information

MDPI

Article

A Maximum Value for the Kullback–Leibler Divergence between Quantized Distributions

Vincenzo Bonnici

Department of Mathematical, Physical and Computer Sciences, University of Parma, Parco Area delle Scienze 53/A (Campus Scienze e Tecnologie), 43124 Parma, Italy; vincenzo.bonnici@unipr.it

Abstract: The Kullback–Leibler (KL) divergence is a widely used measure for comparing probability distributions, but it faces limitations such as its unbounded nature and the lack of comparability between distributions with different quantum values (the discrete unit of probability). This study addresses these challenges by introducing the concept of quantized distributions, which are probability distributions formed by distributing a given discrete quantity or *quantum*. This study establishes an upper bound for the KL divergence between two quantized distributions, enabling the development of a normalized KL divergence that ranges between 0 and 1. The theoretical findings are supported by empirical evaluations, demonstrating the distinct behavior of the normalized KL divergence compared to other commonly used measures. The results highlight the importance of considering the quantum value when applying the KL divergence, offering insights for future advancements in divergence measures.

Keywords: Kullback–Leibler divergence; entropic divergence; statistical distance; quantized distribution; probability distributions

Citation: Bonnici, V. A Maximum Value for the Kullback–Leibler Divergence between Quantized Distributions. *Information* **2024**, *15*, 547. https://doi.org/10.3390/info15090547

Academic Editor: Gabriel Luque

Received: 6 August 2024
Revised: 21 August 2024
Accepted: 2 September 2024
Published: 6 September 2024

1. Introduction

The Kullback–Leibler divergence (KL), also called entropic divergence, is a widely used measure for comparing two discrete probability distributions [1]. Such a divergence is derived from the notion of entropy, and it aims at evaluating the amount of information that is gained by switching from one distribution to another. The applications of the divergence span several scientific areas, for example, for testing random variables [2–4], for selecting the right sample size [5], for optimizing sampling in bioinformatics [6] or for analyzing magnetic resonance images [7]. However, the entropic divergence has two important properties that limit its applicability. It has also been applied as a cost function in predictive machine-learning approaches [8] based on the well-established random-forest model, or in artificial neural networks [9], for example, for clustering data points [10] or for generative models [11]. It can not be used as a metric because it is not symmetric, i.e., $KL(P||Q) \neq KL(Q||P)$, where P and Q are two probability distributions. Moreover, its value is 0 if equal distributions are compared, but it is shown not to have an upper bound to its possible value. One of the reasons is that it results in an infinite divergence if the probability of a specific event is equal to 0 in Q but is greater than 0 in P. Although infinite divergence is discarded, an upper bound for the entropic divergence has not been established.

The search for bounded divergences is an important topic in information theory, and some attempts have been made in the past few years. For example, the main goal of the so-called Jensen–Shannon divergence (JS) [12] is to provide a notion of symmetric divergence, but it is also shown to be upper-bounded by the value 1 if the base of the used logarithm is 2. It is a metric but its values are not uniformly distributed within the range $[0 \ldots 1]$, as is empirically shown in this study. Kullback–Leibler and Jensen–Shannon measures are in the class of f-divergences [13], which aim at representing the divergence as an average of the

odds ratioweighted by a function f. Each divergence has a specific meaning and behavior, and the relation among different types of f-divergence is a well-studied topic [14]. The Hellinger distance [15] is one of the most used measures among the f-divergences, together with Kl and JS. It avoids infinite divergences by definition, and it is bounded between 0 and 1. More generally, the KL divergence is shown to be related to several other types of divergences [16].

The present work introduces a new class of discrete probability distributions called quantized distributions. The name comes from the fact that the probabilities reported by such a class of distributions are formed by quanta of probability. The final aim of the present study is to show that, given a quantized distribution P, there exists another quantized distribution U that maximizes the entropic divergence from P. Thus, for each distribution P, an upper bound to the divergence form P can be obtained by constructing U. The assumptions are that infinite divergences must be avoidable and that the two distributions must be formed by distributing a given discrete quantity, namely, the same quantum must form them. This last property highlights an important previously unaccounted aspect of the KL divergence. Because such a bound can only be assessed under this condition, KL divergence should only be applied between probability distributions formed by the same quantum. These theoretical results allow the introduction of a notion of entropic divergence that is normalized in the range $[0 \dots 1]$, independently from the base of the used logarithm. Such a measure is compared with the more commonly used notions of divergence and distance between distributions by showing that it behaves in a precise way. Furthermore, it is empirically shown that its values are better distributed in the range $[0 \dots 1]$ with respect to the compared measures. In conclusion, the novel aspects of this study are (i) the introduction of the concept of quantized distributions, (ii) the establishment of an upper bound for the KL divergence between quantized distributions, and (iii) the proposal of a normalized KL divergence.

2. Preliminaries

A *finite* (thus discrete) multiplicity distribution is defined as a function f, which distributes a given discrete quantity M to a finite set C of $|C|$ distinct cells. Thus, $\sum_{c \in C} f(c) = M$. This class of distributions is often represented via Ferrers diagrams [17], in which the distributed quantity is a finite set of M dots that are assigned to cells. A multiplicity distribution is commonly transformed into discrete probability (frequency) distributions by converting it to a distribution such that the sum of its outcomes equals 1. Thus, a finite discrete probability distribution P is obtained by dividing the assigned quantities for the total quantity, namely $P(c \in C) = \frac{f(c)}{M}$. In this context, the term *quantum* signifies that the distribution is defined on a finite, discrete domain, and the assigned values are composed of quanta, which are discrete, unitary pieces of information.

Definition 1 (Quantized distribution)**.** *A quantized (probability) distribution (QD) is a finite discrete probability distribution that assigns a probability value to each of the n values of a variable. The probability values are positive, non-zero multiples of the fraction* $1/M$*, called the quantum of the distribution, for a given* $M \in \mathbb{N}$*. The value n is also called the cardinality of the distribution.*

It has to be noticed that since quantized distributions are probability distributions, the sum of the assigned probabilities must equal 1. Furthermore, this defines a special type of probability distribution. In fact, in general, it is not required that a probability distribution is sourced by a discrete quantity M distributed over a finite set of cells. Such a type of distribution is of great importance in the field of Computer Science, where probabilities are estimated by looking at frequencies calculated from discrete quantities, for example, for representing biological information [18–20]. However, it can be easily shown that the class of quantized distribution covers all the possible discrete finite probability distributions. For distributions where assigned probabilities are rational numbers, rescaling is always possible. This is achieved by setting the quantum value as 1 divided by the common denominator of

the assigned probabilities. The rest of the discrete finite probability distributions can be approximated by using an arbitrarily small epsilon for their discretization.

Remark 1. *For each multiplicity distribution D, there exists a quantized distribution P, and vice versa.*

In fact, given a multiplicity distribution, it can always be converted to a quantized distribution by dividing the assigned values by their sum. Vice versa, a quantized distribution can be represented as a function that assigns values that are an integer multiple (a multiplicity) of the quantum.

Because of the strict relation between frequency/probability and multiplicity distributions, from now on and without loss of generality, the assigned probability values, $\frac{f(c)}{M}$, will be interchanged with their multiplicity/integer-quantity counterpart, $f(c)$, depending on the purpose of the context in which they are recalled. Similarly, Ferret diagrams and their dot-based representation represent this type of distribution.

Remark 2. *Two distributions P and Q, defined on the same domain C, are considered equal, thus not distinct if $\forall c \in C \Rightarrow P(c) = Q(c)$.*

Proposition 1. *The total number of distinct, thus not equal, quantized distributions that can be formed by arranging a quantity M in n distinct cells is $\binom{M-1}{M-n}$.*

Proof. Given a set S of x elements, the number of y-combinations, namely the number of subsets of y distinct elements of S, if given by $\binom{x}{y}$. The number of y-combinations with repetitions, namely the number of sequences of y non-necessarily distinct elements of S, is given by $\binom{x+y-1}{y}$ [21].

Quantized distributions require that at each cell, a minimum value of $1/M$ must be assigned. Switching from quantized to multiplicity distributions implies that a quantity of n elements, out of M, does not participate in the arrangement process since a fixed minimum value of 1 is assigned to each cell. Thus, the quantity that is arranged equals $M - n$. Each dot must be assigned to a given cell, and no dot can remain unassigned. Thus, the arrangement process can be seen as an assignment of one specific cell to each of the $M - n$ dots by allowing a cell to be assigned to multiple dots. Compared to classical combinatorial problems, we are not assigning dots to cells but cells to dots. Thus, this means counting the number of $(M - n)$-combinations with repetitions of a set of n elements, which is given by

$$\binom{M-n+n-1}{M-n} = \binom{M-1}{M-n} \tag{1}$$

□

The present study aims to show that for each of these distributions, there exists another distribution that maximizes the value of the entropic divergence. The proof of it, which is given in the next section, requires that the elements of the domain must be ordered according to the values assigned to them.

Definition 2 (Ordered quantized distribution)**.** *Given a quantized distribution P, an ordered quantized distribution (OQD) is obtained by assigning an integer index i, with $1 \leq i \leq |C|$, to each domain element $c \in C$ such that $P(c_i) \geq P(c_{i+1})$. $P(c_i)$ is also referred to as P_i.*

It must be noted that Definition 2 is based on a monotonically decreasing order, but without loss of generality; a monotonically increasing order can also be used. Furthermore, in what follows, the greatest value of the distribution is considered to be placed in the leftmost position. Consequently, the lowest value is considered to be placed in the rightmost position of the distribution.

Remark 3. *Two ordered distributions P and Q, defined on the same domain C, are equal, thus not distinct, if* $\forall i : 1 \leq i \leq |C| \Rightarrow P_i = Q_i$.

