HAL
open science

Normalized information-based divergences

Jean-François Coeurjolly, Rémy Drouilhet, Jean-François Robineau

To cite this version:

Jean-François Coeurjolly, Rémy Drouilhet, Jean-François Robineau. Normalized information-based divergences. Problems of Information Transmission, 2007, 43 (3), pp.167-189. 10.1134/S0032946007030015 . hal-00022566v2

HAL Id: hal-00022566

https://hal.science/hal-00022566v2

Submitted on 13 Nov 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

hal-00022566, version 2 - 13 Nov 2006

Normalized information-based divergences

By J.-F. Coeurjolly, R. Drouilhet and J.-F. Robineau
University of Grenoble 2, France

November 13, 2006

Abstract

This paper is devoted to the mathematical study of some divergences based on the mutual information 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.

Keywords and phrases. Information theory, entropy distance, information distance, triangular inequality, redundancy.

1 Introduction

Shannon information theory, usually just called *information* theory was introduced in 1948, Shannon (1948). The theory aims at providing a 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 by some encoding way. In the Shannon approach, the objects to be encoded are assumed to be outcomes of a known source. Shannon 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 noticed 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, Crutchfield (1990), Hillman (1998) introduced the *entropy distance* defined as the sum of the conditional entropies. Other interesting measures are the *information distance* Bennett et al. (1998) and its normalized version named *similarity metric* introduced by Li et al. (2004) 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 the uncomputable notion of Kolmogorov complexity.

Recent papers have demonstrated useful application of suitable version of the similarity metric in areas as diverse as genomics, virology, languages, literature, music, handwritten digits and astronomy, Cilibrasi and Vitányi (2005b). To apply the metric to real data, the authors have to replace the use of the noncomputable Kolmogorov complexity by an approximation using standard real-world compressors : GenCompress for genomics, Li et al. (2001), the *Normalized Compression Distance (NCD)* for music clustering, Cilibrasi et al. (2003), the *Normalized Google Distance (NGD)* for automatic meaning discovery, Cilibrasi and Vitányi (2005a), are examples of effective compressors. To include the information distance and the similarity metric in a framework based on information theory concepts, we make use of the principle that *expected Kolmogorov complexity equals Shannon entropy* and interested reader can refer to Grünwald and Vitányi (2004), Leung-Yan-Cheong and Cover (1978), Hammer et al. (2000) for more details. Consequently, the entropy and information distances are both expressed in terms

of conditional entropies: the first one as their sum and the second one as their maximum. Kraskov et al. (2003) gives a proof of the triangular inequality for these distances and their respective normalized versions.

In the supervised learning framework, the use of some selection method of covariables amoung a large number is required when it is assumed that the data size is too small with respect to the number of the available covariables in order to apply any existing discriminant analysis method. Such a problem has been widely treated, Liu and Motoda (1998). The approach undertaken by Robineau (2004) is mainly based on three kinds of methodological tools. The first one is a supervised quantization method consisting in the simplification of covariables too complex (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 not be too complex. The second one is a more usual step by step selection method combining the simplified covariables together in order to detect cluster of data of the same class. The last one is aimed at detecting redundancy among the covariables set. These three tasks may be realized using the entropy or information distances (or their normalized versions). Let us emphasize some properties allowing to understand the usefulness of these criterions 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. Robineau (2004) proposes then 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 of $I_{\boldsymbol{X},\boldsymbol{Y}}$ reached when $\boldsymbol{X}$ and $\boldsymbol{Y}$ share exactly the same information. The notion of normalized information-based divergence $\delta_{\boldsymbol{X},\boldsymbol{Y}}$ derives directly by dividing the associated information-based divergence $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ by the complexity term $C_{\boldsymbol{X},\boldsymbol{Y}}$. The normalized version d^E and d^I of D^E and D^I are particular examples. Other examples are given in Robineau (2004). Amoung them, one is of particular interest since its complexity term C^S is the mean of the marginal entropies.

The associated (non-normalized) information-based divergence Δ^S is not so different from D^E since it corresponds to its half. Nevertheless, the expression of its complexity term C^S really differs from the complexity term C^E of D^E (i.e. the joint entropy). For pratical purposes, we may argue that D^I, D^E and Δ^S are not well-suited in 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 IB-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 definition and their main properties of the entropy and information distances (and their normalized version). Similarly to Granger et al. (2004), we extract the main characteristics to define some general concept of information divergence which could be theoretically applied in a more general setting (continuous, discrete, ...). Section 3 concentrates itself on categorical data (and in particular discrete) random vectors, as it is usually the case in most of applications using entropy or information distance. We give the definition of (normalized) information-based divergence and propose several examples. We study their mathematical properties in a general context and propose some sufficient conditions for these divergences to verify some triangular's 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 sequel, $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are three elements of such a set Γ.

2.1 Some notation

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

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

In the same way, one can define the joint entropy of $\boldsymbol{X}$ and $\boldsymbol{Y}$ denoted by $H_{\boldsymbol{X},\boldsymbol{Y}}$, the conditional entropy of $\boldsymbol{X}$ (resp. $\boldsymbol{Y}$) by $\boldsymbol{Y}$ (resp. $\boldsymbol{X}$) denoted by $H_{\boldsymbol{X}|\boldsymbol{Y}}$ (resp. $H_{\boldsymbol{Y}|\boldsymbol{X}}$). 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, the following relations hold (see *e.g.* Cover and Thomas (1991)):

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

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

2.2 Definition and characteristics

We now shall present some measures allowing to overcome some drawbacks of the mutual information. As a first generalization, several authors noticed 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 litterature. The first one called "entropy distance" is derived from the domain of information theory. The second one called "information distance" originates in works around the Kolmogorov complexity. Both measures are defined (when they exist) for two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ by:

- Entropy distance:

$$D^{E}_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X}|\boldsymbol{Y}} + H_{\boldsymbol{Y}|\boldsymbol{X}} \tag{3}$$

- Information distance:

$$D^{I}_{\boldsymbol{X},\boldsymbol{Y}} = \max\left(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}}\right). \tag{4}$$

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

$$D^{E}_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X},\boldsymbol{Y}} - I_{\boldsymbol{X},\boldsymbol{Y}} \qquad \text{and} \qquad D^{I}_{\boldsymbol{X},\boldsymbol{Y}} = \max\left(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}\right) - I_{\boldsymbol{X},\boldsymbol{Y}}. \tag{5}$$

The quantities $H_{\boldsymbol{X},\boldsymbol{Y}}$ and $\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ are upper-bounds of the mutual information $I_{\boldsymbol{X},\boldsymbol{Y}}$ that 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 $\boldsymbol{X}$ (resp. $\boldsymbol{Y}$) predicts $\boldsymbol{Y}$ (resp. $\boldsymbol{X}$) with probability 1.

