stats

MDPI

Article

Generalized Mutual Information

Zhiyi Zhang

Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA; zzhang@uncc.edu

Received: 28 April 2020; Accepted: 7 June 2020; Published: 10 June 2020

Abstract: Mutual information is one of the essential building blocks of information theory. It is however only finitely defined for distributions in a subclass of the general class of all distributions on a joint alphabet. The unboundedness of mutual information prevents its potential utility from being extended to the general class. This is in fact a void in the foundation of information theory that needs to be filled. This article proposes a family of generalized mutual information whose members are indexed by a positive integer n, with the nth member being the mutual information of nth order. The mutual information of the first order coincides with Shannon's, which may or may not be finite. It is however established (a) that each mutual information of an order greater than 1 is finitely defined for all distributions of two random elements on a joint countable alphabet, and (b) that each and every member of the family enjoys all the utilities of a finite Shannon's mutual information.

Keywords: mutual information; Shannon's entropy; conditional distribution of total collision; generalized entropy; generalized mutual information

MSC: primary 60E10; secondary 94A15, 82B30

1. Introduction and Summary

This article proposes a family of generalized mutual information whose members are indexed by a positive integer n, with the nth member being the mutual information of nth order. The mutual information of the first order coincides with Shannon's, which may or may not be finite. It is however established that each mutual information of an order greater than 1 is finitely defined for all distributions of two random elements on a joint countable alphabet, and that each and every member of the family enjoys several important utilities of a finite Shannon's mutual information.

Let Z be a random element on a countable alphabet $\mathscr{Z} = \{z_k; k \geq 1\}$ with an associated distribution $\mathbf{p} = \{p_k; k \geq 1\}$. Let the cardinality or support on $\mathscr{Z}$ be denoted $K = \sum_{k\geq 1} 1[p_k > 0]$, where $1[\cdot]$ is the indicator function. K is possibly finite or infinite. Let $\mathscr{P}$ denote the family of all distributions on $\mathscr{Z}$. Let (X, Y) be a pair of random elements on a joint countable alphabet $\mathscr{X} \times \mathscr{Y} = \{(x_i, y_j); i \geq 1, j \geq 1\}$ with an associated joint probability distribution $\mathbf{p}_{X,Y} = \{p_{i,j}; i \geq 1, j \geq 1\}$, let the two marginal distributions be respectively denoted $\mathbf{p}_X = \{p_{i,\cdot} = \sum_{j\geq 1} p_{i,j}; i \geq 1\}$ and $\mathbf{p}_Y = \{p_{\cdot,j} = \sum_{i\geq 1} p_{i,j}; j \geq 1\}$. Let $\mathscr{P}_{X,Y}$ denote the family of all distributions on $\mathscr{X} \times \mathscr{Y}$. Shannon [1] offers two fundamental building blocks of information theory, Shannon's entropy $H = H(Z) = -\sum_{k\geq 1} p_k \log p_k$, where the logarithm is 2-based; and mutual information $MI = MI(X, Y) = H(X) + H(Y) - H(X, Y)$, where $H(X)$, $H(Y)$ and $H(X, Y)$ are entropies respectively defined with the distributions $\mathbf{p}_X$, $\mathbf{p}_Y$ and $\mathbf{p}_{X,Y}$.

Mutual information plays a central role in the theory and the practice of modern data science for three basic reasons. First, the definition of MI does not rely on any metrization on an alphabet, nor does it require the letters of the alphabet to be ordinal. This generality allows it to be defined and used in data spaces beyond the real coordinate space $\mathbb{R}^n$, where random variables (as opposed to random elements) reside. Second, when X and Y are random variables assuming real values, that is, the joint alphabet is metrized, $MI(X,Y)$ captures linear as well as any non-linear stochastic association between X and Y. See Chapter 5 of [2] for examples. Third, it offers a single-valued index measure for the stochastic association between two random elements, more specifically $MI(X,Y) \geq 0$ for any probability distribution of X and Y on a joint alphabet and $MI(X,Y) = 0$ if and only if X and Y are independent, under a wide class of general probability distributions.

