y}}$ (to maximize). Let us assume that we have already got some $\boldsymbol{X}_1$ and its associated measure $\Delta_{\boldsymbol{Y},\boldsymbol{X}_1}$ (respectively, $\delta_{\boldsymbol{Y},\boldsymbol{X}_1}$). After evaluating $\Delta_{\boldsymbol{Y},\boldsymbol{X}_2}$ (respectively, $\delta_{\boldsymbol{Y},\boldsymbol{X}_2}$), we may be interested in describing conditions under which $\boldsymbol{X}_2$ is better (or worse) than $\boldsymbol{X}_1$.

Proposition 11. *The following cases may occur.*

Case 1. We choose $\boldsymbol{X}_2$ instead of $\boldsymbol{X}_1$ when

$$\begin{aligned} \Delta_{\boldsymbol{Y},\boldsymbol{X}_2} < \Delta_{\boldsymbol{Y},\boldsymbol{X}_1} \quad &\Longleftrightarrow \quad C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} < I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}, \\ \delta_{\boldsymbol{Y},\boldsymbol{X}_2} < \delta_{\boldsymbol{Y},\boldsymbol{X}_1} \quad &\Longleftrightarrow \quad \frac{C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1}}{C_{\boldsymbol{Y},\boldsymbol{X}_1}} < \frac{I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}}{I_{\boldsymbol{Y},\boldsymbol{X}_1}}. \end{aligned} \tag{39}$$

Case 2. We keep $\boldsymbol{X}_1$ and reject $\boldsymbol{X}_2$ when

$$\begin{aligned} \Delta_{\boldsymbol{Y},\boldsymbol{X}_2} \geq \Delta_{\boldsymbol{Y},\boldsymbol{X}_1} \quad &\Longleftrightarrow \quad C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} \geq I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}, \\ \delta_{\boldsymbol{Y},\boldsymbol{X}_2} \geq \delta_{\boldsymbol{Y},\boldsymbol{X}_1} \quad &\Longleftrightarrow \quad \frac{C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1}}{C_{\boldsymbol{Y},\boldsymbol{X}_1}} \geq \frac{I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}}{I_{\boldsymbol{Y},\boldsymbol{X}_1}}. \end{aligned}$$

This result implies automatically that properties **[P8]** and **[P9]** are satisfied. Let us comment more precisely on the previous proposition.

- Case 1 holds when
  1. $\boldsymbol{X}_2$ is simpler than $\boldsymbol{X}_1$ (i.e., $C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} < 0$) and $\boldsymbol{X}_2$ is at least as informative as $\boldsymbol{X}_1$ (i.e., $I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1} \geq 0$).
  2. $\boldsymbol{X}_2$ and $\boldsymbol{X}_1$ have the same complexity (i.e., $C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} = 0$) and $\boldsymbol{X}_2$ is more informative than $\boldsymbol{X}_1$ (i.e., $I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1} > 0$).
  3. $\boldsymbol{X}_2$ is simpler and less informative than $\boldsymbol{X}_1$ and such that the absolute (respectively, relative) excess of complexity is less than the absolute (respectively, relative) gain of information, i.e., $C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} < I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1} < 0$ (respectively, $\frac{C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1}}{C_{\boldsymbol{Y},\boldsymbol{X}_1}} < \frac{I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}}{I_{\boldsymbol{Y},\boldsymbol{X}_1}} < 0$).
  4. $\boldsymbol{X}_2$ is more complex and more informative than $\boldsymbol{X}_1$ and such that the absolute (respectively, relative) excess of complexity is less than the absolute (respectively, relative) gain of information, i.e., $0 < C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} < I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}$ (respectively, $0 < \frac{C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1}}{C_{\boldsymbol{Y},\boldsymbol{X}_1}} < \frac{I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}}{I_{\boldsymbol{Y},\boldsymbol{X}_1}}$).
- Case 2 holds when
  1. $\boldsymbol{X}_2$ is at least as complex as $\boldsymbol{X}_1$ (i.e., $C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} \geq 0$) and $\boldsymbol{X}_2$ is at most as informative as $\boldsymbol{X}_1$ (i.e., $I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1} \leq 0$).