These measures satisfy

$$D^{E}_{\boldsymbol{X},\boldsymbol{Y}} \leq H_{\boldsymbol{X},\boldsymbol{Y}} \qquad \text{and} \qquad D^{I}_{\boldsymbol{X},\boldsymbol{Y}} \leq \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}), \tag{6}$$

where the equality holds if the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. As noticed by Kaltchenko (2004), Li and Vitányi (1997) argued that in Bioinformatics an unnormalized distance may not be a proper evolutionary distance measure. It would put two long and complex sequences that differ only by a tiny fraction of the total information as dissimilar as two short sequences that differ by the same absolute amount and are completely random with respect to one another. To overcome this problem within the algorithmic framework Li and Vitányi (1997) form two normalized versions of distances D^{E} and D^{I}. Their Shannon version have been proposed and studied by Kraskov et al. (2003)

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

$$d^{I}_{\boldsymbol{X},\boldsymbol{Y}} = \frac{\max\left(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}}\right)}{\max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})}$$

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* $d^{E}_{\boldsymbol{X},\boldsymbol{Y}} = 0$ *(resp.* $d^{I}_{\boldsymbol{X},\boldsymbol{Y}} = 0$*) when* $H_{\boldsymbol{X}} = H_{\boldsymbol{Y}} = 0$*.*

We are encouraged to define the following class of equivalence: the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be equivalent if $\boldsymbol{X}$ (resp. $\boldsymbol{Y}$) predicts $\boldsymbol{Y}$ (resp. $\boldsymbol{X}$) with probability 1 and one will denote

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

Due to the previous convention

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

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 $d^I_{\boldsymbol{X},\boldsymbol{Y}}$.*

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

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

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

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

To our knowledge, these results have been proved by Kraskov et al. (2003). Proofs are very similar to proofs of Li et al. (2003) who consider the algorithmic version of these distances. The proof is then omitted, but in Section 3.3, we propose a result extending this one in the sense that we give conditions on measures that can be written as (8) and (9) to constitute a metric.

2.3 Concept of information divergence

We can exhibit from the previous study related to D^I, D^E, d^I and d^E, some characteristics useful for an attempt to define the concept of information divergence denoted by Δ in a more general setting. Let us first consider a similarity measure $\mathcal{I}_{\boldsymbol{X},\boldsymbol{Y}}$ (not necessarily the mutual information) 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 minimum (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 maximum 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 a triangular inequality: $\Delta_{\boldsymbol{X},\boldsymbol{Y}} \leq \Delta_{\boldsymbol{X},\boldsymbol{Z}} + \Delta_{\boldsymbol{Z},\boldsymbol{Y}}$.

[P7] invariant under continuous and strictly increasing transformations $\boldsymbol{\varphi}(\cdot)$, $\boldsymbol{\psi}(\cdot)$ of the vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, whenever they are quantitative random vectors.

There exists a large litterature on the discussion of criteria satisying the previous stated properties. We may cite Ullah (1996), or a recent work of Granger et al. (2004) who propose to detect the dependence between two possibly nonlinear processes through the Bhattacharya-Matusita-Hellinger measure of dependence given by

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

where f_1 (resp. f_2) is the joint density (resp. the product of marginal densities) of $\boldsymbol{X}$ and $\boldsymbol{Y}$. This measure, that has the other advantage to be applicable to 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 respectively the complexity of vectors $\boldsymbol{X}$, $\boldsymbol{Y}$ and $(\boldsymbol{X},\boldsymbol{Y})$. 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 $\mathcal{C}_{\boldsymbol{Y},\boldsymbol{X}_1} = \mathcal{C}_{\boldsymbol{Y},\boldsymbol{X}_2}$): $\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} < \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}$ 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. $\mathcal{I}_{\boldsymbol{Y},\boldsymbol{X}_1} = \mathcal{I}_{\boldsymbol{Y},\boldsymbol{X}_2}$): $\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} < \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}$ 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 the fact that

[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 $\Delta_{\boldsymbol{X}_1,\boldsymbol{X}_2} \simeq 0$) then 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]** the previous statements 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. The properties **[P2]**-**[P3]** and the form (11) implies that $\mathcal{C}_{\boldsymbol{X},\boldsymbol{Y}}$ is an upper bound of $\mathcal{I}_{\boldsymbol{X},\boldsymbol{Y}}$ reached when $\boldsymbol{X}$ and $\boldsymbol{Y}$ share exactly the same information.

In the rest of this paper we concentrate ourself on criteria described by (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 respectively called an information-based divergence and a normalized information-based divergence (in short IB-divergence and NIB-divergence) if they can respectively be written*

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

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

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 IB-divergence $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ (resp. a NIB-divergence $\delta_{\boldsymbol{X},\boldsymbol{Y}}$) satisfies properties **[P1]**-**[P4]** (resp. **[P1]**-**[P5]**). In the rest of the paper, the term $C_{\boldsymbol{X},\boldsymbol{Y}}$ is expressed as

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

where $f_C(\cdot,\cdot,\cdot)$ is a nonnegative function. Under such an expression of $C_{\boldsymbol{X},\boldsymbol{Y}}$, the property **[P7]** is ensured since the conditional entropies and the 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 for which we adopt the following convention: an IB-divergence (resp. a NIB-divergence) satisfying the triangular inequality is denoted D(resp. d) rather than Δ(resp. δ) . Moreover, each example will be particularized by some discriminating additonal letter in the same manner as D^E and D^I (resp. d^E and d^I) which clearly constitute IB-divergences (resp. NIB-divergences).

In Robineau (2004), we investigate about 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 NIB-divergences:

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

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

Their non normalized version 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 IB-divergence or NIB-divergence with complexity terms of the form:

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

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 it is not ambiguous we set $m = m_{\boldsymbol{X},\boldsymbol{Y}}$ and $M = M_{\boldsymbol{X},\boldsymbol{Y}}$. To be convinced that IB-divergences and NIB-divergences with complexity terms of the form (17) satisfy (ii) of Definition 2, let us notice that

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

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

We pay now particular attention on the complexity terms $C^{D,\alpha}$, $C^{S,\alpha}$, $C^{R,\alpha}$ and $C^{P,\alpha}$ of the form (17) respectively with $g^{D}(\cdot) = 1/\cdot$, $g^{S}(\cdot) = \cdot$, $g^{R}(\cdot) = \sqrt{\cdot}$ 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 $C^{1/2}_{\boldsymbol{X},\boldsymbol{Y}} = g^{-1}\left(\frac{1}{2}g(H_{\boldsymbol{X}}) + \frac{1}{2}g(H_{\boldsymbol{Y}})\right)$. When $\alpha = \frac{1}{2}$, $\Delta^{\bullet,\alpha}$ and $\delta^{\bullet,\alpha}$ will be simply denoted by $\Delta^{\bullet}$ and $\delta^{\bullet}$ where $\bullet$ stands for S, R, P and D.