Multiple unordered distributions may produce the same ordered distribution leading them to belong to the same class of equivalence that is defined by such a shared ordered output. Formally, $\mathbb{Q}_{M,n}$ is defined as the complete set of QDs that can be formed by arranging a quantity of M into n cells. $\mathbb{O}_{M,n}$ is defined as the complete set of OQDs that can be formed by arranging a quantity of M into n cells. Then, the function that transforms an unordered QD into an ordered QD, $ord : \mathbb{Q}_{M,n} \mapsto \mathbb{O}_{M,n}$, can be shown to be a surjective function. Thus, each class of equivalence is represented by a given distribution $O \in \mathbb{O}_{M,n}$, and it is formed by a subset of $\mathbb{Q}_{M,n}$, referred to as $\mathbb{Q}^O_{M,n}$, such that $\forall P \in \mathbb{Q}^O_{M,n} : ord(P) = O$.

We are interested in counting the number of classes, which also equals the number of distinct ordered distributions.

Proposition 2. *The total number of distinct ordered distributions that can be formed by arranging a quantity of M in n cells equals the number of partitions of the integer M for representing it as a sum of n integer addends.*

Proof. Similarly to unordered distributions (see Proposition 1), ordered distributions assign a minimum quantity of 1 to each cell. The number of partitions of an integer x to represent it as a sum of y addends, the value of which can not be 0, can be obtained by the recursive formula $p_y(x) = p_y(x-y) + p_{y-1}(x-1)$, with $p_y(x) = 0$ if $y > x$ and $p_y(0) = 0$ [22]. Thus, we can use the formula to evaluate the number of ordered distributions by setting $x = M$ because it is the total arranged quantity, and $y = n$ because we want to represent such an integer as a sum of exactly n non-zero addends (namely, non-empty cells). □

The search for a maximum value of the KL divergence between two quantized distributions P and Q (presented in the next section) is based on the fact that the same quantum value must form both distributions. However, there are plenty of practical situations where this assumption is not verified, and the two distributions need to be transformed into two *comparable* distributions before calculating the divergence.

Proposition 3. *Given two quantized distributions, P and Q, formed by two different quanta,* $1/M_P$ *and* $1/M_Q$*, respectively, they can always be transformed into two distributions formed by the same quantum.*

Proof. Since M_P and M_Q are two positive integer numbers, the least common multiple (lcm) between them can be used for re-scaling the two distributions such that the same quantum value will form them. The new distributions are formed by the same quantum, that is $M = 1/lcm(M_P, M_Q)$. The values of these distributions are always in the form x/M, with x being a positive integer. Thus, the values of the new distributions can be re-scaled as $x(M/M_Q)/M$. It is trivial to show that the new distributions maintain their status of quantized distribution. □

3. Upper Bound of the Entropic Divergence

Given two probability distributions, the entropic divergence, also called the Kullback–Leibler (KL) divergence from the authors who discovered it [1], quantifies the information gained by switching from one distribution to another. For two probability distributions, P and Q, that are defined on the same domain C, the divergence of P from Q is defined as:

$$KL(P||Q) = \sum_{c \in C} p(c) log_2 \frac{P(c)}{Q(c)} \tag{2}$$

The divergence is not symmetric, thus $KL(P||Q) \neq KL(Q||P)$, and its possible value ranges between 0 and $+\infty$. In fact, the divergence is 0 if the two distributions are equal in their outcomes, namely $P(c) = Q(c), \forall c \in C$. Gibbs' inequality [23] demonstrates that it has no

upper bound. However, such an affirmation has been shown by comparing two *general* distributions and by stating that the entropic divergence is a difference between the two quantities $-\sum_{c\in C} P(c)log_2P(c)$ and $-\sum_{c\in C} P(c)log_2Q(c)$, which implies

$$-\sum_{c\in C} P(c)log_2P(c) \leq -\sum_{c\in C} P(c)log_2Q(c) \tag{3}$$

and thus

$$KL(P||Q) = \sum_{c\in C} P(c)log_2\frac{P(c)}{Q(c)} \geq 0 \tag{4}$$

Given two positive numbers M and n, the previous section establishes that the sets $\mathbb{Q}_{M,n}$ and $\mathbb{O}_{M,n}$ are finite. Consequently, for any P within either of these sets, there exists a distribution U in $\mathbb{Q}_{M,n}$ or $\mathbb{O}_{M,n}$ that maximizes $KL(P||U)$. Here, we are interested in finding such a distribution U. It must be noted that P and U are quantized distributions formed by the same quantum. This assumption is crucial for obtaining an upper bound on the divergence from a given distribution P in practical situations.

The general concept of distribution, and thus of probability distribution, is independent of a given ordering of the elements in C. In this perspective, ordered quantized distributions are used without loss of generality. The KL formula for ordered distributions can be written as:

$$KL(P||Q) = \sum_{1\leq i\leq n} P_i log_2\frac{P_i}{Q_i} \tag{5}$$

It must be pointed out that the ordering does not affect the value of the KL divergence. This means that the distribution that maximizes the KL value from a given distribution P also maximizes the KL for all the unordered distributions within the same class of equivalence of P defined in the previous section. Thus, the goal is to define the shape of the distribution U, which maximizes the entropic divergence to P.

It is required that the compared distributions, P and U, must be defined on the set C and that for each element, the two distributions are non-zero valued, namely $P_i > 0$ and $U_i > 0$ for $1 \leq i \leq n$, in order to avoid infinite divergences. This constraint, together with the discretization of the quantity that is distributed to the cells, implies that at each cell, at least a quantity equal to 1 is assigned, that is, $P_i, U_i \geq 1/M$ for every i. Thus, the quantity that must be arranged to construct the distribution U is M^-n.

The entropic divergence is a sum of terms in the form $P_i log_2(P_i/U_i)$. If $P_i < U_i$, then a negative contribution is given to the sum because of the logarithmic function, while positive contributions are given for $P_i \geq U_i$. Thus, the aim is to reduce the number of positions with negative contributions. Each term is mediated by the P_i factor. Thus, it is preferable to assign positive contributions to the greatest P_i values. On the contrary, negative contributions should be assigned to the smallest P_i values. This means that if P is monotonically decreasing in order (from left to right), then positive contributions should be on the left side of the distributions, and negative terms should be on the right side. Furthermore, the greater P_i with respect to U_i, the higher the value of the divergence. This translates to trying to increase the difference between the greatest P_i values and their corresponding U_i counterparts as much as possible. Of course, reducing the quantity assigned to the initial positions of U results in increasing the quantity assigned to the right-most positions of it.

All of these considerations lead to the intuition that the distribution that maximizes the entropic divergence is the one that minimizes the quantity assigned to positions from 1 to $n-1$ and that assigns the remaining amount to the last position. Since the minimum quantity is equal to 1, such a distribution assigns the remaining $M-n+1$ quantity to the last position n. In what follows, it is shown that if P is monotonically decreasing ordered, then such a distributional shape maximizes the entropic divergence independently of how the quantity is distributed in P. This fact also implies that such maximization is independent of the ordering of P. It is only necessary that the quantity $M-n+1$ is

assigned to the position i, rather than n, where P_i is minimal. However, the ordering helps prove the initial statement.

From here on, the maximizing distribution is always referred to as U, and any other competitor distribution is referred to as Q. The proof that the entropic divergence from U to P is greater than the divergence from any other distribution Q is split into two parts. Firstly, a special case is addressed, then the proof of the general case is given.

The special case is presented in Figure 1. A total amount of $M = 11$ elements are arranged into $n = 5$ cells to compose the distributions. As introduced above, the P distribution has a monotonically decreasing order, and the U distributions assign a quantity of $M - n + 1$ to the last cell. The special case is represented by the Q distribution, which assigns a quantity of 2 to the $(n-1)$-th position and a quantity of $M - n$ to the last position. For all the distributions, for every cell, a minimal quantity of 1 is assigned. The goal is to show that:

$$KL(P||U) > KL(P||Q) \tag{6}$$

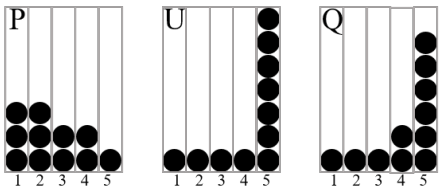

Figure 1. First special case. Each element is represented as a dot that is assigned to one of the cells. A total of 11 elements are assigned to a total of 5 cells for each of the three distributions, P, U and Q, that are present in the case.

From cell 1 to cell 3, the two divergences have an equal contribution; thus, they differ in terms of the last two terms. Thus, the inequality can be written as:

$$P_4 log_2 \frac{P_4}{U_4} + P_5 log_2 \frac{P_5}{U_5} > P_4 log_2 \frac{P_4}{Q_4} + P_5 log_2 \frac{P_5}{Q_5} \tag{7}$$

that is

$$\begin{array}{c} P_4 log_2 P_4 - P_4 log_2 U_4 + P_5 log_2 P_5 - P_5 log_2 U_5 \\ > \\ P_4 log_2 P_4 - P_4 log_2 Q_4 + P_5 log_2 P_5 - P_5 log_2 Q_5 \end{array} \tag{8}$$

that is

$$\begin{array}{r} -P_4 log_2 U_4 - P_5 log_2 U_5 > -P_4 log_2 Q_4 - P_5 log_2 Q_5 \\ -\frac{2}{11} log_2 \frac{1}{11} - \frac{1}{11} log_2 \frac{5}{11} > -\frac{2}{11} log_2 \frac{2}{11} - \frac{1}{11} log_2 \frac{4}{11} \\ -\frac{1}{11} log_2 5 > -\frac{2}{11} log_2 2 - \frac{1}{11} log_2 4 \\ -log_2 5 > -2 log_2 2 - log_2 4 \\ -log_2 5 > -4 \end{array} \tag{9}$$

which is true.

A general proof of this special case, independently from the values of M and n, is given in Appendix A.

Moving forward, the final goal is to show that U maximizes the divergence with respect to any possible distribution Q obtained by arranging the $M - n$ quantity in all the cells.