However, mutual information MI, in its current form, may not be finitely defined for joint distributions in a subclass of $\mathscr{P}_{X,Y}$, partially due to the fact that any or all of the three Shannon's entropies in the linear combination may be unbounded. The said unboundedness prevents the potential utility of mutual information from being fully realized, and hence there is a deficiency of MI, which leaves a void in $\mathscr{P}_{X,Y}$. (More detailed arguments are provided in Section 2 below). This article introduces a family of generalized mutual information indexed by a positive integer $n \in \mathbb{N}$, denoted $\mathbb{I} = \{MI_n; n \geq 1\}$, each of whose members, MI_n, is referred to as the nth order mutual information. All members of $\mathbb{I}$ are finitely defined for each and every $\mathbf{p}_{X,Y} \in \mathscr{P}_{X,Y}$, except $MI_1 = MI$, and all of them preserve the utilities of Shannon's mutual information when it is finite.

The said deficiency of MI is due to the fact that Shannon's entropy may not be finite for "thick-tailed" distributions (with p_k decaying slowly in k) in $\mathscr{P}$. To address the deficiency of MI, the issue of unboundedness of Shannon's entropy on a subset of $\mathscr{P}$ must be addressed, through some generalization in one way or the other. The effort to generalize Shannon's entropy has been long and extensive in the existing literature. The main perspective in the generalization in the existing literature is based on axiomatic characterization of Shannon's entropy. Interested readers may refer to [3,4] for details and references therewithin. In a nutshell, with respect to the functional form, $H = \sum_{k\geq1} h(p_k)$, under certain desirable axioms, for example, [5,6], $h(p) = -p\log p$ is uniquely determined up to a multiplicative constant; if the strong additivity axiom is relaxed to be one of the weaker versions, say α-additivity or composability, then $h(p)$ may be of other forms, which give rise to Rényi's entropy [7], and the Tsallis entropy [8]. However, all such generalization effort does not seem to lead to an information measure on a joint alphabet that would possess all the desirable properties of MI, in particular $MI(X,Y) = 0$ if and only if X and Y are independent, which is supported by an argument via the Kullback–Leibler divergence [9].

Toward repairing the said deficiency of MI, a new perspective of generalizing Shannon's entropy is introduced in this article. In the new perspective, instead of searching for alternative forms of $h(p)$ in $H = \sum_{k\geq1} h(p_k)$ under weaker axiomatic conditions, it is sought to apply Shannon's entropy not to the original underlying distribution $\mathbf{p}$ but to distributions induced by $\mathbf{p}$. One particular set of such induced distributions is a family, each of whose members is referred to as a conditional distribution of total collision (CDOTC) indexed by $n \in \mathbb{N}$. It is shown that Shannon's entropy defined with every CDOTC induced by any $\mathbf{p} \in \mathscr{P}$ is bounded above, provided that $n \geq 2$. The boundedness of the generalized entropy allows mutual information to be defined for any CDOTC of degree $n \geq 2$ for any $\mathbf{p}_{X,Y} \in \mathscr{P}_{X,Y}$. The resulting mutual information is referred to as the nth order mutual information index and is denoted MI_n, which is shown to possess all the desired properties of MI but with boundedness guaranteed. The main results are given and established in Section 3 after several motivating arguments for the generalization of mutual information in Section 2.

2. Generalization Motivated

To further motivate the generalization of mutual information in this article, let the definition of mutual information be considered in a broader perspective. Inherited from the Kullback–Leibler divergence, mutual information on a joint alphabet, $MI(X,Y) = \sum_{i\geq1,j\geq1} p_{i,j}\log(p_{i,j}/(p_{i,\cdot} \times p_{\cdot,j}))$, is unbounded for a large subclass of distributions in $\mathscr{P}_{X,Y}$. Example 1 below demonstrates the existence of such a subclass of joint distributions.

Example 1. *Let $\mathbf{p} = \{p_k; k \geq 1\}$ be a probability distribution with $p_k > 0$ for every k but unbounded entropy. Let $\mathbf{p}_{X,Y} = \{p_{i,j}; i \geq 1$ and $j \geq 1\}$ be such that $p_{i,j} = p_i$ for all $i = j$ and $p_{i,j} = 0$ for all $i \neq j$, hence $\mathbf{p}_X = \{p_{i,\cdot} = p_i; i \geq 1\}$ and $\mathbf{p}_Y = \{p_{\cdot,j} = p_j; j \geq 1\}$. Then $MI(X,Y) = \sum_{i\geq 1, j\geq 1} p_{i,j} \log(p_{i,j}/(p_{i,\cdot} \times p_{\cdot,j})) = -\sum_{k\geq 1} p_k \log p_k = \infty$.*

One of the most attractive properties of mutual information is that mutual information $MI(X,Y)$ is finitely defined for all joint distributions such that $p_{i,j} = p_{i,\cdot} \times p_{\cdot,j}$ for all $i \geq 1$ and $j \geq 1$ and $MI(X,Y) = 0$ if and only if the two random elements X and Y are independent. However, the utility of mutual information is beyond a mere indication of whether it is zero or not. The magnitude of mutual information is also of essential importance, although Shannon did not elaborate on that in his landmark paper [1]. The said importance is perhaps best illustrated by the notion of the standardized mutual information defined as $\kappa(X,Y) = MI(X,Y)/H(X,Y)$ and Theorem 1 below.