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

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

Clearly, the previous representation of $\Delta^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}$ (resp. $\delta^{D,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}$) as a convex combination of $\Delta^{\min}_{\boldsymbol{X},\boldsymbol{Y}}$ and $D^{I}_{\boldsymbol{X},\boldsymbol{Y}}$ (resp. $\delta^{\min}_{\boldsymbol{X},\boldsymbol{Y}}$ and $d^{I}_{\boldsymbol{X},\boldsymbol{Y}}$) introduces a degree of freedom that could be useful for practical purposes in prediction framework where $\boldsymbol{Y}$ could represent some class variable. According to the parameter α one may favour to take into account between one or two prediction terms amoung $H_{\boldsymbol{X}|\boldsymbol{Y}}$ and $H_{\boldsymbol{Y}|\boldsymbol{X}}$ (resp. $\frac{H_{\boldsymbol{X}|\boldsymbol{Y}}}{H_{\boldsymbol{X}}}$ and $\frac{H_{\boldsymbol{Y}|\boldsymbol{X}}}{H_{\boldsymbol{Y}}}$). This possibility to introduce a non uniform 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 $C^{\gamma,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} = ||\left(\alpha^{\frac{1}{\gamma}}m, (1-\alpha)^{\frac{1}{\gamma}}M\right)||_{\gamma}$, where $||\boldsymbol{x}||_{\gamma} = \left(\sum_{i=1}^{2}|x_i|^{\gamma}\right)^{1/\gamma}$ denotes the norm of some vector $\boldsymbol{x}$ of length 2. Note that for any $0 \leq \alpha \leq 1$, we have*

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

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 transformation 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. Before, some preliminary result is given.

Lemma 3 *Let $C^{(1)}$ and $C^{(2)}$ two complexity terms of the form (17) with function g_1 and g_2. Assume either that the function $g_1 \circ g_2^{-1}$ is concave or that the function $g_2 \circ g_1^{-1}$ is convex, then $C^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} \leq C^{(2)}_{\boldsymbol{X},\boldsymbol{Y}}$*

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 may assert

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

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

Proposition 4 *For any $\Delta^{(1)}$, $\Delta^{(2)}$ IB-divergences or any $\delta^{(1)}$, $\delta^{(2)}$ NIB-divergences with respective complexity terms $C^{(1)}$ and $C^{(2)}$, the following equivalence holds:*

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

Since, for any $0 \leq \alpha \leq \alpha' \leq 1$,

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

the associated IB-divergences and NIB-divergences are then 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

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

and the right-hand side is direct. Since $g^P \circ (g^D)^{-1}(\cdot) = -\log(\cdot)$, $g^R \circ (g^P)^{-1}(\cdot) = \exp(\frac{1}{2}\cdot)$ and $g^S \circ (g^R)^{-1}(\cdot) = (\cdot)^2$ are convex functions, (24) is a direct consequence of Lemma 3. ■

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

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

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

The following proposition gives a larger class of examples of IB-divergences and NIB-divergences.

Proposition 5 *Let $(\alpha^{(j)})_{j=1,\ldots,J}$ be some vector of probability weights for some $J \geq 1$.*

(i) Let $\delta^{(1)}, \ldots, \delta^{(J)}$, J NIB-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 a NIB-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 $\Delta^{(1)}, \ldots, \Delta^{(j)}$*, J IB-divergences and* $\delta^{(1)}, \ldots, \delta^{(j)}$*, J NIB-divergences with complexity terms* $C^{(1)}_{\boldsymbol{X},\boldsymbol{Y}}, \ldots, C^{(J)}_{\boldsymbol{X},\boldsymbol{Y}}$ *then the measures defined by*

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

are also respectively an IB-divergence and a NIB-divergence.

The proof is immediate.

3.2 Around the property [P3]

The fact that an IB-divergence Δ(resp. NIB-divergence δ) satisfies the property **[P3]** may be expressed by: $\Delta_{\boldsymbol{X},\boldsymbol{Y}} = 0 \Leftrightarrow D^I_{\boldsymbol{X},\boldsymbol{Y}} = 0$ (resp. $\delta_{\boldsymbol{X},\boldsymbol{Y}} = 0 \Leftrightarrow d^I_{\boldsymbol{X},\boldsymbol{Y}} = 0$). In fact, **[P3]** should be extended to the more useful assumption: $\Delta_{\boldsymbol{X},\boldsymbol{Y}}$ (or $\delta_{\boldsymbol{X},\boldsymbol{Y}}$) is near from 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 IB-divergence Δ(resp. a NIB-divergence δ):

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

$$\Delta_{\boldsymbol{X},\boldsymbol{Y}} \leq \varepsilon \Longrightarrow D^I_{\boldsymbol{X},\boldsymbol{Y}} \leq \gamma \quad (\text{resp.} \quad \delta_{\boldsymbol{X},\boldsymbol{Y}} \leq \varepsilon \Longrightarrow d^I_{\boldsymbol{X},\boldsymbol{Y}} \leq \gamma).$$

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

$$D^I_{\boldsymbol{X},\boldsymbol{Y}} \leq \gamma \Longrightarrow \Delta_{\boldsymbol{X},\boldsymbol{Y}} \leq \varepsilon \quad (\text{resp.} \quad d^I_{\boldsymbol{X},\boldsymbol{Y}} \leq \gamma \Longrightarrow \delta_{\boldsymbol{X},\boldsymbol{Y}} \leq \varepsilon).$$

An IB-divergence Δ(resp. a NIB-divergence δ) inherits of the previous property if it satisfies:

[P3bis(Υ, k_1, k_2)] there exists some positive constants k_1, k_2 ($k_1 \leq k_2$) such that for all $(\boldsymbol{X}, \boldsymbol{Y}) \in \Upsilon \subset \Gamma^2$:

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

Among our examples, we assert that D^E and d^E both satisfy $[\mathbf{P3bis}(\Gamma^2, 1, 2)]$ that is

$$D^I_{\boldsymbol{X},\boldsymbol{Y}} \leq D^E_{\boldsymbol{X},\boldsymbol{Y}} \leq 2D^I_{\boldsymbol{X},\boldsymbol{Y}} \quad (\text{resp. } d^I_{\boldsymbol{X},\boldsymbol{Y}} \leq d^E_{\boldsymbol{X},\boldsymbol{Y}} \leq 2d^I_{\boldsymbol{X},\boldsymbol{Y}}).$$

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 $[\mathbf{P3bis}(\Upsilon, k_1, 1)]$ then the associated δ also satisfies $[\mathbf{P3bis}(\Upsilon, k_1, 1)]$ since

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

And so in the rest of this section, the results presented hereafter will be only expressed for IB-divergences.

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 two following subsets of $\mathcal{D}^g$:

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

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

In the sequel, two results ensuring that an IB-divergence Δ^α of the form (17) satisfies $[\mathbf{P3bis}(\Upsilon, k_1, k_2)]$, are proposed. The difference relies upon the framework: the constants k_1 and k_2 differ whenever the set Υ differs.

Proposition 6 *For any $\Theta \in \mathcal{E}_1^g$ the IB-divergence Δ^α satisfies* $[\mathbf{P3bis}(\Upsilon_\Theta, \alpha^\wedge \frac{\kappa^g_{\inf,\Theta}}{\kappa^g_{\sup,\Theta}}, 1)]$ *with* $\Upsilon_\Theta = \left\{(\boldsymbol{X}, \boldsymbol{Y}) \in \Gamma^2 : H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}, I_{\boldsymbol{X},\boldsymbol{Y}} \in \Theta\right\}$.