2. $\boldsymbol{X}_2$ is simpler and less informative than $\boldsymbol{X}_1$ and such that the absolute (respectively, relative) excess of complexity is greater than or equal to the absolute (respectively, relative) gain of information, i.e., $0 > C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} \geq I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}$ (respectively, $0 > \frac{C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1}}{C_{\boldsymbol{Y},\boldsymbol{X}_1}} \geq \frac{I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}}{I_{\boldsymbol{Y},\boldsymbol{X}_1}}$).
3. $\boldsymbol{X}_2$ is more complex and more informative than $\boldsymbol{X}_1$ and such that the absolute (respectively, relative) excess of complexity is greater than or equal to the absolute (respectively, relative) gain of information, i.e., $C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} \geq I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1} > 0$ (respectively, $\frac{C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1}}{C_{\boldsymbol{Y},\boldsymbol{X}_1}} \geq \frac{I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}}{I_{\boldsymbol{Y},\boldsymbol{X}_1}} > 0$).

Proposition 12. *Any complexity term C^α of the form* (17) *with $\alpha \in]0,1[$ satisfies* **[P10]**.

Proof. Without loss of generality, the function $g(\cdot)$ that defines C^α is assumed to be increasing. It is clear that $H_{\boldsymbol{X}_2} \geq H_{\boldsymbol{X}_1} \Rightarrow C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_2} \geq C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_1}$ and respectively (by symmetry) $H_{\boldsymbol{X}_1} \geq H_{\boldsymbol{X}_2} \Rightarrow C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_1} \geq C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_2}$, which is equivalent to $C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_2} > C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_1} \Rightarrow H_{\boldsymbol{X}_2} > H_{\boldsymbol{X}_1}$. Thus, it remains to check that $C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_2} = C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_1} \Rightarrow H_{\boldsymbol{X}_2} = H_{\boldsymbol{X}_1}$:

$$\begin{aligned} C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_2} - C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_1} = 0 \quad &\Longleftrightarrow \quad \alpha(g(m_1) - g(m_2)) + (1-\alpha)(g(M_1) - g(M_2)) = 0 \\ &\Longleftrightarrow \quad g(m_1) = g(m_2) \quad \text{and} \quad g(M_1) = g(M_2) \\ &\Longleftrightarrow \quad m_1 = m_2 \quad \text{and} \quad M_1 = M_2 \\ &\Longleftrightarrow \quad H_{\boldsymbol{X}_2} = H_{\boldsymbol{X}_1}. \end{aligned} \tag{40}$$

Equation (40) is obtained by observing that $g(m_1) - g(m_2)$ and $g(M_1) - g(M_2)$ are of the same sign. $\triangle$

Remark 7. The complexity terms C^E and C^I (which corresponds to the case of any C^α with $\alpha = 0$) do not satisfy property **[P10]** in the general case. Indeed, there is no implication for C^E, and one can only prove that $H_{\boldsymbol{X}_1} \geq H_{\boldsymbol{X}_2} \Rightarrow C^I_{\boldsymbol{Y},\boldsymbol{X}_1} \geq C^I_{\boldsymbol{Y},\boldsymbol{X}_2}$. However, we point out that when $I_{\boldsymbol{Y},\boldsymbol{X}_1} = I_{\boldsymbol{Y},\boldsymbol{X}_2}$, then both C^E and C^I satisfy **[P10]**.

More specifically, two frameworks may be of special interest.