Proposition 4. *Let P be an OQD obtained by distributing a quantity M to n cells. Let U be a QD, which assigns all the free quantity $M - n$ to the n-th cell and the minimum quantity of* 1 *to*

each cell. Let Q be a QD that assigns the free quantity in a way different from U, in addition to the minimum quantity of 1 *for each cell. Then,* $KL(P||U) > KL(P||Q)$*, independent of how the quantity is arranged in Q.*

Proof. The initial $n-1$ cells of U have a value equal to $1/M$, and the last cell has a value equal to $\frac{M-n+1}{M}$. Instead, for what concerns Q, a quantity equal to $1+x_i$, for $x_i \geq 0$, is assigned to each cell, such that $\sum_{1\leq i\leq n} x_i = M-n$.

The following inequality must be verified:

$$\left(\sum_{1\leq i\leq n-1} P_i log_2 \frac{P_i}{\frac{1}{M}}\right) + P_n log_2 \frac{P_n}{\frac{M-n+1}{M}} > \sum_{1\leq i\leq n} P_i log_2 \frac{P_i}{\frac{1+x_i}{M}} \tag{10}$$

that is

$$\begin{gathered}\left(\sum_{1\leq i\leq n-1} P_i log_2 \frac{P_i}{\frac{1}{M}}\right) + P_n log_2 \frac{P_n}{\frac{M-n+1}{M}} \\ > \\ \left(\sum_{1\leq i\leq n-1} P_i log_2 \frac{P_i}{\frac{1+x_i}{M}}\right) + P_n log_2 \frac{P_n}{\frac{1+x_n}{M}}\end{gathered} \tag{11}$$

The left side of the inequality is composed of a series of terms $P_i log_2 \frac{P_i}{\frac{1}{M}}$, each of which equals $P_i log_2 P_i - P_i log_2 1 + P_i log_2 M$, and the entire inequality can be written as

$$\begin{gathered}\sum_{1\leq i\leq n-1}\left(P_i log_2 P_i - P_i log_2 1 + P_i log_2 M\right) \\ + P_n log_2 P_n - P_n log_2 (M-n+1) + P_n log_2 M \\ > \\ \sum_{1\leq i\leq n-1}\left(P_i log_2 P_i - P_i log_2 (x_i+1) + P_i log_2 M\right) \\ + P_n log_2 P_n - P_n log_2 (x_n+1) + P_n log_2 M\end{gathered} \tag{12}$$

that is

$$\begin{gathered}\sum_{1\leq i\leq n-1}\left(P_i log_2 1\right) - P_n log_2 (M-n+1) \\ > \\ \sum_{1\leq i\leq n-1}\left(-P_i log_2 (x_i+1)\right) - P_n log_2 (x_n+1)\end{gathered} \tag{13}$$

that is

$$-P_n log_2 (M-n+1) > -\sum_{1\leq i\leq n} P_i log_2 (x_i+1) \tag{14}$$

Since P is ordered, for each position i it happens that $P_i = P_{i-1} - \epsilon_i$, namely $P_{i-1} = P_i + \epsilon_i$. The inequality can be written as

$$\begin{aligned} P_n log_2 (M-n+1) < \\ (P_n) log_2 (x_n+1) + \\ (P_n+\epsilon_n) log_2 (x_{n-1}+1) + \\ (P_n+\epsilon_n+\epsilon_{n-1}) log_2 (x_{n-2}+1) + \\ \cdots \\ (P_n+\epsilon_n+\cdots+\epsilon_{n-n+2}) log_2 (x_{n-n+1}+1)\end{aligned} \tag{15}$$

The arguments of the logarithms are always greater than 1, thus the values of the logarithms are always positive. Moreover, the factors that multiply the logarithms are always positive because they are probabilities. The inequality can be written as

$$P_n log_2(M-n+1) < nP_n\left(\sum_{1\leq i\leq n} log_2(x_i+1)\right) + c \tag{16}$$

with $c \geq 0$ and $\sum_{1\leq i\leq n}(x_i+1) = M-n$. Considering that the sum of logarithms is greater than the logarithm of the sum [24], it is now trivial to show that the inequality is always satisfied. □

The previous proof is given for an ordered distribution P. However, the final inequality is independent of the ordering. In fact, it compares the quantity $M-n+1$ (that is, the one that makes U the distribution of interest) with the sum of the x_i+1 terms independent of their position and specific value. P is ordered, and U assigns by construction the additional $M-n$ quantity to the cell where P has the lowest assigned value.

The retrieving of an upper bound for the entropic divergence is here shown to be possible under two main conditions: (i) no zero values are assigned by the two distributions; (ii) the compared distributions are quantized distributions over the same quantum value $1/M$. The first condition is often ensured in practical applications, where pseudo-counts are used to avoid infinite divergences. The second condition emerges from this study. It states that the entropic divergence acquires a more powerful meaning when applied to *comparable* distributions. The term *comparable* refers to sharing the same quantum value. This aspect should be taken into account in future developments of divergences.

4. A Notion of Normalized Entropic Divergence

The retrieving of the maximizing distribution is exploited to normalize the entropic divergence in the range $[0,1]$, both included. Given two distributions P and Q, the normalized entropic divergence is calculated as

$$NKL(P||Q) = \frac{KL(P||Q)}{KL(P||U)} \tag{17}$$

where U is the distribution for which the maximum entropic divergence from P is obtained. This maximizing distribution is constructed based on the results from the previous section. Specifically, it distributes a minimum value of 1 to each cell, and the remaining quantity $M-n$ is assigned to the cell for which the value in P is the minimum.

In what follows, the proposed normalized entropic divergence is compared with the most-used notions of entropic divergence, plus a measure that is highly suitable for comparing multiplicity distributions. The comparison is performed by looking at three different aspects: (i) the difference in the values that the measures output on comparing two distributions (see Section 4.2); (ii) the spread of output values within the output range (see Section 4.4; (iii) the diversity of the measures in assigning a rank (see Section 4.5). The relation between the measures and the properties of the compared distributions is investigated too (see Section 4.3).

The investigations are empirically conducted by computationally generating the distributions. The source code for generating the unordered and ordered distributions and the computational experiments are available at the following link https://github.com/vbonnici/KL-maxima (accessed on 1 September 2024).

4.1. Compared Measures

The proposed divergence is compared with the unnormalized one, namely $KL(P||Q)$, and with the commonly used symmetric divergence, also called Jensen–Shannon divergence (JS). The JS divergence is defined as

$$JSD(P,Q) = \frac{KL(P||A) + KL(Q||A)}{2} \tag{18}$$

with $A = \frac{P+Q}{2}$, and it is known to be upper-bounded by 1 if the base of the logarithm is 2 [12].

Another important divergence is the Hellinger distance, which is defined as

$$HE(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{1\leq i\leq n}(\sqrt{P_i} - \sqrt{Q_i})^2} \tag{19}$$

and it can also be written as $HE^2(P,Q) = 1 - \sum_{1\leq i\leq n}\sqrt{P_iQ_i}$. Important properties of such a divergence are that it implicitly avoids infinite divergences and it is bounded in the range $[0\ldots1]$.

The generalized Jaccard similarity is a measure suitable for comparing multiplicity distributions. It is defined as:

$$J(P,Q) = \frac{\sum_{1\leq i\leq n} min(P_i,Q_i)}{\sum_{1\leq i\leq n} max(P_i,Q_i)} \tag{20}$$

It can be shown that such a measure ranges from 0 to 1, both included. The minimum value is reached when the two distributions have no multiplicity in common, which means that $P_i = 0$ when $Q_i \neq 0$ and vice versa. It reaches the maximum value when the two distributions have equal values. It is a notion of similarity. Therefore, it is in contrast with the meaning of entropic divergence. Thus, for this study, it is converted as $\widetilde{J}(P,Q) = 1 - J(P,Q)$ to have it as a notion of distance.

The generalized Jaccard distance is directly applied to multiplicity distributions, while entropic divergences are applied after converting the distributions into probability/frequency distributions.

4.2. Direct Comparison of Output Values

In what follows, scatter plots are used for investigating differences between the compared measures when they are applied to measure the distance/divergence between two quantized distributions. Each point within the scatter plot represents a specific pair of distributions that are put in comparison. The position of the point within the plot depends on the values of the compared measures. For example, in Figure 2a, the classical Kullback–Leibler divergence is compared with the proposed unnormalized Kullback–Leibler divergence, respectively, on the axis of ordinates and the axis of abscissae. Thus, the coordinates of a point representing a specific pair of distributions are given by the KL divergence and the NKL divergence between them. Since a relatively huge number of two-by-two distribution comparisons are made, many points overlap in the same area of the plot. Thus, the chart is also equipped with two histograms located beside the axes that report the number of instances that fall within a given range of values.

Figure 2 reports the relations between the proposed normalized divergence and the other investigated measures. Calculations were performed by setting a number of cells equal to 5 and a total distributed quantity of 15. The experiment generated 1001 unordered distributions, of which 30 were monotonically decreasing ordered. Thus, a total of 1001 two-by-two distribution comparisons were performed.