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

$$\begin{aligned} g^{-1}\left(\frac{g(x+z)+g(y+z)}{2}\right) - z &= \left(\alpha(g(x+z)-g(z)) + (1-\alpha)(g(y+z)-g(z))\right)(g^{-1})'(c_1) \\ &= \alpha|g'(c_2)||(g^{-1})'(c_1)| \times x + (1-\alpha)|g'(c_3)||(g^{-1})'(c_1)| \times y, \end{aligned}$$

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, we obtain for all x, y, z:

$$g^{-1}\left(\frac{g(x+z)+g(y+z)}{2}\right) - z \geq \alpha^{\wedge} \frac{\kappa^g_{\inf,\Theta}}{\kappa^g_{\sup,\Theta}} \max(x,y)$$

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

Proposition 7 *For any $\Theta \in \mathcal{E}^g_2$ the IB-divergence Δ^{α} satisfies* ***[P3bis(***$\Gamma^2_{\Theta}, 1-\alpha\frac{\kappa^g_{\sup,\Theta}}{\kappa^g_{\inf,\Theta}}, 1$***)]*** *with $\Gamma_{\Theta} = \{\boldsymbol{Z} \in \Gamma : H_{\boldsymbol{Z}} \in \Theta\}$.*

Proof.

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

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

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

which leads to the result. ■

For sake of simplicity, we denote by $\kappa^{\bullet}_{\inf,\Theta}$ and $\kappa^{\bullet}_{\sup,\Theta}$ instead of $\kappa^{g^{\bullet}}_{\inf,\Theta}$ and $\kappa^{g^{\bullet}}_{\sup,\Theta}$

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

Proposition 8 *$\Delta^{\bullet,\alpha}$ satisfies* ***[P3bis(***$\Upsilon_{\Theta}, k_1^{a,\bullet}, 1$***)]*** *(from Proposition 6),* ***[P3bis(***$\Gamma^2_{\Theta}, k_1^{b,\bullet}, 1$***)]*** *(from Proposition 7) and* ***[P3bis(***$\Gamma^2_{\Theta}, k_1^{c,\bullet}, 1$***)]*** *where $\bullet$ stands for S, R, P and D, and*

$\bullet$	Θ	$\kappa^{\bullet}_{\inf,\Theta}$	$\kappa^{\bullet}_{\sup,\Theta}$	$k_1^{a,\bullet} = \alpha^{\wedge}\frac{\kappa_{\inf,\Theta}}{\kappa^{\bullet}_{\sup,\Theta}}$	$k_1^{b,\bullet} = 1 - \alpha\frac{\kappa_{\sup,\Theta}}{\kappa^{\bullet}_{\inf,\Theta}}$	$k_1^{c,\bullet}$
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* $\rho < \frac{1}{\alpha^2}$*)*	$(1-\alpha)\left(1 - \frac{\alpha}{(1+\frac{1}{\sqrt{\rho}})^2}\right)$
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* $\rho < \frac{1}{\alpha}$*)*	$\frac{\rho^{\alpha}-1}{\rho-1}$
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 $0 < c_1 \leq c_2 < +\infty$, $\rho = \frac{c_2}{c_1}$.

Proof. The computations of $k_1^{a,\bullet}$ and $k_1^{b,\bullet}$ derive from Proposition 6 and 7. Hence, let us concentrate only on $k_1^{c,\bullet}$ for the complexity terms C^R, C^P and C^D. Let us denote by $m = \min(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$ and by $M = \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}})$.

- Complexity term C^R:

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

And so,

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

The result is obtained by noticing that

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

- Complexity term C^P : by using a Taylor expansion with integral rest, one obtains

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

And so,

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

which leads to the result.

- Complexity term C^D :

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

■

3.3 Around the triangular inequality's property

The question arises now whether an IB-divergence or a NIB-divergence satisfies the property **[P6]** that is a triangular inequality. The following proposition establishes sufficient conditions for such measures to constitute a metric.

Lemma 9

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

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

Proof. From 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 derives from (2). ■

Proposition 10 *Assume the complexity term defining an IB-divergence satisfies the following property:*
(i)

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

Then, the associated IB-divergence satisfies the triangular 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
(ii)

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

then the associated NIB-divergence satisfies also a triangular inequality, that is

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

Proof. Since the following 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), we have immediately (33). Moreover, the following equation is valid

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

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

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

From (36) it follows

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

■

Remark 4 *In Proposition 10, there is no implication between (32) and (34). Indeed, one may check that the NIB-divergence δ^S (with $\alpha = 1/2$ for example) satisfies the first one but not the second one. Now consider a NIB-divergence with complexity term $C_{\boldsymbol{X},\boldsymbol{Y}} = \max\left(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}\right) + H_{\boldsymbol{X}|\boldsymbol{Y}} H_{\boldsymbol{Y}|\boldsymbol{X}}$. By choosing $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ such that $H_{\boldsymbol{X}|\boldsymbol{Y}} = H_{\boldsymbol{Y}|\boldsymbol{X}} = I_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{Z}}/3 = H_{\boldsymbol{X},\boldsymbol{Y}}/3 > 2$, one asserts that (34) is satisfied but not (32).*

Remark 5 *Let us consider a NIB-divergence δ with complexity term $C_{\boldsymbol{X},\boldsymbol{Y}} = C'_{\boldsymbol{X},\boldsymbol{Y}} + \max\left(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}\right)$ 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 a triangular inequality if C' also satisfies a triangular inequality. However, this is not a necessary condition. Indeed, the triangular 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 11

(i) The measures D^E, D^I satisfy the condition (32) and so are metrics.

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

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

(iv) Let $(\alpha^{(j)})_{j=1,\ldots,J}$ be some vector of probability weights for some $J \geq 1$. Let $\Delta^{(1)}, \ldots, \Delta^{(J)}$, J IB-divergences (resp. $\delta^{(1)}, \ldots, \delta^{(J)}$, J NIB-divergences) with complexity terms $C^{(1)}_{\boldsymbol{X},\boldsymbol{Y}}, \ldots, C^{(J)}_{\boldsymbol{X},\boldsymbol{Y}}$ satisfying (32) (resp. (32) and (34)) then these measures defined by (27) satisfy a triangular inequality.