- $\boldsymbol{X}_2$ is as informative as $\boldsymbol{X}_1$ (i.e., $I_{\boldsymbol{Y},\boldsymbol{X}_1} = I_{\boldsymbol{Y},\boldsymbol{X}_2}$): we expect to select the random variable with the smallest entropy. This is effectively what happens when **[P10]** is satisfied; from Proposition 12 and Remark 7 (in this framework), this is the case for $C^\bullet$ with $\bullet = I, S, R, P, D$ in the general case and for C^E in this framework, since $H_{\boldsymbol{Y},\boldsymbol{X}_2} - H_{\boldsymbol{Y},\boldsymbol{X}_1} = H_{\boldsymbol{X}_2} - H_{\boldsymbol{X}_1}$.
- $\boldsymbol{X}_1 = g(\boldsymbol{X}_2)$ with g some surjective (but not injective) mapping: $\boldsymbol{X}_2$ is more complex than $\boldsymbol{X}_1$, and $\boldsymbol{X}_2$ is at least as informative as $\boldsymbol{X}_1$. Consequently, this case is nontrivial since both the absolute (respectively, relative) excess of complexity and absolute (respectively, relative) gain of information are competing. Let us give two important examples of such a context.
  1. Quantization problem: given a quantized version $\boldsymbol{X}_1$ of some (continuous) random variable with its associated partition $\mathcal{A}_1$, the problem is to know whether some new quantized version $\boldsymbol{X}_2$ with an associated partition $\mathcal{A}_2$ finer than $\mathcal{A}_1$ should be preferred to predict $\boldsymbol{Y}$.
  2. Variable selection problem: assume that one wants to construct an ascending selection method. The vector $\boldsymbol{X}_1$ could represent some selected set of covariables and $\boldsymbol{X}_2 = (\boldsymbol{X}_1, \boldsymbol{X}'_2)$ a larger set of covariables. The aim is therefore to know if $\boldsymbol{X}'_2$ should be integrated to the selected set or not.

Some simple algorithms of quantization and selection methods are proposed in [15] using these results.

4.2. Around the Redundancy of Two Random Vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$

In the future use of an I-divergence or NI-divergence, one would expect that if two discrete-valued random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the same (or almost the same) information with respect to an I-divergence or NI-divergence, then both have the same effect on the prediction of another vector $\boldsymbol{Y}$. This requirement, expressed by property **[P11]**, could be used for example in a variable selection problem in the context of discrimination to detect redundant variables.

In order to make property **[P11]** applicable for practical purpose, we may find interesting to have a bound on the difference $|\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}|$ (respectively, $|\delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \delta_{\boldsymbol{Y},\boldsymbol{X}_2}|$) expressed in terms of $D^I_{\boldsymbol{X}_1,\boldsymbol{X}_2}$ (respectively, $d^I_{\boldsymbol{X}_1,\boldsymbol{X}_2}$). More precisely, the question may arise whether there exists a function $h(\cdot)$ satisfying $h(x) \to 0$ as $x \to 0$ and such that $|\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}| \le h(D^I_{\boldsymbol{X}_1,\boldsymbol{X}_2})$ (respectively, $|\delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \delta_{\boldsymbol{Y},\boldsymbol{X}_2}| \le h(d^I_{\boldsymbol{X}_1,\boldsymbol{X}_2})$). Here, according to our examples, we only concentrate on linear functions $h(\cdot)$.

We then propose to translate property **[P11]** to an I-divergence Δ (respectively, NI-divergence δ) as follows:

[P11′($\boldsymbol{\Upsilon}, \boldsymbol{k}$)] There exists some positive constant k such that for all $(\boldsymbol{X}_1, \boldsymbol{X}_2) \in \Upsilon \subset \Gamma^2$ we have

$$|\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}| \le k D^I_{\boldsymbol{X}_1,\boldsymbol{X}_2}. \tag{41}$$

As a first answer, let us note that if the I-divergence (respectively, NI-divergence) satisfies the triangle inequality **[P6′($\boldsymbol{\Gamma^2}$, 1)]** or **[P3′($\boldsymbol{\Upsilon}, \boldsymbol{k_1}, \boldsymbol{k_2}$)]**, then it satisfies **[P11′($\boldsymbol{\Upsilon}, \boldsymbol{k_2}$)]** due to the following equivalent expression of the triangle inequality:

$$|D_{\boldsymbol{Y},\boldsymbol{X}_1} - D_{\boldsymbol{Y},\boldsymbol{X}_2}| \le D_{\boldsymbol{X}_1,\boldsymbol{X}_2} \qquad (\text{respectively,} \quad |d_{\boldsymbol{Y},\boldsymbol{X}_1} - d_{\boldsymbol{Y},\boldsymbol{X}_2}| \le d_{\boldsymbol{X}_1,\boldsymbol{X}_2}).$$