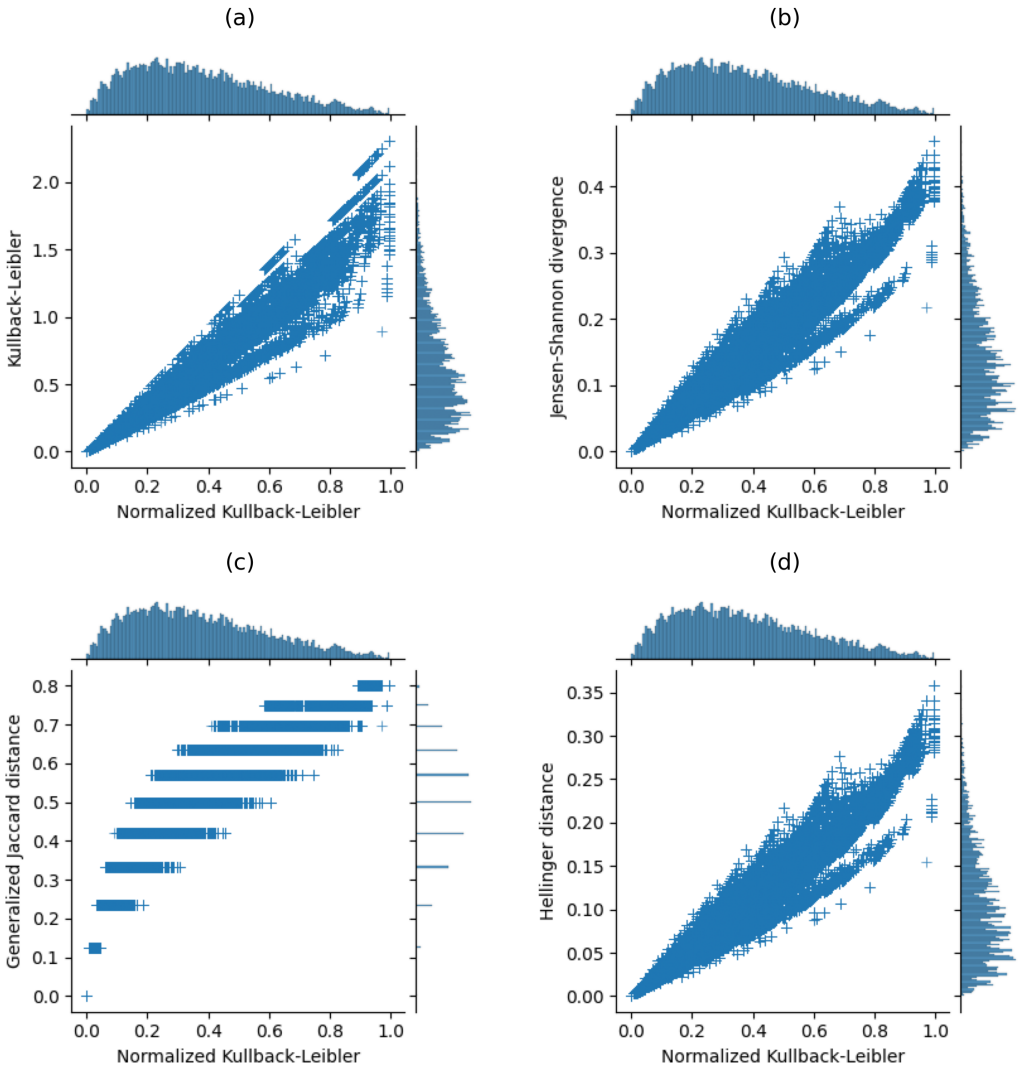

Figure 2. Relation between the proposed normalized Kullback–Leibler divergence and (**a**) unnormalized Kullback–Leibler divergence; (**b**) symmetric Kullback–Leibler divergence; (**c**) generalized Jaccard distance; (**d**) Hellinger distance.

The proposed measure is more correlated with the non-symmetric divergence than the other measures. The Pearson correlation coefficient [25] reaches a value of 0.97 between the proposed divergence and the unnormalized one and a correlation value of 0.96 between the proposed measure and the symmetric divergence. The complete list of Pearson correlation coefficients between the compared measures is reported in Table A1.

4.3. Relation with Distributional Properties

Entropic divergences and other measures can be used to prioritize elements with respect to their deviance from randomness or, generically, from a background model. Thus, it can be interesting to study how the rank assigned to elements, based on their divergence, changes when the four different measures are used. In what follows, the uniform distribution is used as the background model, and the measure of divergence from it is calculated for the set of ordered distributions that can be formed by taking into account the same quantity that is distributed in the uniform shape. For the experiments, a number of cells equal to 8 and a total quantity of 32 have been considered. In this way, the uniform distribution assigns a quantity of four to each cell. The difference with respect to the previous experiments, where 5 cells and 15 elements are considered, is because the previous experiment generates only 30 distinct ordered distributions, which is a relatively

small number. On the contrary, a setup with 8 cells and 32 elements generates a high number of unordered distributions (2,629,575) that leads to a huge number of two-by-two comparisons. As a pro, the new setup generates 919 ordered distributions, which can be considered sufficient to draw experimental conclusions.

The correlation between the measures and the properties of the compared distributions is investigated. Entropy, coefficient of variation, skewness and the Kurtosis index are the considered properties.

Figure 3 shows the relation between the four investigated measures and the entropy of the ordered distribution that is compared with the uniform distribution. The simple Kullback–Leibler divergence is the measure that better correlates with the entropy, followed by the proposed normalized divergence. Table A2 reports the correlations between the measures and the entropy. The numeric correlations confirm what is shown by the graphics.

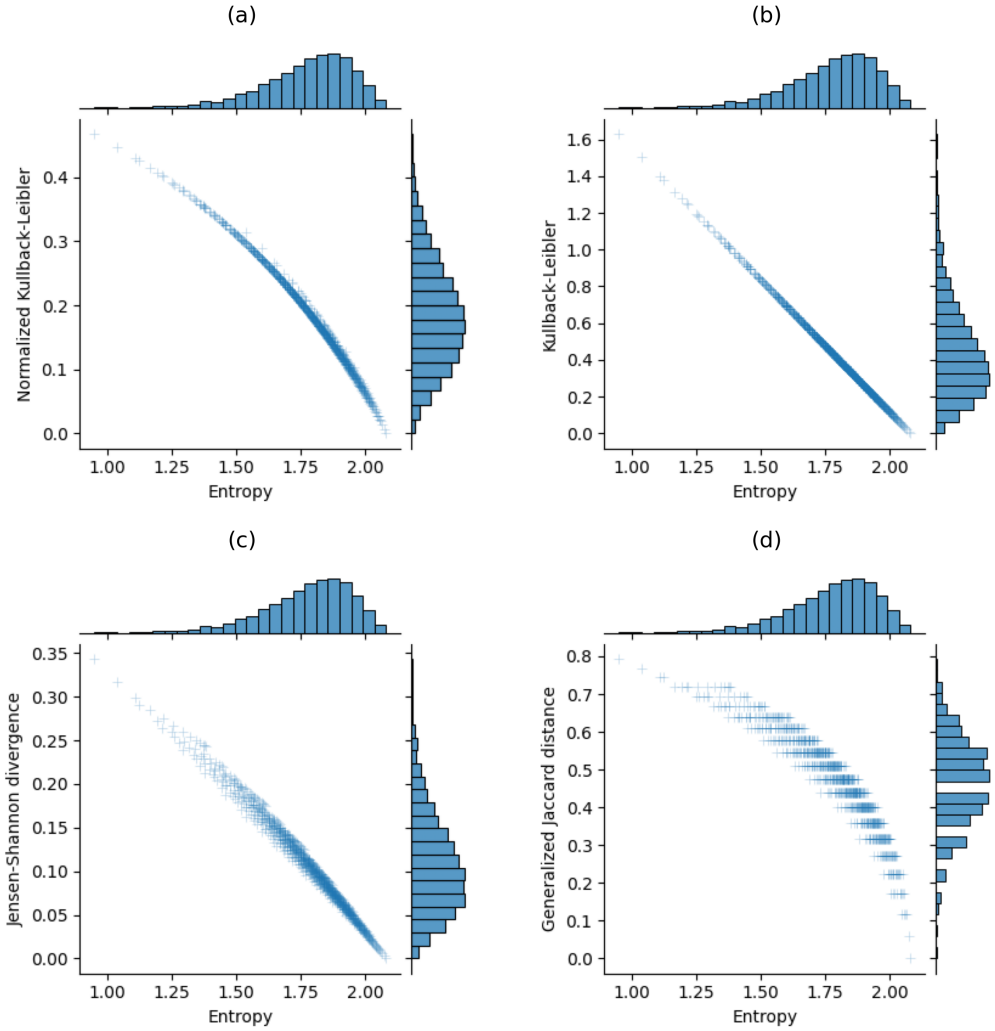

Figure 3. Scatter plots generated by putting in relation four of the investigated measures and the entropy of the set of monotonically ordered distributions, generated with 8 cells and 32 dots, and the corresponding uniform distribution: (**a**) normalized Kullback–Leibler divergence, (**b**) unnormalized Kullback–Leibler divergence; (**c**) Jensen–Shannon divergence; (**d**) generalized Jaccard distance.

Figure 4 shows the relation between the four measures and the coefficient of variation of the ordered distribution that is compared with the uniform one. Pearson correlation coefficients are reported in Table A2 of Appendix C. Differently from entropy-related correlations, the proposed normalized measure is the one that better correlates with the coefficient of variation, followed by the unnormalized entropic divergence. Moreover, unlike

the unnormalized Kullback–Leibler divergence and the Jensen–Shannon divergence, the proposed normalized divergence forms a sigmoid curve rather than an exponential trend.