Proof. (i) and (ii) Equation (29) corresponds exactly to (32) for $C^E_{\boldsymbol{X},\boldsymbol{Y}} = H_{\boldsymbol{X},\boldsymbol{Y}}$. And since $H_{\boldsymbol{X},\boldsymbol{Z}} \geq \max(H_{\boldsymbol{X}}, H_{\boldsymbol{Z}})$, 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 is sufficient to prove (32) which is quite obvious. Indeed,

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

(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 $(1-2\alpha)(C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} - H_{\boldsymbol{Z}}) + C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}} < C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}$. And so, (32) can never be satisfied for $\alpha > \frac{1}{2}$. Now, if $\alpha \leq \frac{1}{2}$, we have

$$C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} > (1-\alpha)(m+M) \geq C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}$$

- $H_{\boldsymbol{Z}} > M$:

$$C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} = (2\alpha - 1)H_{\boldsymbol{Z}} + (1-\alpha)(m+M) \geq \alpha + (1-\alpha)M = C^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Y}}.$$

- $m \leq H_{\boldsymbol{Z}} \leq M$:

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

(iv) trivial. ■

We assert 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 the 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 such that $H_{\boldsymbol{Z}} = \frac{1+\alpha}{\alpha} H_{\boldsymbol{X}} = \frac{1+\alpha}{\alpha} H_{\boldsymbol{Y}}$. This leads to

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

which is in contradiction with (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 a triangular inequality but our tool cannot be applied to prove it. We propose to weaken the property **[P6]** in the following way in order to obtain more results. An IB-divergence or NIB-divergence satisfies

[P6bis(Υ,c)] There exists $c \geq 1$ such that for all $(\boldsymbol{X}, \boldsymbol{Y}), (\boldsymbol{Y}, \boldsymbol{Z}), (\boldsymbol{X}, \boldsymbol{Z}) \in \Upsilon$

$$\Delta_{\boldsymbol{X},\boldsymbol{Y}} \leq c \times \left(\Delta_{\boldsymbol{X},\boldsymbol{Z}} + \Delta_{\boldsymbol{Y},\boldsymbol{Z}} \right).$$

Property **[P6]** is then equivalent to **[P6bis(Γ^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 **[P6bis(Γ^2,1)]**. When $\Upsilon \subsetneq \Gamma^2$ the property **[P6bis]** is in some sense local whereas it is global (as a classical triangular inequality) when $\Upsilon = \Gamma^2$.

Let us notice that if an IB-divergence (or NIB-divergence) satisfies **[P3bis(Υ, k_1, k_2)]**, then **[P6bis(Υ,$\frac{k_2}{k_1}$)]** is satisfied since

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

We then inherit a lot of results from Proposition 8 related to our examples. In particular $\Delta^{\bullet,\alpha}$ and $\delta^{\bullet,\alpha}$ (where $\bullet$ stands for S, R, P and D) both satisfy **[P6bis(Υ_Θ,$\frac{1}{k_1^a}$)]**, **[P6bis(Γ^2_Θ,$\frac{1}{k_1^b}$)]** and **[P6bis(Γ^2_Θ,$\frac{1}{k_1^c}$)]**.

In the rest of this section, we attempt to ensure the global property **[P6bis(Γ^2,c)]**. From Proposition 8 (with $\Theta = \mathbb{R}^+$), we assert that the divergences $\Delta^{S,\alpha}$ (when $\alpha > \frac{1}{2}$) and $\delta^{S,\alpha}$ (resp. $\Delta^{R,\alpha}$ and $\delta^{R,\alpha}$) satisfy **[P6bis(Γ^2,$\frac{1}{1-\alpha}$)]** (resp. **[P6bis(Γ^2,$\frac{1}{(1-\alpha)^2}$)]**).

When $\alpha \leq \frac{1}{2}$, we could improve the previous on $\Delta^{R,\alpha}$ by proving that it satisfies [**P6bis**(Γ^2,$\frac{1}{\alpha^2+(1-\alpha)^2}$)], in the same spirit of the proof leading to [**P3bis**]. Indeed,

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

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 notice that

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

We now give a further and general result allowing us, in particular, to improve [**P6bis**(Γ^2, $\frac{1}{1-\alpha}$)] for $\Delta^{S,\alpha}$ when $\alpha > \frac{1}{2}$.

Proposition 12 *Let us consider the following assumptions on a complexity term: there exists a constant $c \geq 1$ such that*

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

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

*If an IB-divergence satisfies (37) or a NIB-divergence satisfies (38), then they satisfy [**P6bis**(Γ^2,c)].*

Proof. Let us introduce

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

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

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

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

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

which leads to

$$\delta_{\boldsymbol{X},\boldsymbol{Y}} \leq \frac{c \times \left(C_{\boldsymbol{X},\boldsymbol{Z}} - I_{\boldsymbol{X},\boldsymbol{Z}}\right) + c \times \left(C_{\boldsymbol{Y},\boldsymbol{Z}} - I_{\boldsymbol{Y},\boldsymbol{Z}}\right)}{\max\left(C_{\boldsymbol{X},\boldsymbol{Z}}, C_{\boldsymbol{Y},\boldsymbol{Z}}\right)} \leq c \times \delta_{\boldsymbol{X},\boldsymbol{Z}} + c \times \delta_{\boldsymbol{Y},\boldsymbol{Z}}.$$

■

Corollary 13

The measures $\Delta^{S,\alpha}$ *for* $\alpha > \frac{1}{2}$ *satisfy* ***[P6bis(***Γ^2***,***$\frac{\alpha}{1-\alpha}$***)]***

Proof. Let us concentrate on $\Delta^{S,\alpha}$ for $\alpha > \frac{1}{2}$. Let $A = cC^{S,\alpha}_{\boldsymbol{X},\boldsymbol{Z}} + cC^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{Z}} - H_{\boldsymbol{Z}} - (c-1)\left(I_{\boldsymbol{X},\boldsymbol{Z}} + I_{\boldsymbol{Y},\boldsymbol{Z}}\right)$. Without loss of generality, we assume $H_x \leq H_{\boldsymbol{Y}}$. We distinguish three cases:

- $H_{\boldsymbol{Z}} \leq H_{\boldsymbol{X}} \leq H_{\boldsymbol{Y}}$: we have

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

  Then,

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

  as soon as $c \geq \frac{\alpha}{1-\alpha}$.

- $H_{\boldsymbol{X}} \leq H_{\boldsymbol{Y}} \leq H_{\boldsymbol{Z}}$: we have

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

  Then,

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

  as soon as $c \geq \frac{\alpha}{1-\alpha}$.

- $H_{\boldsymbol{X}} < H_{\boldsymbol{Z}} < H_{\boldsymbol{Y}}$: we have

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

  Then,

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

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

■

Remark 6 *The tool presented in Proposition 12 cannot be applied to the IB-divergence* $\Delta^{D,\alpha}$ *and the NIB-divergence* $\delta^{D,\alpha}$*. Indeed, let us give some* $c \geq 1$ *and let us consider the quantity*

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

In fact, one can always find $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ *such that for all* $c \geq 1$*, 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* $\alpha H_{\boldsymbol{Z}} + (1-\alpha) H_{\boldsymbol{Y}} = 3cH_{\boldsymbol{Y}}$*. Then, it is easy to see that* $A = H_{\boldsymbol{Z}}\left(\frac{1}{3} + \frac{1}{3} - 1\right) < 0$*. In the same manner, the tool is inapplicable to the IB-divergence* $\Delta^{P,\alpha}$ *and the NIB-divergence* $\delta^{P,\alpha}$*. Indeed, let us give* $\boldsymbol{Z}$ *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*

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

The following result is an extension of Proposition 12 well-suited to be applied to $\delta^{D,\alpha}$.