A priori, if an I-divergence or NI-divergence only satisfies **[P6′($\boldsymbol{\Gamma^2}, \boldsymbol{c}$)]** with some $c > 1$, then this property does no more seem to be true: indeed, for all $\boldsymbol{Y}$, $\boldsymbol{X}_1$, and $\boldsymbol{X}_2$, one may prove for an I-divergence, for instance, that

$$|\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}| \le c\Delta_{\boldsymbol{X}_1,\boldsymbol{X}_2} + (c-1)\min(\Delta_{\boldsymbol{Y},\boldsymbol{X}_1}, \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}) \not\le c\Delta_{\boldsymbol{X}_1,\boldsymbol{X}_2}.$$

Actually, this apparently disappointing result only expresses that the "redundancy" property cannot (always) be derived from a triangle-type inequality.

The following proposition gives some sufficient conditions on a complexity term ensuring that the associated Δ and δ satisfy property **[P11′]**.

Proposition 13. (i) *Assume that there exists some positive constant $\varkappa_1$ such that the complexity term of an I-divergence for all $(\boldsymbol{X}_1, \boldsymbol{X}_2) \in \Upsilon$ satisfies*

$$|C_{\boldsymbol{Y},\boldsymbol{X}_1} - C_{\boldsymbol{Y},\boldsymbol{X}_2}| \le \varkappa_1 |H_{\boldsymbol{X}_1} - H_{\boldsymbol{X}_2}|. \tag{42}$$

Then Δ satisfies **[P11′($\boldsymbol{\Upsilon}, \boldsymbol{1 + \varkappa_1}$)]**.

(ii) *If in addition there exists some positive constant $\varkappa_2$ such that for all $(\boldsymbol{X}_1, \boldsymbol{X}_2) \in \Upsilon$*

$$\max(C_{\boldsymbol{Y},\boldsymbol{X}_1}, C_{\boldsymbol{Y},\boldsymbol{X}_2}) \ge \varkappa_2 C^I_{\boldsymbol{X}_1,\boldsymbol{X}_2}, \tag{43}$$

then the associated NI-divergence satisfies **[P11′($\boldsymbol{\Upsilon}, \dfrac{\boldsymbol{1 + \varkappa_1}}{\boldsymbol{\varkappa_2}}$)]**.

Proof. (i) Let us first write

$$|\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}| \le |I_{\boldsymbol{Y},\boldsymbol{X}_1} - I_{\boldsymbol{Y},\boldsymbol{X}_2}| + |C_{\boldsymbol{Y},\boldsymbol{X}_1} - C_{\boldsymbol{Y},\boldsymbol{X}_2}|. \tag{44}$$

Now note that

$$I_{\boldsymbol{Y},\boldsymbol{X}_1} \geq I_{\boldsymbol{Y},\boldsymbol{X}_2} + I_{\boldsymbol{X}_1,\boldsymbol{X}_2} - H_{\boldsymbol{X}_2},$$

from which one can deduce

$$|I_{\boldsymbol{Y},\boldsymbol{X}_1} - I_{\boldsymbol{Y},\boldsymbol{X}_2}| \leq \max(H_{\boldsymbol{X}_1}, H_{\boldsymbol{X}_2}) - I_{\boldsymbol{X}_1,\boldsymbol{X}_2} = \max(H_{\boldsymbol{X}_1|\boldsymbol{X}_2}, H_{\boldsymbol{X}_2|\boldsymbol{X}_1}) = D^I_{\boldsymbol{X}_1,\boldsymbol{X}_2}. \tag{45}$$

The result is then obtained by combining (42), (44), and (45).