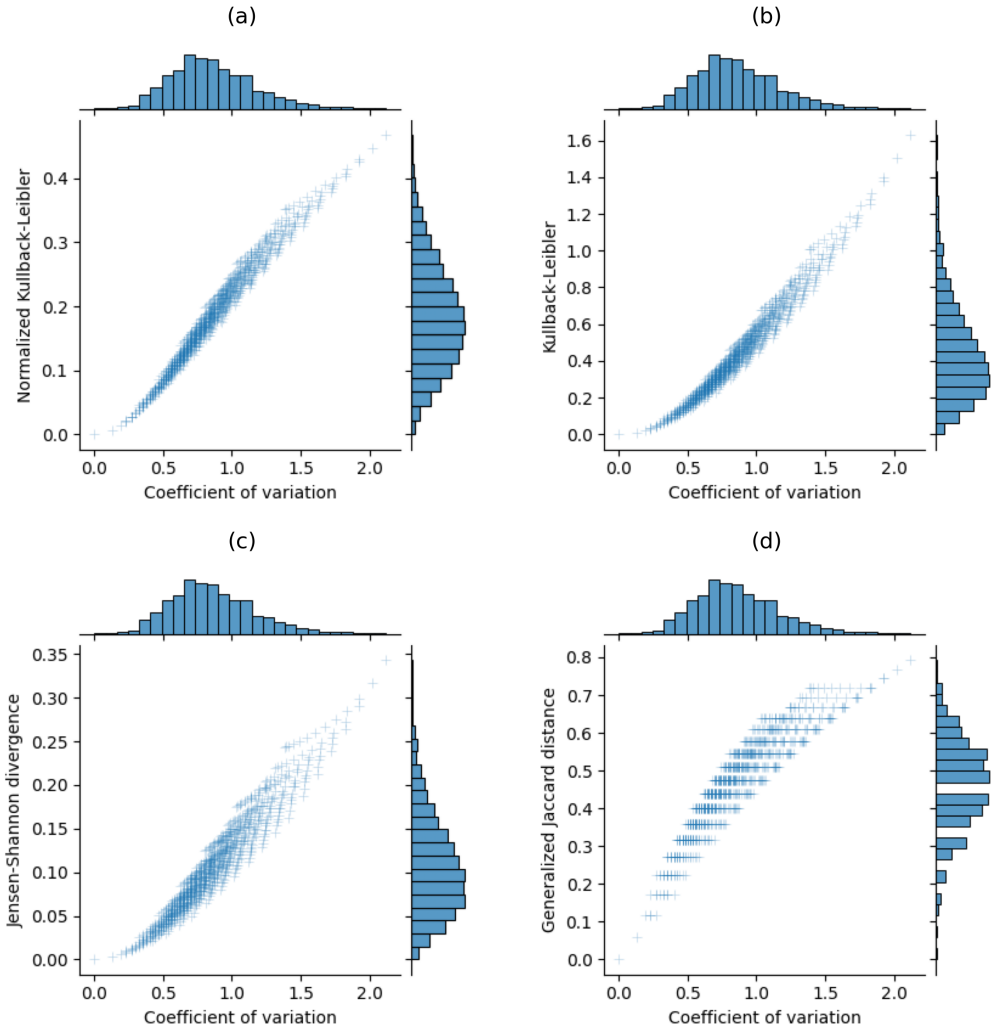

Figure 4. Scatter plots generated by putting in relation four of the investigated measures and the coefficient of variation of the set of monotonically ordered distributions, generated with 8 cells and 32 dots, and the corresponding uniform distribution: (**a**) normalized Kullback–Leibler divergence, (**b**) unnormalized Kullback–Leibler divergence; (**c**) Jensen–Shannon divergence; (**d**) generalized Jaccard distance.

Entropy and coefficient of variation are the distributional properties that better correlate with the investigated measures. In Figures A1 and A2 of Appendix C, it is shown that the skewness and the Kurtosis's index of the compared unordered distribution weakly correlate with the measures. However, both distributional properties form shapes similar to grids when they are plotted. This behavior is possibly due to the discrete nature of the compared distributions.

4.4. Outcome Spread Diversity

Scatter plots and histograms of the proposed figures show interesting behaviors of the investigated measures related to how the output values of these measures spread along the output range.

For example, visible clusters are formed by the generalized Jaccard distance (see Figure 2). This behavior directly emerges from Equation (20) since the Jaccard distance tends to flatten the punctual comparison among the elements in the domain of the dis-

tribution into a sum of values of multiplicity. The divergences do not seem to produce such clusters; however, it can be helpful to investigate such a phenomenon more properly. Distances between consecutive values of the two measures have been taken into account. Given a set of n comparisons, a vector of size n is built from the values of the specific measure on such comparisons. The vector is sorted and then runs within the vector, reporting that the same values are substituted with one single value. The differences between adjacent positions of the vector are extracted. Then, the mean and the standard deviation are computed. The elimination of the runs on the vector of the generalized Jaccard measure decreases the size of the vector from 1001×1001 to 11, as can be observed in the figure. The distances of the generalized Jaccard measure have a mean equal to 0.08 and a standard deviation of 0.02. On the contrary, the distances of the normalized entropic divergence have an average of 0.00004 and a standard deviation of 0.0005. Thus, it seems that the divergence is not forming clusters.

Regarding the experiments presented in Section 4.3, the compared measures have different output ranges. By definition, the unnormalized entropic divergence and the Jensen–Shannon divergence have no upper bound; on the contrary, the proposed measure and the generalized Jaccard distance are expected to range between 0 and 1. The proposed normalized divergence ranges from 0 to circa 0.5 because one of the two compared distributions is always the uniform distribution. The monotonically ordered distribution that diverges more from the uniform distribution is the one that assigns all the available quantity to the first cell. Such a distribution is completely opposed to U, and the uniform distribution is in the middle of them. Thus, the divergence from the distribution to the uniform one is half of the divergence from U. The generalized Jaccard distance is influenced by the fact that values close to 1 can not be reached because the compared distributions have no term equal to 0. The maximum observable distance is 0.8.

Table 1 shows the maximum value that each investigated measure reaches with the varying numbers of cells and dots with which distributions are built. All the measures have a minimum value of 0 because the uniform distribution is among the distributions that are compared to itself. The proposed normalized divergence takes values that are close to 0.5 but never equal to such a value. The reason resides in the discretized nature of the compared distributions. However, some pattern emerges from the table. The values of the measures are directly related to the number of dots that are distributed. The smaller the number of dots is, the higher the value of the proposed normalized measure is. This behavior is opposite to the three other measures, which increase their value by increasing the number of distributed dots. Intuitively, the distribution that maximizes the divergence/distance from the uniform distribution is the one that assigns all the available dots to the first cell; thus, it is specular to U. Computational experiments also confirm this intuition. The fact that the measure takes different values depends on the ratio between the dots that are assigned to the first cell and the number of cells. For example, the uniform distributions obtained for 6 cells and 12 dots and for 7 cells and 14 dots are almost identical. Both of them assign two dots to each cell. However, the number of available dots, after assigning one dot to each cell, is six in the first case and seven in the second case. Thus, the difference between the two generalized Jaccard distances is $\frac{2}{7}$ versus $\frac{2}{8} + \frac{1}{2} = \frac{3}{4}$ because except for the first cell, all the other cells carry a value of $\frac{1}{2}$ for both configurations. The configuration with seven cells has an additional cell. This difference, notably, leads to a different resulting value. Similar considerations can be made for the other measures.

The difference in how the measures spread the values, along with the range from 0 to the maximum value, is summarized in Table 2. Each experiment regards a specific number of cells and dots, as for the previous analysis. The average value divided by the maximum value is used as a measure of spread. The closer the resultant measurement is to 0.5, the greater the spread of the values. On the contrary, if the measurement tends to 0, then the values are more concentrated towards 0, and similarly, they are concentrated towards the maximum if the measurement tends to 1. The proposed normalized divergence better tends to 0.5, with an average value of 0.4296, along with the complete set of experiments.

The unnormalized KL tends to 0 more than the Jensen–Shannon divergence, which is in contrast with the mode observed in the figures. The generalized Jaccard distance tends more to the maximum value, with an average of 0.6.

Table 1. Maximum values of the five investigated measures by varying the number of cells and dots by which the distributions are formed by.

Cells	Dots	Norm. KL	Unnorm. KL	Jensen–Shannon	Hellinger	Gen. Jaccard
6	12	0.5078	0.6376	0.1395	0.0989	0.5882
6	18	0.4498	1.0876	0.2399	0.1719	0.7143
6	24	0.4297	1.3629	0.3046	0.2201	0.7692
6	30	0.4164	1.5480	0.3500	0.2546	0.8000
7	14	0.5151	0.7143	0.1518	0.1082	0.6000
7	21	0.4687	1.2057	0.2578	0.1857	0.7273
7	28	0.4502	1.5038	0.3257	0.2364	0.7826
7	35	0.4374	1.7033	0.3731	0.2726	0.8136
8	16	0.5233	0.7831	0.1622	0.1161	0.6087
8	24	0.4845	1.3103	0.2727	0.1973	0.7368
8	32	0.4672	1.6280	0.3429	0.2500	0.7925
8	40	0.4546	1.8397	0.3919	0.2876	0.8235
9	18	0.5315	0.8455	0.1711	0.1230	0.6154
9	27	0.4981	1.4043	0.2851	0.2072	0.7442
9	36	0.4815	1.7391	0.3573	0.2616	0.8000
9	45	0.4691	1.9614	0.4076	0.3002	0.8312
10	20	0.5394	0.9027	0.1789	0.1291	0.6207
10	30	0.5098	1.4897	0.2958	0.2158	0.7500
10	40	0.4938	1.8395	0.3696	0.2716	0.8060
10	50	0.4815	2.0713	0.4208	0.3112	0.8372

Table 2. Average divided by maximum value of the five investigated measures by varying the number of cells and dots by which the distributions are formed by.