Proposition 14 *Let us assume that there exists two positive integer* I *and* J *such that a NIB-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 $\left(\alpha^{(j)}\right)_{j=1,\cdots,J}$ *is some vector of probability weights. By denoting* $S_{\boldsymbol{X},\boldsymbol{Y}} = \sum_{i=1}^{I} S^{(i)}_{\boldsymbol{X},\boldsymbol{Y}}$ *and* $U_{\boldsymbol{X},\boldsymbol{Y}} = \max_{i=1,\cdots,I} U^{(i)}_{\boldsymbol{X},\boldsymbol{Y}}$*, if there exists some real number* $c \geq 1$ *such that for any* $j = 1, \cdots, J$ *the following assumptions are satisfied:*

(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* ***[P6bis(***Γ^2***,c)]****.*

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

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

It follows that

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

■

Corollary 15 *The measure $\delta^{D,\alpha}$ satisfies* ***[P6bis(Γ^2,$\frac{1}{\alpha^\wedge}$)]****.*

Proof. We have

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

By identification with notation introduced in Proposition 14, we have $I = J = 2$, $S^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} = \alpha \min\left(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}}\right)$, $S^{(2)}_{\boldsymbol{X},\boldsymbol{Y}} = (1-\alpha) \max\left(H_{\boldsymbol{X}|\boldsymbol{Y}}, H_{\boldsymbol{Y}|\boldsymbol{X}}\right)$, $U^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} = \min\left(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}\right)$, $U^{(2)}_{\boldsymbol{X},\boldsymbol{Y}} = \max\left(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}\right)$, $C^{(1)}_{\boldsymbol{X},\boldsymbol{Y}} = \min\left(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}\right)$ and $C^{(2)}_{\boldsymbol{X},\boldsymbol{Y}} = \max\left(H_{\boldsymbol{X}}, H_{\boldsymbol{Y}}\right)$. Let us fix c to the value $\frac{1}{\alpha^\wedge}$. We have

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

Clearly from (29)

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

And one also has

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

It follows that $A^{(1)}$ fullfills conditions (i) and (ii) of Proposition 14 with $c = \frac{1}{\alpha^\wedge}$. The proof is strictly similar for $A^{(2)}$. ■

4 Prediction framework

We pay attention on properties related to the prediction of some fixed random vector $\boldsymbol{Y}$.

4.1 Prediction framework

Recall that our purpose is to find the random vector $\boldsymbol{X}$ that minimizes $\Delta_{\boldsymbol{Y},\boldsymbol{X}}$ (resp. $\delta_{\boldsymbol{Y},\boldsymbol{X}}$) 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 imagine that we already get some $\boldsymbol{X}_1$ and its associated measure $\Delta_{\boldsymbol{Y},\boldsymbol{X}_1}$ (resp. $\delta_{\boldsymbol{Y},\boldsymbol{X}_1}$). After evaluating $\Delta_{\boldsymbol{Y},\boldsymbol{X}_2}$ (resp. $\delta_{\boldsymbol{Y},\boldsymbol{X}_2}$), we may be interested in describing the conditions under which $\boldsymbol{X}_2$ is better or worse than $\boldsymbol{X}_1$:

Proposition 16 *Two situations may occur*
Case 1: we choose $\boldsymbol{X}_2$ instead of $\boldsymbol{X}_1$ when

$$\Delta_{\boldsymbol{Y},\boldsymbol{X}_2} < \Delta_{\boldsymbol{Y},\boldsymbol{X}_1} \Longleftrightarrow C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} < I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1} \tag{39}$$

$$\delta_{\boldsymbol{Y},\boldsymbol{X}_2} < \delta_{\boldsymbol{Y},\boldsymbol{X}_1} \Longleftrightarrow \frac{C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1}}{C_{\boldsymbol{Y},\boldsymbol{X}_1}} < \frac{I_{Y,\boldsymbol{X}_2} - I_{Y,\boldsymbol{X}_1}}{I_{Y,\boldsymbol{X}_1}} \tag{40}$$

Case 2: we keep $\boldsymbol{X}_1$ and reject $\boldsymbol{X}_2$ when

$$\Delta_{\boldsymbol{Y},\boldsymbol{X}_2} \geq \Delta_{\boldsymbol{Y},\boldsymbol{X}_1} \Longleftrightarrow C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} \geq I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1} \tag{41}$$

$$\delta_{\boldsymbol{Y},\boldsymbol{X}_2} \geq \delta_{\boldsymbol{Y},\boldsymbol{X}_1} \Longleftrightarrow \frac{C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1}}{C_{\boldsymbol{Y},\boldsymbol{X}_1}} \geq \frac{I_{Y,\boldsymbol{X}_2} - I_{Y,\boldsymbol{X}_1}}{I_{Y,\boldsymbol{X}_1}} \tag{42}$$

This result implies automatically that the properties **[P8]** and **[P9]** are satisfied. Let us comment more precisely 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 (resp. relative) excess of complexity is lower than the absolute (resp. relative) gain of

information that is $C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} < I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1} < 0$ (resp. $\frac{C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1}}{C_{\boldsymbol{Y},\boldsymbol{X}_1}} < \frac{I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}}{I_{\boldsymbol{Y},\boldsymbol{X}_1}} < 0$).

4. $\boldsymbol{X}_2$ is more complex and more informative than $\boldsymbol{X}_1$ and such that the absolute (resp. relative) excess of complexity is lower than the absolute (resp. relative) gain of information that is $0 < C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1} < I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}$ (resp. $0 < \frac{C_{\boldsymbol{Y},\boldsymbol{X}_2} - C_{\boldsymbol{Y},\boldsymbol{X}_1}}{C_{\boldsymbol{Y},\boldsymbol{X}_1}} < \frac{I_{\boldsymbol{Y},\boldsymbol{X}_2} - I_{\boldsymbol{Y},\boldsymbol{X}_1}}{I_{\boldsymbol{Y},\boldsymbol{X}_1}}$).

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

Proposition 17 *Any complexity term C^α of the form (17) satisfies **[P10]**.*

Proof. Without loss of generality the function $g(\cdot)$ defining C^α is assumed to be an increasing function. Hence, $H_{\boldsymbol{X}_2} \geq H_{\boldsymbol{X}_1}$ implies that $C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_2} \geq C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_1}$. Now, let us assume $C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_2} \geq C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_1}$. We assert by denoting $m_i = \min(H_{\boldsymbol{Y}}, H_{\boldsymbol{X}_i})$ and $M_i = \max(H_{\boldsymbol{Y}}, H_{\boldsymbol{X}_i})$ for $i = 1, 2$

$$\begin{aligned} C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_2} \geq C^\alpha_{\boldsymbol{Y},\boldsymbol{X}_1} &\Longleftrightarrow g^{-1}\left(\alpha g(m_1) + (1-\alpha) g(M_1)\right) \leq g^{-1}\left(\alpha g(m_2) + (1-\alpha) g(M_2)\right) \\ &\Longleftrightarrow \alpha\left(g(m_1) - g(m_2)\right) + (1-\alpha)\left(g(M_1) - g(M_2)\right) \leq 0 \end{aligned}$$