(ii) We can obtain the following result:

$$\begin{aligned}|\delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \delta_{\boldsymbol{Y},\boldsymbol{X}_2}| &\leq \frac{\min(C_{\boldsymbol{Y},\boldsymbol{X}_1}, C_{\boldsymbol{Y},\boldsymbol{X}_2})(|I_{\boldsymbol{Y},\boldsymbol{X}_1} - I_{\boldsymbol{Y},\boldsymbol{X}_2}| + |C_{\boldsymbol{Y},\boldsymbol{X}_1} - C_{\boldsymbol{Y},\boldsymbol{X}_2}|)}{C_{\boldsymbol{Y},\boldsymbol{X}_1} C_{\boldsymbol{Y},\boldsymbol{X}_2}} \\ &\leq \frac{|I_{\boldsymbol{Y},\boldsymbol{X}_1} - I_{\boldsymbol{Y},\boldsymbol{X}_2}| + |C_{\boldsymbol{Y},\boldsymbol{X}_1} - C_{\boldsymbol{Y},\boldsymbol{X}_2}|}{\max(C_{\boldsymbol{Y},\boldsymbol{X}_1}, C_{\boldsymbol{Y},\boldsymbol{X}_2})}.\end{aligned}$$

The result then comes from (42), (43), and (45). $\triangle$

Let us apply the previous result to our different examples.

Corollary 4. *Let $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \Gamma_\Theta$ with $\Theta = [c_1, c_2]$; define γ_i, $i = 1, 2$, such that $c_i = \gamma_i H_{\boldsymbol{Y}}$. Then*

$$|\Delta^\bullet_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta^\bullet_{\boldsymbol{Y},\boldsymbol{X}_2}| \leq (1 + \varkappa^\bullet_{1,\Theta}) D^I_{\boldsymbol{X}_1,\boldsymbol{X}_2} \qquad \text{and} \qquad |\delta^\bullet_{\boldsymbol{Y},\boldsymbol{X}_1} - \delta^\bullet_{\boldsymbol{Y},\boldsymbol{X}_2}| \leq \frac{1 + \varkappa^\bullet_{1,\Theta}}{\varkappa^\bullet_{2,\Theta}} d^I_{\boldsymbol{X}_1,\boldsymbol{X}_2}, \tag{46}$$

where $\bullet$ stands for S, R, P, and D, and where the different constants are expressed by

$\bullet$	$\varkappa^\bullet_{1,\Theta}$	$\varkappa^\bullet_{2,\Theta}$
S	$\alpha^\vee$	$(1-\alpha) + \alpha\gamma_{1,2}$
R	$\alpha^{\vee 2} + \dfrac{\alpha(1-\alpha)}{\sqrt{\gamma_1}}$	$\left((1-\alpha) + \alpha\sqrt{\gamma_{1,2}}\right)^2$
P	$\max\left(\dfrac{1-\alpha}{\gamma_1^\alpha}, \dfrac{\alpha}{\gamma_1^{1-\alpha}}, \mathbf{1}_{]0,1]}(\gamma_1)\right)$	$\gamma^\alpha_{1,2}$
D	$\dfrac{\alpha^\vee}{(\alpha^\wedge)^2} \dfrac{1}{(1+\gamma_{1,2})^2}$	$\left(\dfrac{\alpha}{\gamma_{1,2}} + (1-\alpha)\right)^{-1}$

with $\gamma_{1,2} = \min\left(\gamma_1, \dfrac{1}{\gamma_2}\right)$.

Proof. For the sake of simplicity, let us denote $m_i = \min(H_{\boldsymbol{Y}}, H_{\boldsymbol{X}_i})$ and $M_i = \max(H_{\boldsymbol{Y}}, H_{\boldsymbol{X}_i})$ for $i = 1, 2$. The proofs mainly rely upon the two following tools:

- Since $m_1 - m_2$ and $M_1 - M_2$ are of the same sign, we have

  $$|(m_1 - m_2)| + |M_1 - M_2| = |(m_1 - m_2) + (M_1 - M_2)| = |H_{\boldsymbol{X}_1} - H_{\boldsymbol{X}_2}|$$

  and then for any $f_1, f_2 \geq 0$

  $$|f_1(m_1 - m_2) + f_2(M_1 - M_2)| \leq f^\vee |H_{\boldsymbol{X}_1} - H_{\boldsymbol{X}_2}|, \tag{47}$$

  with $f^\vee = \max(f_1, f_2)$.
- For $i = 1, 2$, we have

  $$m_i \geq \left\{ \begin{array}{c} \min(1, \gamma_1) M_i \\ \min\left(1, \dfrac{1}{\gamma_2}\right) M_i \end{array} \right\} \geq \min\left(1, \gamma_1, \frac{1}{\gamma_2}\right) M_i = \gamma_{1,2} M_i,$$

  since $\gamma_{1,2} < 1$ as a direct consequence of $\gamma_1 < \gamma_2$.