Cells	Dots	Norm. KL	Unnorm. KL	Jensen–Shannon	Hellinger	Gen. Jaccard
6	12	0.5301	0.4089	0.4418	0.4379	0.6203
6	18	0.4403	0.3217	0.3540	0.3495	0.5979
6	24	0.3987	0.2865	0.3159	0.3110	0.5848
6	30	0.3739	0.2672	0.2941	0.2886	0.5766
7	14	0.5277	0.3987	0.4365	0.4315	0.6277
7	21	0.4396	0.3139	0.3523	0.3466	0.6069
7	28	0.3965	0.2778	0.3131	0.3071	0.5910
7	35	0.3709	0.2583	0.2910	0.2846	0.5815
8	16	0.5318	0.3931	0.4379	0.4314	0.6483
8	24	0.4390	0.3066	0.3505	0.3436	0.6144
8	32	0.3956	0.2711	0.3117	0.3046	0.5967
8	40	0.3694	0.2514	0.2891	0.2818	0.5860
9	18	0.5332	0.3871	0.4369	0.4292	0.6578
9	27	0.4404	0.3017	0.3508	0.3427	0.6218
9	36	0.3961	0.2660	0.3114	0.3033	0.6022
9	45	0.3692	0.2462	0.2885	0.2803	0.5906
10	20	0.5362	0.3832	0.4384	0.4294	0.6708
10	30	0.4421	0.2977	0.3514	0.3423	0.6284
10	40	0.3972	0.2621	0.3119	0.3029	0.6076
10	50	0.3696	0.2422	0.2886	0.2796	0.5951
avg		0.4349	0.3071	0.3483	0.3414	0.6103

4.5. Differences in Ranking Outcomes

Lastly, the difference in the ranking produced by the four measures has been investigated. Experimental results were obtained using 8 cells and 32 dots. As in the previous

experiment, the uniform distribution was compared to the set of monotonically decreasing ordered distributions. Then, distributions were ranked depending on the value each measure assigned to them. Figure 5 compares the normalized entropic divergence and the three other measures in assigning the rank to the distributions. Each point in one of the three plots is a given distribution whose coordinates, in the Cartesian plane, are given by the rank assigned by the two compared measures. These charts show how different a ranking can be when applying different measures. A mathematical method for comparing rankings is Spearman's rank correlation coefficient [26], whose values are reported in Table A3 of Appendix D. The reported correlations may appear significantly high. However, there is a discordance between the measures from circa 0.05 to 0.001, which means that from 5% to 0.1% of the elements are ranked differently. Such a difference may, for example, lead to different empirical p-values, which may change the results of a study.

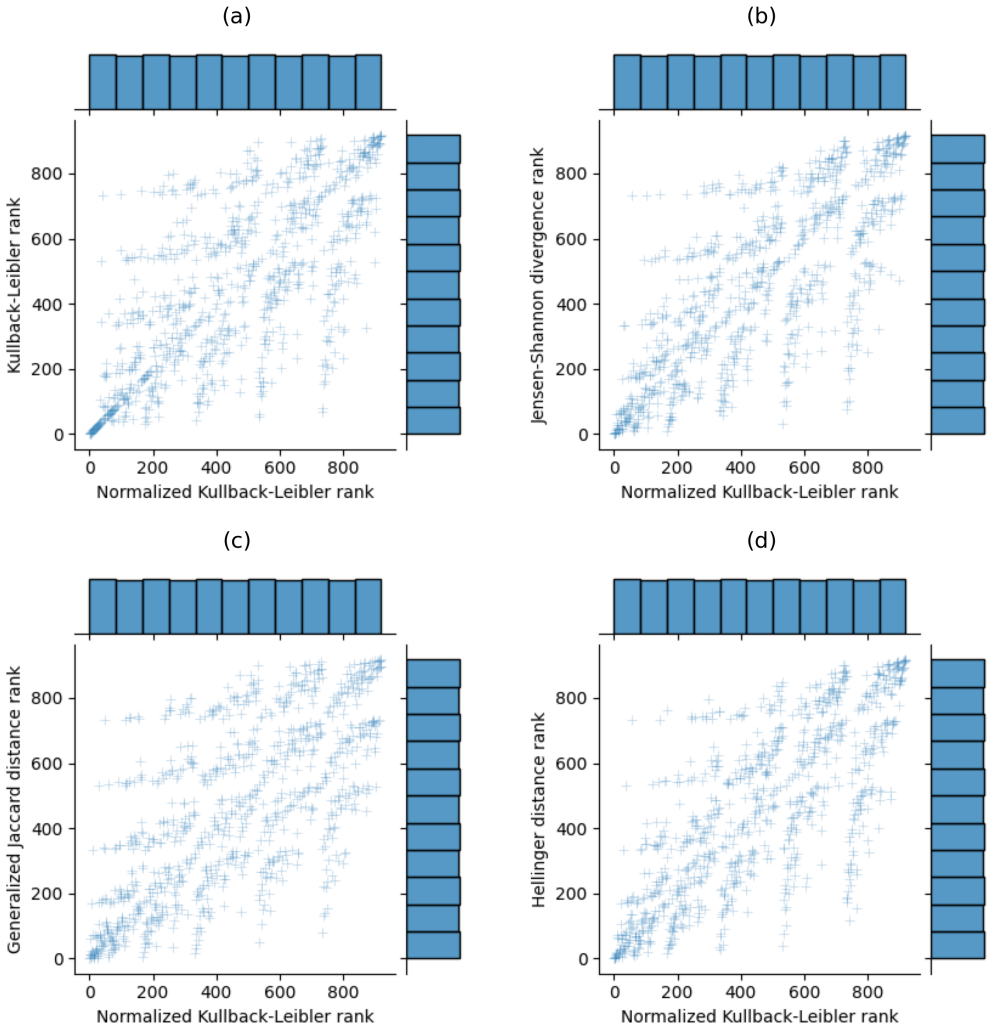

Figure 5. Scatter plots obtained by considering the rank assigned by the proposed normalized Kullback–Leibler and the other investigated measures: (**a**) unnormalized Kullback–Leibler divergence, (**b**) Jensen-Shannon divergence, (**c**) generalized Jaccard distance and (**d**) Hellinger distance. The complete set of monotonically ordered distributions generated with 8 cells and 32 dots was used for extracting the rankings.

5. Conclusions

This study demonstrates that for any probability distribution P, there exists another distribution U that maximizes the entropic divergence from P, provided infinite divergences are avoided. P and U must have been generated by distributing a given discrete quantity.

In the realm of quantum theory, the real world is composed of discretized quantities called quanta. Thus, quantized probability distributions are, in their essence, multiplicity distributions. This implies that the findings presented here have broad applicability.

Here, the shape of the distribution U is characterized, and it is used to provide a notion of entropic divergence normalized between 0 and 1. Empirical evaluation of such a normalized divergence with respect to other commonly used measures is reported. The evaluation demonstrates that the proposed divergence exhibits distinct behavior, differing from established measures, as the properties of the compared distributions vary.

This study highlights an important aspect of entropic divergence. An upper bound to the divergence is obtainable only if the two compared distributions are formed by the same quantum. Future developments of divergence should take this aspect into account.

Funding: This research received no external funding.

Institutional Review Board Statement: This study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: Scripts for reproducing all the analyses presented in this manuscript are available at https://github.com/vbonnici/KL-maxima (accessed 1 September 2024).

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KL	Kullback–Leibler divergence
NKL	Normalized Kullback–Leibler divergence
QD	Quantized Distribution
OQD	Ordered Quantized Distribution
JS	Jensen–Shannon divergence

Appendix A. Special Case 1: General Proof

Proposition A1. *Let P be an OQD obtained by distributing a quantity M to n cells. Let U be a QD, which assigns all the free quantity $M-n$ to the n-th cell and the minimum quantity of 1 to each cell. Let Q be a QD, which assigns a quantity $M-n-1$ to the n-th cell and a quantity of 1 to the $(n-1)$-th cell, in addition to the minimum quantity of 1 to each cell. Then, $KL(P||U) > KL(P||Q)$.*

Proof. A first consideration is that from position 1 to $n-2=3$, the two divergences have identical contributions; thus, they can be ignored in the comparison. Therefore, it has to be proven that:

$$P_{n-1} log_2 \frac{P_{n-1}}{U_{n-1}} + P_n log_2 \frac{P_n}{U_n} > P_{n-1} log_2 \frac{P_{n-1}}{Q_{n-1}} + P_n log_2 \frac{P_n}{Q_n} \tag{A1}$$

By construction, $U_n = \frac{M-n+1}{M}$ and $U_{n-1} = \frac{1}{M}$, while $Q_n = \frac{M-n}{M}$ and $Q_{n-1} = \frac{2}{M}$. Thus, Equation (A1) can be written as:

$$P_{n-1} log_2 \frac{P_{n-1}}{\frac{1}{M}} + P_n log_2 \frac{P_n}{\frac{M-n+1}{M}} > P_{n-1} log_2 \frac{P_{n-1}}{\frac{2}{M}} + P_n log_2 \frac{P_n}{\frac{M-n}{M}} \tag{A2}$$

that is

$$\begin{aligned} P_{n-1} log_2 P_{n-1} - P_{n-1} log_2 \frac{1}{M} + P_n log_2 P_n - P_n log_2 \frac{M-n+1}{M} > \\ P_{n-1} log_2 P_{n-1} - P_{n-1} log_2 \frac{2}{M} + P_n log_2 P_n - P_n log_2 \frac{M-n}{M} \end{aligned} \tag{A3}$$

and therefore, by removing equal terms from the left and right sides of the inequality,

$$-P_{n-1}log_2\frac{1}{M} - P_nlog_2\frac{M-n+1}{M} > -P_{n-1}log_2\frac{2}{M} - P_nlog_2\frac{M-n}{M} \tag{A4}$$

that is

$$\begin{aligned} -P_{n-1}log_2(1) + P_{n-1}log_2(M) - P_nlog_2(M-n+1) + P_nlog_2(M) > \\ -P_{n-1}log_2(2) + P_{n-1}log_2(M) - P_nlog_2(M-n) + P_nlog_2(M) \end{aligned} \tag{A5}$$

therefore, since $log_2(1) = 0$ and by removing equal terms,

$$-P_nlog_2(M-n+1) > -P_{n-1}log_2(2) - P_nlog_2(M-n) \tag{A6}$$

For this specific case, the difference between P_n and P_{n-1} is given by a single element. However, since P is ordered, it can be assumed that there is a discretized gap between the two positions such that $P_{n-1} = P_n + \epsilon$, for $\epsilon \in \mathbb{N}, \geq 0$. Thus, the inequality can be written by also changing its verse as

$$P_nlog_2(M-n+1) < (P_n+\epsilon)log_2(2) + P_nlog_2(M-n) \tag{A7}$$

that is

$$P_nlog_2(M-n+1) < P_nlog_2(2) + \epsilon log_2(2) + P_nlog_2(M-n) \tag{A8}$$

that is

$$P_nlog_2(M-n+1) - P_nlog_2(2) - P_nlog_2(M-n) < \epsilon log_2(2) \tag{A9}$$

that is

$$P_n\left(log_2(M-n+1) - log_2(2) - log_2(M-n)\right) < \epsilon log_2(2) \tag{A10}$$

It can be assumed that $P_n = k\epsilon$, for a given factor $k \in \mathbb{R}, > 0$; thus, P_n can be greater or smaller than ϵ. In addition, $log_2(M-n+1) - log_2(2) - log_2(M-n)$ equals $log_2\frac{M-n+1}{2(M-n)}$. Thus, the inequality can be written as

$$k\epsilon\left(log_2\frac{M-n+1}{2(M-n)}\right) < \epsilon log_2(2) \tag{A11}$$

and, therefore

$$k\left(log_2\frac{M-n+1}{2(M-n)}\right) < log_2(2) \tag{A12}$$

If $M-n > 1$, which is always true because a minimum amount of 1 is assigned to each cell and the two distributions must be different, then $\frac{M-n+1}{2(M-n)}$ is always less than 1. This implies that $log_2\frac{M-n+1}{2(M-n)}$ is always less than or equal to zero. Thus, the inequality is always satisfied independently from the value of k, which must be in any case ≥ 0.