Now, assume moreover that $H_{\boldsymbol{X}_1} > H_{\boldsymbol{X}_2}$, then the right-hand side is

$$\begin{cases} = (1-\alpha)(g(H_{\boldsymbol{X}_1} - g(H_{\boldsymbol{X}_2})) > 0 & \text{if } H_{\boldsymbol{Y}} \leq H_{\boldsymbol{X}_2} < H_{\boldsymbol{X}_1} \\ > g(m_1) - g(m_2) = 0 & \text{if } H_{\boldsymbol{X}_2} < H_{\boldsymbol{Y}} < H_{\boldsymbol{X}_1} \\ = \alpha(g(H_{\boldsymbol{X}_2}) - g(H_{\boldsymbol{X}_1})) > 0 & \text{if } H_{\boldsymbol{X}_2} < H_{\boldsymbol{X}_1} \leq H_{\boldsymbol{Y}} \end{cases}.$$

This leads to a contradiction which implies that $H_{\boldsymbol{X}_2} \geq H_{\boldsymbol{X}_1}$. ■

Remark 7 *The complexity terms C^E and C^I do not satisfy the 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, one can 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]** which is satisfied from Proposition 17 and Remark 7 (in this framework)

  $C^\bullet$ 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 not trivial since both absolute (resp. relative) excess of complexity and absolute (resp. 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. variables selection problem: suppose 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 so 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 Robineau (2004) 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 IB-divergence or NIB-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 IB-divergence or NIB-divergence, then both have the same effect on the prediction of another vector $\boldsymbol{Y}$. This requirement, expressed by the property **[P11]**, could be used for example in a variables selection problem in the context of discrimination to detect redundant variables.

In order to make the property **[P11]** applicable for practical purpose, we may find interesting to have a bound of the difference $|\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}|$ (resp. $|\delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \delta_{\boldsymbol{Y},\boldsymbol{X}_2}|$) expressed in terms of $D^I_{\boldsymbol{X}_1,\boldsymbol{X}_2}$ (resp. $d^I_{\boldsymbol{X}_1,\boldsymbol{X}_2}$). 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 $|\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}| \leq h(D^I_{\boldsymbol{X}_1,\boldsymbol{X}_2})$ (resp. $|\delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \delta_{\boldsymbol{Y},\boldsymbol{X}_2}| \leq h(d^I_{\boldsymbol{X}_1,\boldsymbol{X}_2})$). Here, according to our examples, we only concentrate ourself on linear function $h(\cdot)$.

We then propose to translate the property **[P11]** on an IB-divergence Δ(resp. a NIB-divergence δ) by:

[P11bis(Υ, k)] there exists some positive constant k such that for all $(\boldsymbol{X}_1, \boldsymbol{X}_2) \in \Upsilon \subset \Gamma^2$ such that

$$|\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}| \leq k\, D^I_{\boldsymbol{X}_1,\boldsymbol{X}_2} \tag{43}$$

As a first answer, let us precise that if the IB-divergence (resp. NIB-divergence) satisfies a triangular inequality **[P6bis($\Gamma^2, 1$)]** and **[P3bis(Υ, k_1, k_2)]** then it satisfies **[P11bis(Υ, k_2)]** due to the equivalent expression of the triangular inequality as

$$|D_{\boldsymbol{Y},\boldsymbol{X}_1} - D_{\boldsymbol{Y},\boldsymbol{X}_2}| \leq D_{\boldsymbol{X}_1,\boldsymbol{X}_2} \quad \text{(resp.} \quad |d_{\boldsymbol{Y},\boldsymbol{X}_1} - d_{\boldsymbol{Y},\boldsymbol{X}_2}| \leq d_{\boldsymbol{X}_1,\boldsymbol{X}_2}).$$

A priori, if an IB-divergence or NIB-divergence only satisfies **[P6bis(Γ^2, 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 IB-divergence by instance that

$$|\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}| \leq c \times \Delta_{\boldsymbol{X}_1,\boldsymbol{X}_2} + (c-1)\min\left(\Delta_{\boldsymbol{Y},\boldsymbol{X}_1}, \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}\right) \nleq c \times \Delta_{\boldsymbol{X}_1,\boldsymbol{X}_2}.$$

Actually, this apparent disappointing result only expresses that a "redundancy" property cannot (always) be derived from a triangular's type inequality.

The following proposition gives some sufficient conditions required on some complexity term ensuring that the associated $\Delta$and δsatisfies the property **[P11bis]**

Proposition 18 (i) *Assume there exists some positive constant* κ_1 *such that the complexity term of an IB-divergence satisfies for all* $(\boldsymbol{X}_1, \boldsymbol{X}_2) \in \Upsilon$

$$\left|C_{\boldsymbol{Y},\boldsymbol{X}_1} - C_{\boldsymbol{Y},\boldsymbol{X}_2}\right| \leq \kappa_1 \left|H_{\boldsymbol{X}_1} - H_{\boldsymbol{X}_2}\right|, \tag{44}$$

then Δ*satisfies [**P11bis**$(\Upsilon, 1+\kappa_1)$]*
(ii) *If in addition, there exists some positive constant* κ_2 *such that for all* $(\boldsymbol{X}_1, \boldsymbol{X}_2) \in \Upsilon$

$$\max\left(C_{\boldsymbol{Y},\boldsymbol{X}_1}, C_{\boldsymbol{Y},\boldsymbol{X}_2}\right) \geq \kappa_2 \times C^I_{\boldsymbol{X}_1,\boldsymbol{X}_2} \tag{45}$$

*then the associated NIB-divergence satisfies [**P11bis**$\left(\Upsilon, \frac{1+\kappa_1}{\kappa_2}\right)$]*

Proof. (i) Let us start to write

$$\left|\Delta_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta_{\boldsymbol{Y},\boldsymbol{X}_2}\right| \leq \left|I_{\boldsymbol{Y},\boldsymbol{X}_1} - I_{\boldsymbol{Y},\boldsymbol{X}_2}\right| + \left|C_{\boldsymbol{Y},\boldsymbol{X}_1} - C_{\boldsymbol{Y},\boldsymbol{X}_2}\right|. \tag{46}$$

Now, notice 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

$$\left|I_{\boldsymbol{Y},\boldsymbol{X}_1} - I_{\boldsymbol{Y},\boldsymbol{X}_2}\right| \leq \max\left(H_{\boldsymbol{X}_1}, H_{\boldsymbol{X}_2}\right) - I_{\boldsymbol{X}_1,\boldsymbol{X}_2} = \max\left(H_{\boldsymbol{X}_1|\boldsymbol{X}_2}, H_{\boldsymbol{X}_2|\boldsymbol{X}_1}\right) = D^I_{\boldsymbol{X}_1,\boldsymbol{X}_2}. \tag{47}$$

The result is then obtained by combining (44), (46) and (47).