For different cases, we need in particular to check (43), which holds whenever there exists some positive constant $\varkappa$ such that $C^{\bullet,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_i} \geq \varkappa M_i$ for $i = 1, 2$, since $\max(C^{\bullet,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_1}, C^{\bullet,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_2}) \geq \varkappa \max(M_1, M_2) \geq \varkappa C^I_{\boldsymbol{X}_1,\boldsymbol{X}_2}$.

- Complexity term $C^{S,\alpha}$: we have

$$\begin{aligned} |C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_1} - C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_2}| &= |\alpha m_1 + (1-\alpha)M_1 - \alpha m_2 - (1-\alpha)M_2| \\ &= |\alpha(m_1 - m_2) + (1-\alpha)(M_1 - M_2)| \\ &\leq \alpha^{\vee} |H_{\boldsymbol{X}_1} - H_{\boldsymbol{X}_2}| \end{aligned}$$

  from (47), with $f_1 = \alpha$ and $f_2 = (1-\alpha)$. Moreover, for $i = 1, 2$,

$$C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_i} = \alpha m_i + (1-\alpha)M_i \geq ((1-\alpha) + \alpha\gamma_{1,2})M_i.$$

- Complexity term $C^{R,\alpha}$: we have

$$\left|C^{R,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_1} - C^{R,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_2}\right| = \left|\alpha^2(m_1 - m_2) + (1-\alpha)^2(M_1 - M_2) + 2\alpha(1-\alpha)\sqrt{H_{\boldsymbol{Y}}}\left(\sqrt{H_{\boldsymbol{X}_1}} - \sqrt{H_{\boldsymbol{X}_2}}\right)\right|.$$

  Furthermore, we may obtain

$$|\alpha^2(m_1 - m_2) + (1-\alpha)^2(M_1 - M_2)| \leq \alpha^{\vee 2} |H_{\boldsymbol{X}_1} - H_{\boldsymbol{X}_2}|$$

  and

$$\left|\sqrt{H_{\boldsymbol{Y}}}\left(\sqrt{H_{\boldsymbol{X}_1}} - \sqrt{H_{\boldsymbol{X}_2}}\right)\right| = \frac{\sqrt{H_{\boldsymbol{Y}}}}{2\sqrt{\min(H_{\boldsymbol{X}_1}, H_{\boldsymbol{X}_2})}} |H_{\boldsymbol{X}_1} - H_{\boldsymbol{X}_2}| \leq \frac{1}{2\sqrt{\gamma_1}} |H_{\boldsymbol{X}_1} - H_{\boldsymbol{X}_2}|.$$

  Hence,

$$\left|C^{R,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_1} - C^{R,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_2}\right| \leq \left(\alpha^{\vee 2} + \frac{\alpha(1-\alpha)}{\sqrt{\gamma_1}}\right) |H_{\boldsymbol{X}_1} - H_{\boldsymbol{X}_2}|$$

  from (47), with $f_1 = \alpha^2$ and $f_2 = (1-\alpha)^2$. Moreover, one can prove for $i = 1, 2$

$$C^{R,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_i} = \left(\alpha\sqrt{m_i} + (1-\alpha)\sqrt{M_i}\right)^2 \geq \left((1-\alpha) + \alpha\sqrt{\gamma_{1,2}}\right)^2 M_i.$$

- Complexity term $C^{P,\alpha}$: we have (by assuming that $H_{\boldsymbol{X}_2} > H_{\boldsymbol{X}_1}$)