More generally, Equation (A12) can be written as:

$$k\left(log_2\frac{M-n+1}{(1+x)(M-n+1-x)}\right) < log_2(1+x) \tag{A13}$$

because a given quantity $x+1$, that is at least 1 and at most $M-n+1$, is moved from position n to position $n-1$.

In Equation (A13), we can put $M - n + 1 = y$ and thus, to assert that the result of the logarithm must always be less the 0, it has to be shown that

$$\begin{aligned} y &< (1+x)(y-x) \\ y &< y + xy - x - x^2 \\ 0 &< +xy - x - x^2 \\ 0 &< x(y-1) - x^2 \\ 0 &> x(1-y) + x^2 \end{aligned} \tag{A14}$$

The determinant is given by $(1-y)^2 - 4$ that is: equal to 0 for $M = n - 4$, which is impossible because $M > n$; less than 0 for $M < n - 4$, which is still impossible because $M > n$; and greater than 0 for $M > n - 4$. Thus, the determinant is always greater than 0, and the inequality is less than 0, which means that it admits two solutions x_1 and x_2 such that it is valid for $x_1 < x < x_2$. The two solutions are given by $\frac{(y-1) \pm \sqrt{(1-y)^2-4}}{2}$. The determinant can also be written as $(1-y)^2 - 4 = (1-M+n-1)^2 - 4 = (M+n)^2 - 2^2$. For practical applications, the determinant can be approximated to $(M+n)^2$, thus the inequality is satisfied for $(M - n + M + n)/2 < x < (M - n - M - n)/2$, namely $-n < x < M$, which is always true because $x \leq M - n$ by definition. □

The fact that Equation (A13) is always verified implies that, independently of how the quantity is arranged in the last two positions, the distribution U is the one that maximizes the entropic divergence. It also implies two other assertions. The first assertion is that if the number of cells is equal to 2, then U is always the maximizing distribution. The second assertion is that if the quantity is moved from the last cell to a specific other cell, not necessarily the second-last, the U is still the maximizing distribution. The inequality is independent of the specific cell position, and it only requires that $P_i = P_n + \epsilon$ and that $P_n = k\epsilon$, thus $P_i = k\epsilon + \epsilon = \epsilon(k+1)$, which means that P_i must be greater than P_n. This consideration highlights the fact that U is the distribution that assigns all the available quantities to the cell with the smallest probability in P; thus, it is independent of the ordering.

Appendix B. Comparisons between Measures

Table A1 reports Pearson's correlation among the five measures that are compared in the main article. Correlations are calculated by taking into account the values of the measures in computing the divergence (dissimilarity) between ordered distributions. The distributions are built by distributing a quantity of 15 to 5 cells.

Table A1. Pearson's correlation among the investigated measures on two-by-two comparisons of ordered distributions generated by distributing a quantity of 15 to 5 cells.

		Pearson Corr.
Normalized Kullback–Leibler	Kullback–Leibler	0.9893
Normalized Kullback–Leibler	Jensen–Shannon divergence	0.9888
Normalized Kullback–Leibler	Generalized Jaccard distance	0.9549
Normalized Kullback–Leibler	Hellinger distance	0.9881
Kullback–Leibler	Jensen–Shannon divergence	0.9926
Kullback–Leibler	Generalized Jaccard distance	0.9232
Kullback–Leibler	Hellinger distance	0.9932
Jensen–Shannon divergence	Generalized Jaccard distance	0.9441
Jensen–Shannon divergence	Hellinger distance	0.9999
Hellinger distance	Generalized Jaccard distance	0.9411

Appendix C. Correlation with Distributional Properties

Table A2 reports Pearson correlation coefficients among the investigated measures on comparing ordered distributions with the uniform one generated by distributing a quantity of 32 to 8 cells.

Figure A1 shows the relation of the compared measures with the Kurtosis's index, and Figure A2 shows the relation with the skewness. It has to be noticed that some values of the skewness and Kurtosis statistics may appear unexpected. Such an unexpected behavior is because relatively small (in their cardinality) distributions are taken into account. Furthermore, the generated distributions are more similar to exponential distributions than normal ones. For example, only positive values of skewness are expected because the examined distributions are monotonically ordered. However, the distribution whose values are $(7,7,7,7,1,1,1,1)$ has a skewness of 0 because the mean, mode, and median of the distribution have the same value. The distribution $(7,7,6,6,3,1,1,1)$ has a negative skewness because the mode (1) is smaller than the mean (4).

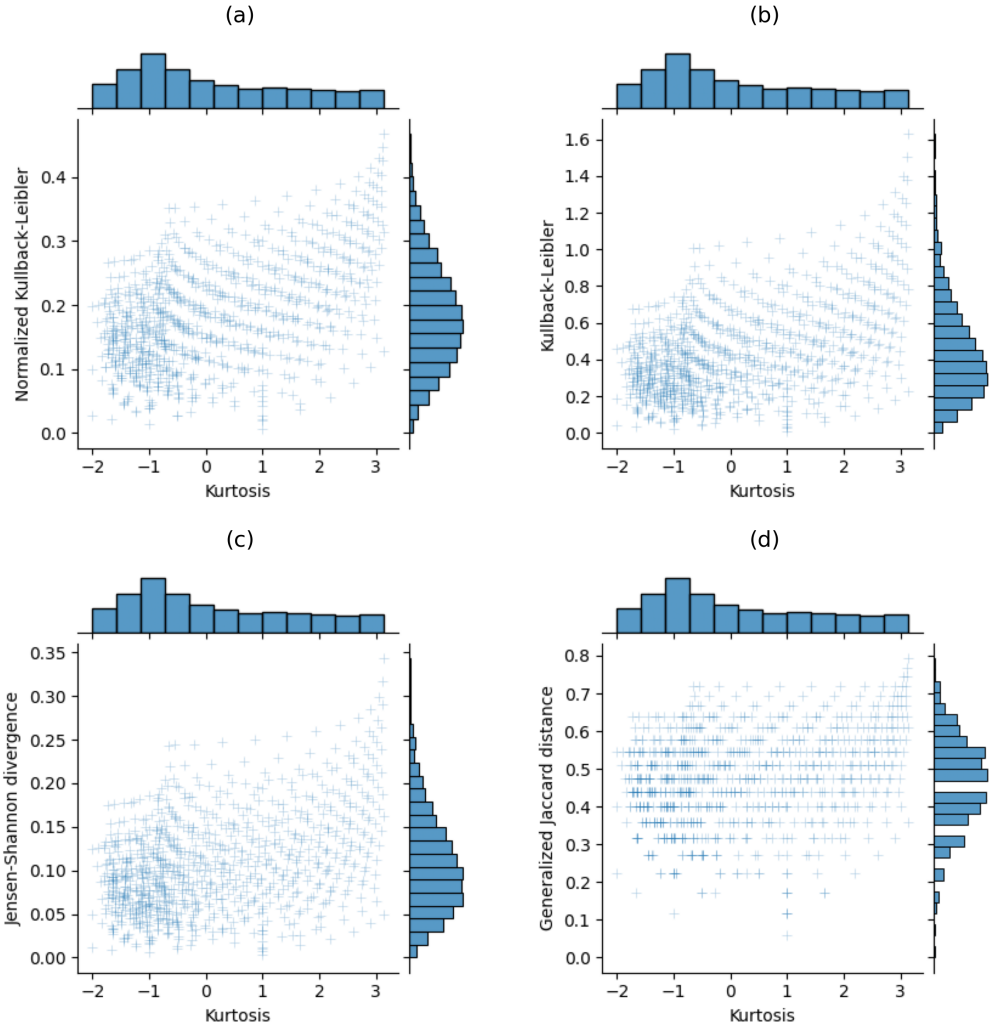

Figure A1. Scatter plots generated by putting in relation four of the investigated measures and the Kurtosis index of the set of monotonically ordered distributions, generated with 8 cells and 32 dots, and the corresponding uniform distribution: (**a**) normalized Kullback–Leibler divergence, (**b**) unnormalized Kullback–Leibler divergence; (**c**) Jensen-Shannon divergence; (**d**) generalized Jaccard distance.