(ii) We can obtain the following result

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

The result then comes from (44), (45) and (47). ■

Let us apply the previous result to our different examples:

Corollary 19 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \Gamma_\Theta$ with $\Theta = [c_1, c_2]$ and define γ_i $(i = 1, 2)$ such that $c_i = \gamma_i H_{\boldsymbol{Y}}$, then*

$$\left|\Delta^\bullet_{\boldsymbol{Y},\boldsymbol{X}_1} - \Delta^\bullet_{\boldsymbol{Y},\boldsymbol{X}_2}\right| \leq \left(1 + \kappa^\bullet_{1,\Theta}\right) D^I_{\boldsymbol{X}_1,\boldsymbol{X}_2} \quad \text{and} \quad \left|\delta^\bullet_{\boldsymbol{Y},\boldsymbol{X}_1} - \delta^\bullet_{\boldsymbol{Y},\boldsymbol{X}_2}\right| \leq \frac{1 + \kappa^\bullet_{1,\Theta}}{\kappa^\bullet_{2,\Theta}} d^I_{\boldsymbol{X}_1,\boldsymbol{X}_2} \quad (48)$$

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

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

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

Proof. For the sake of simplicity, let us denote by $m_i = \min\left(H_{\boldsymbol{Y}}, H_{\boldsymbol{X}_i}\right)$ (resp. $m = \min\left(H_{\boldsymbol{Y}}, H_{\boldsymbol{X}}\right)$) and by $M_i = \max\left(H_{\boldsymbol{Y}}, H_{\boldsymbol{X}_i}\right)$ for $i = 1, 2$ (resp. $M = \max\left(H_{\boldsymbol{Y}}, H_{\boldsymbol{X}}\right)$). Let us notice on the one hand that $|M_1 + m_1 - (M_2 + m_2)| = |H_{\boldsymbol{X}_1} - H_{\boldsymbol{X}_2}|$ and on the other hand that

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

- Complexity term C^S : we have

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

  Moreover,

$$C^{S,\alpha}_{\boldsymbol{Y},\boldsymbol{X}} = \alpha m + (1-\alpha)M \geq \left((1-\alpha) + \alpha\gamma_{1,2}\right) M$$

- Complexity term C^R : we have

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

Furthermore, we may obtain

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

and

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

Hence,

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

Moreover, one can prove

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

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

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

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

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

Moreover, we may obtain

$$C^{P,\alpha}_{\boldsymbol{Y},\boldsymbol{X}} = m^\alpha M^{1-\alpha} \geq \gamma_{1,2}^\alpha M$$

- Complexity term C^D : we have

$$|C^{D,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_1} - C^{D,\alpha}_{\boldsymbol{Y},\boldsymbol{X}_2}| = \frac{\alpha M_1 M_2(m_1 - m_2) + (1-\alpha)m_1 m_2(M_1 - M_2)}{(\alpha M_1 + (1-\alpha)m_1)(\alpha M_2 + (1-\alpha)m_2)} \tag{49}$$

$$\leq \frac{\alpha^\vee}{(\alpha^\wedge)^2} \frac{M_1 M_2}{(m_1 + M_1)(m_2 + M_2)} \left| H_{\boldsymbol{X}_2} - H_{\boldsymbol{X}_1} \right| \tag{50}$$

$$\leq \frac{\alpha^\vee}{(\alpha^\wedge)^2} \frac{1}{(1+\gamma_{1,2})^2} \left| H_{\boldsymbol{X}_2} - H_{\boldsymbol{X}_1} \right| \tag{51}$$

Finally, we also have

$$C^{D,\alpha}_{\boldsymbol{Y},\boldsymbol{X}} = \left(\frac{\alpha}{m} + \frac{1-\alpha}{M} \right)^{-1} \geq \left(\frac{\alpha}{\gamma_{1,2}} + (1-\alpha) \right)^{-1} M.$$

■

Remark 8 *Note that when $\alpha \leq \frac{1}{2}$, the measure $\Delta^{S,\alpha}$ is a metric and so we derive (48) directly from **[P3bis]**.*

References

C. H. Bennett, P. Gács, M. Li, P. M. B. Vitányi, and W. H. Zurek. Information distance. *IEEE Trans. on Info. Theory*, 44(4):1407–1423, 1998.

R. Cilibrasi and P. Vitányi. Automatic meaning discovery using google, 2005a. E-print arxiv.org/abs/

R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *IEEE Trans. on Info. Theory*, 51(4):1523–1545, 2005b.

R. Cilibrasi, P. M. B. Vitányi, and R. de Wolf. Algorithmic clustering of music. *Computing Research Repository*, 2003. Eprint arxiv.org/abs/cs.SD/0303025.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA, 1991.

J. P. Crutchfield. Information and its metric. In L. Lam and H. C. Morris, editors, *Nonlinear Structures in Physical Systems – Pattern Formation, Chaos and Waves*, pages 119–130. Springer Verlag, 1990.

C.W. Granger, E. Maasoumi, and J. Racine. A dependence metric for possibly nonlinear processes. *J. Time Ser. Anal.*, 25(5):649–669, 2004.

P. Grünwald and P. M. B. Vitányi. Shannon information and kolmogorov complexity. *Computing Research Repository*, 2004. Eprint arxiv.org/abs/cs.IT/0410002.

D. Hammer, A. E. Romashchenko, A. Shen, and N. K. Vereshchagin. Inequalities for shannon entropy and kolmogorov complexity. *J. Comput. Syst. Sci*, 60(2):442–464, 2000.

C. Hillman. A formal theory of information: I. statics, February 19 1998. URL `http://citeseer.ist.psu.edu/89520.html`.

A. Kaltchenko. Algorithms for estimation of information distance with application to bioinformatics and linguistics. In *Proceedings of the 2004 Canadian Conference on Electrical and Computer Engineering*, volume 4, page 2255, 2004.

A. Kraskov, H. Stögbauer, R.G. Andrzejak, and P. Grassberger. Hierarchical clustering based on mutual information, 2003. E-print arxiv.org/abs/q-bio/0311039.

S. K. Leung-Yan-Cheong and T. M. Cover. Some equivalences between shannon entropy and kolmogorov complexity. *IEEE Trans. on Info. Theory*, 24(3):331–338, 1978.

M. Li, J. H. Badger, X. Chen, S. Kwong, P. E. Kearney, and H. Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17(1):149–154, 2001.

M. Li, X. Chen, X. Li a. B. Ma, and P. M.B. Vitanyi. The similarity metric. In *Proceedings of the fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-03)*, pages 863–872, New York, January 12–14 2003. ACM Press.

M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi. The similarity metric. *IEEE Trans. on Info. Theory*, 50(12):3250–3264, 2004.

M. Li and P. Vitányi. *An introduction to Kolmogorov complexity and its applications.* Springer, New York, 2nd edition, 1997.

H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining.* Kluwer Academic Publishers, 1998.

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

C. E. Shannon. A mathematical theory of communication (continued). *The Bell System Technical Journal*, 27(4):623–656, October 1948. ISSN 0005-8580.

A. Ullah. Entropy, divergence and distance measures with econometric applications. *J. Stat. Plan. Infer.*, 49:137–162, 1996.

Authors: Jean-François Coeurjolly, Rémy Drouilhet and Jean-François Robineau.
Address: LABSAD, BSHM, 1251 avenue centrale BP 47 - 38040 GRENOBLE Cedex 09.
E-mail addresses:

`Jean-Francois.Coeurjolly@upmf-grenoble.fr`

`Remy.Drouilhet@upmf-grenoble.fr`

Corresponding author: Jean-François Coeujolly.