$$\begin{aligned} \left|C^{P,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_1} - C^{P,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_2}\right| &= \left|m_1^{\alpha} M_1^{1-\alpha} - m_2^{\alpha} M_2^{1-\alpha}\right| \\ &= \begin{cases} H^{\alpha}_{\boldsymbol{Y}}(H^{1-\alpha}_{\boldsymbol{X}_2} - H^{1-\alpha}_{\boldsymbol{X}_1}) & \text{if } H_{\boldsymbol{Y}} \leq \min(H_{\boldsymbol{X}_1}, H_{\boldsymbol{X}_2}), \\ H^{1-\alpha}_{\boldsymbol{Y}}(H^{\alpha}_{\boldsymbol{X}_2} - H^{\alpha}_{\boldsymbol{X}_1}) & \text{if } H_{\boldsymbol{Y}} \geq \max(H_{\boldsymbol{X}_1}, H_{\boldsymbol{X}_2}), \\ H^{\alpha}_{\boldsymbol{Y}} H^{1-\alpha}_{\boldsymbol{X}_2} - H^{\alpha}_{\boldsymbol{X}_1} H^{1-\alpha}_{\boldsymbol{Y}} & \text{otherwise.} \end{cases} \end{aligned}$$

  Note that the third case cannot occur if $\gamma_1 \geq 1$. We have

$$\begin{aligned} \left|C^{P,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_1} - C^{P,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_2}\right| &\leq \begin{cases} \frac{1-\alpha}{\gamma_1^{\alpha}}(H_{\boldsymbol{X}_2} - H_{\boldsymbol{X}_1}) & \text{if } H_{\boldsymbol{Y}} \leq \min(H_{\boldsymbol{X}_1}, H_{\boldsymbol{X}_2}), \\ \frac{\alpha}{\gamma_1^{1-\alpha}}(H_{\boldsymbol{X}_2} - H_{\boldsymbol{X}_1}) & \text{if } H_{\boldsymbol{Y}} \geq \max(H_{\boldsymbol{X}_1}, H_{\boldsymbol{X}_2}), \\ H_{\boldsymbol{X}_2} - H_{\boldsymbol{X}_1} & \text{otherwise} \end{cases} \\ &\leq \max\left(\frac{1-\alpha}{\gamma_1^{\alpha}}, \frac{\alpha}{\gamma_1^{1-\alpha}}, \mathbf{1}_{]0,1]}(\gamma_1)\right) |H_{\boldsymbol{X}_2} - H_{\boldsymbol{X}_1}|. \end{aligned}$$

  Moreover, we may obtain for $i = 1, 2$

$$C^{P,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_i} = m_i^{\alpha} M_i^{1-\alpha} \geq \gamma_{1,2}^{\alpha} M_i.$$

- Complexity term $C^{D,\alpha}$: we have

$$\begin{aligned}\left|C^{D,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_1} - C^{D,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_2}\right| &= \frac{|\alpha M_1 M_2(m_1-m_2) + (1-\alpha)m_1m_2(M_1-M_2)|}{(\alpha M_1 + (1-\alpha)m_1)(\alpha M_2 + (1-\alpha)m_2)} \\ &\le \frac{\alpha^\vee}{(\alpha^\wedge)^2}\frac{M_1M_2}{(m_1+M_1)(m_2+M_2)}|H_{\boldsymbol{X}_2} - H_{\boldsymbol{X}_1}| \qquad (48)\\ &\le \frac{\alpha^\vee}{(\alpha^\wedge)^2}\frac{1}{(1+\gamma_{1,2})^2}|H_{\boldsymbol{X}_2} - H_{\boldsymbol{X}_1}|.\end{aligned}$$

Equation (48) is obtained using (47) with $f_1 = \alpha M_1 M_2$ and $f_2 = (1-\alpha)m_1m_2$. Finally, we also have for $i = 1, 2$

$$C^{D,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_i} = \left(\frac{\alpha}{m_i} + \frac{1-\alpha}{M_i}\right)^{-1} \ge \left(\frac{\alpha}{\gamma_{1,2}} + (1-\alpha)\right)^{-1} M_i. \quad \triangle$$

Remark 8. Note that when $\alpha \le \frac{1}{2}$, the measure $\Delta^{S,\alpha}$ is a metric and so we derive (46) directly from **[P3′]**.

The authors would like to thank the referee whose suggestions and fruitful comments helped them to improve the paper.