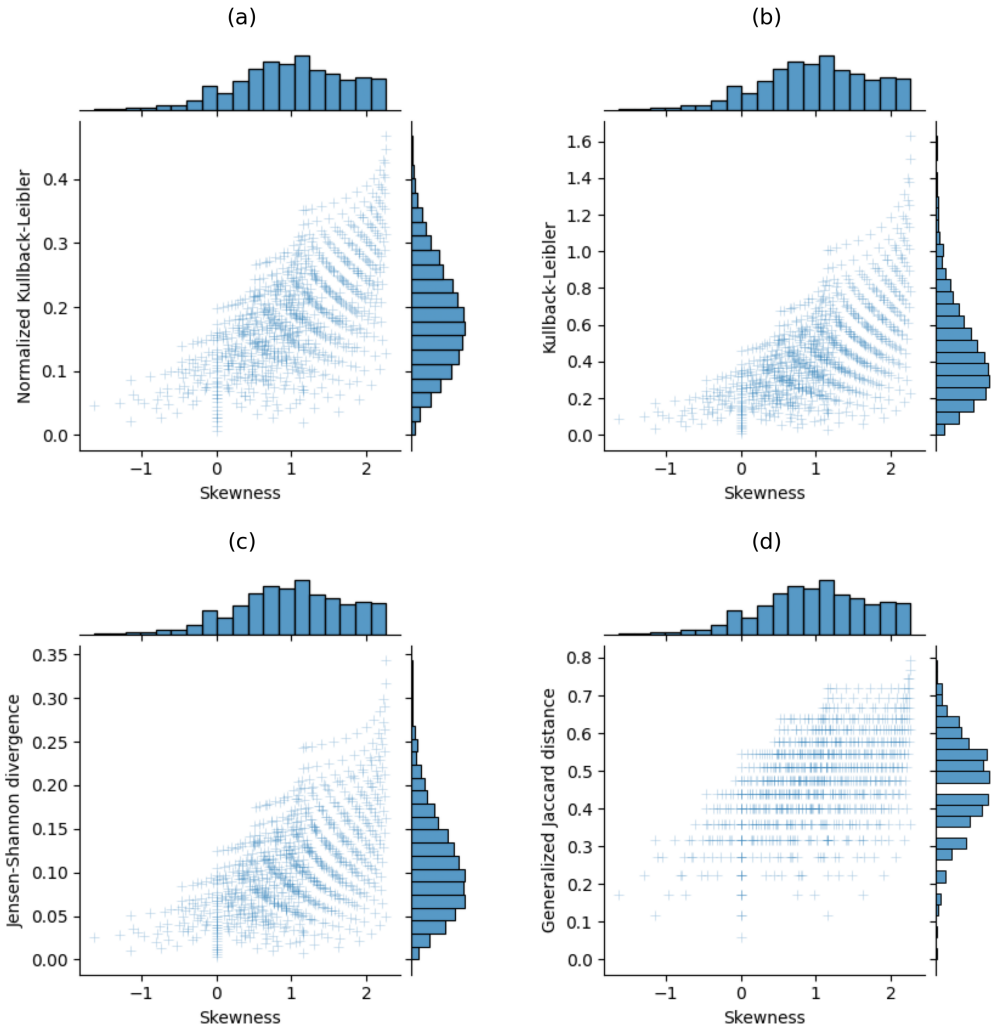

Figure A2. Scatter plots generated by putting in relation four of the investigated measures and the skewness of the set of monotonically ordered distributions, generated with 8 cells and 32 dots, and the corresponding uniform distribution: (**a**) normalized Kullback–Leibler divergence, (**b**) unnormalized Kullback–Leibler divergence; (**c**) Jensen-Shannon divergence; (**d**) generalized Jaccard distance.

Table A2. Pearson correlation coefficients among the investigated measures on comparing ordered distributions with the uniform one generated by distributing a quantity of 32 to 8 cells.

Measure	**Property**	**Correlation**
Normalized Kullback–Leibler	Entropy	−0.9892
Kullback–Leibler	Entropy	−0.9999
Jensen–Shannon divergence	Entropy	−0.9804
Generalized Jaccard distance	Entropy	−0.9232
Hellinger distance	Entropy	−0.9932
		Pearson
Normalized Kullback–Leibler	Coefficient of variation	0.9872
Kullback–Leibler	Coefficient of variation	0.9832
Jensen–Shannon divergence	Coefficient of variation	0.9678
Generalized Jaccard distance	Coefficient of variation	0.9181
Hellinger distance	Coefficient of variation	0.9649

Table A2. *Cont.*

Measure	Property	Correlation
		Pearson
Normalized Kullback–Leibler	Skewness	0.6343
Kullback–Leibler	Skewness	0.6096
Jensen–Shannon divergence	Skewness	0.6554
Generalized Jaccard distance	Skewness	0.5143
Hellinger distance	Skewness	0.5475
		Pearson
Normalized Kullback–Leibler	Kurtosis	0.4715
Kullback–Leibler	Kurtosis	0.4795
Jensen–Shannon divergence	Kurtosis	0.5170
Generalized Jaccard distance	Kurtosis	0.2622
Hellinger distance	Kurtosis	0.3995

Appendix D. Differences in Ranking Outcomes

Table A3 shows the Spearman rank correlations among the investigated measures on comparing ordered distributions with the uniform one generated by distributing a quantity of 32 to 8 cells.

Table A3. Spearman rank correlations among the investigated measures on comparing ordered distributions with the uniform one generated by distributing a quantity of 32 to 8 cells.

		Spearman
Normalized Kullback–Leibler	Kullback–Leibler	0.9989
Normalized Kullback–Leibler	Jensen–Shannon divergence	0.9909
Normalized Kullback–Leibler	Generalized Jaccard distance	0.9695
Normalized Kullback–Leibler	Hellinger distance	0.9905
Kullback–Leibler	Jensen–Shannon divergence	0.9947
Kullback–Leibler	Generalized Jaccard distance	0.9695
Kullback–Leibler	Hellinger distance	0.9946
Jensen–Shannon divergence	Generalized Jaccard distance	0.9742
Jensen–Shannon divergence	Hellinger distance	1.0000
Hellinger distance	Generalized Jaccard distance	0.9728

References

1. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
2. Arizono, I.; Ohta, H. A test for normality based on Kullback—Leibler information. *Am. Stat.* **1989**, *43*, 20–22.
3. Li, Y.; Wang, L. Testing for homogeneity in mixture using weighted relative entropy. *Commun. Stat. Comput.* **2008**, *37*, 1981–1995. [CrossRef]
4. Belov, D.I.; Armstrong, R.D. Automatic detection of answer copying via Kullback-Leibler divergence and K-index. *Appl. Psychol. Meas.* **2010**, *34*, 379–392. [CrossRef]
5. Clarke, B.S. Asymptotic normality of the posterior in relative entropy. *IEEE Trans. Inf. Theory* **1999**, *45*, 165–176. [CrossRef]
6. Lin, X.; Pittman, J.; Clarke, B. Information conversion, effective samples, and parameter size. *IEEE Trans. Inf. Theory* **2007**, *53*, 4438–4456. [PubMed]
7. Volkau, I.; Prakash, K.B.; Ananthasubramaniam, A.; Aziz, A.; Nowinski, W.L. Extraction of the midsagittal plane from morphological neuroimages using the Kullback–Leibler's measure. *Med. Image Anal.* **2006**, *10*, 863–874. [CrossRef] [PubMed]
8. Ahn, S.; Lee, S.E.; Kim, M.h. Random-forest model for drug–target interaction prediction via Kullback–Leibler divergence. *J. Cheminformatics* **2022**, *14*, 67. [CrossRef] [PubMed]
9. Clim, A.; Zota, R.D.; TinicĂ, G. The Kullback-Leibler divergence used in machine learning algorithms for health care applications and hypertension prediction: A literature review. *Procedia Comput. Sci.* **2018**, *141*, 448–453. [CrossRef]
10. Garg, S.; Dalirrooyfard, M.; Schneider, A.; Adler, Y.; Nevmyvaka, Y.; Chen, Y.; Li, F.; Cecchi, G. Information theoretic clustering via divergence maximization among clusters. In Proceedings of the Uncertainty in Artificial Intelligence, PMLR, Pittsburgh, PA, USA, 31 July–4 August 2023; pp. 624–634.
11. Asperti, A.; Trentin, M. Balancing Reconstruction Error and Kullback-Leibler Divergence in Variational Autoencoders. *arXiv* **2002**, arXiv:2002.07514. [CrossRef]

12. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
13. Rényi, A. On measures of entropy and information. In *Contributions to the Theory of Statistics, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20 June–July 30 1960*; University of California Press: Berkeley, CA, USA, 1961; Volume 1.
14. Sason, I.; Verdú, S. f-divergence Inequalities. *IEEE Trans. Inf. Theory* **2016**, *62*, 5973–6006. [CrossRef]
15. Hellinger, E. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *J. Für Die Reine Und Angew. Math. (Crelles J.)* **1909**, *1909*, 210–271. [CrossRef]
16. Cichocki, A.; Amari, S.i. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568. [CrossRef]
17. Pemmaraju, S.; Skiena, S. *Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica®*; Cambridge University Press: Cambridge, UK, 2003.
18. Pinello, L.; Bosco, G.L.; Hanlon, B.; Yuan, G.C. A motif-independent metric for DNA sequence specificity. *BMC Bioinform.* **2011**, *12*, 408. [CrossRef] [PubMed]
19. Manca, V. *Infobiotics*; Springer: Berlin/Heidelberg, Germany, 2013.
20. Zambelli, F.; Mastropasqua, F.; Picardi, E.; D'Erchia, A.M.; Pesole, G.; Pavesi, G. RNentropy: An entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments. *Nucleic Acids Res.* **2018**, *46*, e46. [CrossRef] [PubMed]
21. Feller, W. *An Introduction to Probability Theory and Its Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2008; Volume 2.
22. Stanley, R.P. *Enumerative Combinatorics*, 2nd ed.; Cambridge Studies in Advanced Mathematics; Cambridge University Press: Cambridge, UK, 2011; Volume 1.
23. Brémaud, P. *An Introduction to Probabilistic Modeling*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
24. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 1991; Volume 68, pp. 69–73.
25. Lee Rodgers, J.; Nicewander, W.A. Thirteen ways to look at the correlation coefficient. *Am. Stat.* **1988**, *42*, 59–66. [CrossRef]
26. Daniel, W.W. *Applied Nonparametric Statistics*; Houghton Mifflin: Boston, MA, USA, 1978.