REFERENCES

1. Shannon, C.E., A Mathematical Theory of Communication (continued), *Bell Syst. Tech. J.*, 1948, vol. 27, no. 4, pp. 623–656.
2. Crutchfield J.P., Information and Its Metric, *Nonlinear Structures in Physical Systems: Pattern Formation, Chaos, and Waves (Proc. 2nd Woodward Conf., San Jose State Univ., 1989)*, Lam, L. and Morris, H.C., Eds., New York: Springer, 1990, pp. 119–130.
3. Hillman, C., A Formal Theory of Information: I. Statics, *Preprint*, 1998. Available from `http://citeseer.ist.psu.edu/89520.html`.
4. Bennett, C.H., Gács, P., Li, M., Vitányi, P.M.B., and Zurek, W.H., Information Distance, *IEEE Trans. Inform. Theory*, 1998, vol. 44, no. 4, pp. 1407–1423.
5. Li, M., Chen, X., Li, X., Ma, B., and Vitányi, P.M.B., The Similarity Metric, *IEEE Trans. Inform. Theory*, 2004, vol. 50, no. 12, pp. 3250–3264.
6. Cilibrasi, R. and Vitányi, P.M.B., Clustering by Compression, *IEEE Trans. Inform. Theory*, 2005, vol. 51, no. 4, pp. 1523–1545.
7. Li, M., Badger, J.H., Chen, X., Kwong, S., Kearney, P., and Zhang, H., An Information-Based Sequence Distance and Its Application to Whole Mitochondrial Genome Phylogeny, *Bioinformatics*, 2001, vol. 17, no. 1, pp. 149–154.
8. Cilibrasi, R., Vitányi, P., and de Wolf, R., Algorithmic Clustering of Music, 2003, e-print `arxiv.org/cs.SD/0303025`.
9. Cilibrasi, R. and Vitányi, P., Automatic Meaning Discovery Using Google, 2005, e-print `arxiv.org/cs.CL/0412098`.
10. Grünwald, P. and Vitányi, P., Shannon Information and Kolmogorov Complexity, 2004, e-print `arxiv.org/cs.IT/0410002`.
11. Leung-Yan-Cheong, S.K. and Cover, T.M., Some Equivalences between Shannon Entropy and Kolmogorov Complexity, *IEEE Trans. on Inf. Theory*, 1978, vol. 24, no. 3, pp. 331–338.
12. Hammer, D., Romashchenko, A., Shen, A., and Vereshchagin, N., Inequalities for Shannon Entropy and Kolmogorov Complexity, *J. Comput. Syst. Sci.*, 2000, vol. 60, no. 2, pp. 442–464.

13. Kraskov, A., Stögbauer, H., Andrzejak, R.G., and Grassberger, P., Hierarchical Clustering Based on Mutual Information, 2003, e-print arxiv.org/q-bio/0311039.

14. Liu, H. and Motoda, H., *Feature Selection for Knowledge Discovery and Data Mining*, Boston: Kluwer, 1998.

15. Robineau, J.-F., Méthodes de sélection de variables (parmi un grand nombre) dans un cadre de discrimination, *PhD Thesis*, Grenoble, France: Univ. Joseph Fourier, 2004.

16. Granger, C.W., Maasoumi, E., and Racine, J., A Dependence Metric for Possibly Nonlinear Processes, *J. Time Ser. Anal.*, 2004, vol. 25, no. 5, pp. 649–669.

17. Cover, T.M. and Thomas, J.A., *Elements of Information Theory*, New York: Wiley, 1991.

18. Kaltchenko, A., Algorithms for Estimation of Information Distance with Application to Bioinformatics and Linguistics, in *Proc. 2004 Canadian Conf. on Electrical and Computer Engineering*, 2004, vol. 4, p. 2255. Available from arxiv.org/cs.CC/0404039.

19. Li, M. and Vitányi, P., *An Introduction to Kolmogorov Complexity and Its Applications*, New York: Springer, 1997, 2nd ed.

20. Li, M, Chen, X., Li, X., Ma, B., and Vitányi, P., The Similarity Metric, in *Proc. 14th Annual ACM–SIAM Symp. on Discrete Algorithms (SODA-03), New York*, 2003, pp. 863–872.

21. Ullah, A., Entropy, Divergence and Distance Measures with Econometric Applications, *J. Stat. Plan. Inference*, 1996, vol. 49, no. 1, pp. 137–162.