Remark 1. *There are several variants of standardized mutual information proposed in the existing literature. Interested readers may refer to [10–13]. Not all variants of the standardized mutual information have the properties given in Theorem 1. A summary of standardized mutual information is found in Chapter 5 of [2].*

However, before stating Theorem 1, Definition 1 below is needed.

Definition 1. *Random elements $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$ are said to have a one-to-one correspondence, or to be one-to-one corresponded, under a joint probability distribution $\mathbf{p}_{X,Y}$ on $\mathscr{X} \times \mathscr{Y}$, if:*

1. *for every i satisfying $\mathrm{P}(X = x_i) > 0$, there exists a unique j such that $\mathrm{P}(Y = y_j | X = x_i) = 1$, and*
2. *for every j satisfying $\mathrm{P}(Y = y_j) > 0$, there exists a unique i such that $\mathrm{P}(X = x_i | Y = y_j) = 1$.*

Theorem 1. *Let (X,Y) be a pair of random elements on alphabet $\mathscr{X} \times \mathscr{Y}$ with joint distribution $\mathbf{p}_{X,Y} \in \mathscr{P}_{X,Y}$ such that $H(X,Y) < \infty$. Then:*

1. *$0 \leq \kappa(X,Y) \leq 1$,*
2. *$\kappa(X,Y) = 0$ if and only if X and Y are independent, and*
3. *$\kappa(X,Y) = 1$ if and only if X and Y are one-to-one corresponded.*

A proof of Theorem 1 can be found on page 159 of [2]. Theorem 1 essentially maps the independence of X and Y (the strongest form of unrelatedness) to $\kappa = 0$, one-to-one correspondence (the strongest form of relatedness) to $\kappa = 1$, and everything else in between. In so doing, the magnitude of mutual information is utilized in measuring the degree of dependence in pairs of random elements, which could lead to all sorts of practical tools for evaluating, ranking, and selecting variables in data space.

It is important to note that the condition of $H(X,Y) < \infty$ is essential in Theorem 1, since obviously, without it, κ may not be well defined. In fact, if $H(X,Y) < \infty$ is not imposed, and even observing reasonable conventions such as $1/\infty = 0$ and $0/\infty = 0$, the statements of Theorem 1 may not be true. To see this, consider the following constructed example.

Example 2. *Let $\mathbf{p} = \{p_k; k \geq 1\}$ be a probability distribution with $p_k > 0$ for every k but unbounded entropy. Let $\mathbf{p}_{X,Y} = \{p_{i,j}; i = 1$ or 2 and $j \geq 1\}$ be such that*

$$p_{i,j} = \begin{cases} p_j & i = 1 \text{ and } j \text{ is odd} \\ p_j & i = 2 \text{ and } j \text{ is even} \\ 0 & \text{otherwise,} \end{cases}$$

hence $\mathbf{p}_X = \{p_{1,\cdot}, p_{2,\cdot}\} = \{\sum_{k=odd} p_k, \sum_{k=even} p_k\}$ *and* $\mathbf{p}_Y = \{p_{\cdot,j} = p_j; j \geq 1\}$*. X and Y are obviously not independent, and*

$$0 < MI(X,Y) = \sum_{i\geq 1, j\geq 1} p_{i,j} \log(p_{i,j}/(p_{i,\cdot} \times p_{\cdot,j})) = H(X) < \infty.$$

It follows that $\kappa = MI(X,Y)/H(X,Y) = H(X)/H(X,Y) = 0$ *but in this case* $MI(X,Y) > 0$*. Therefore Part 2 of Theorem 1 fails.*

Example 2 indicates that mutual information in its current form is deprived of the potential utility of Theorem 1 for a large class of joint distributions and therefore leaves much to be desired.

Another argument for the generalization of mutual information can be made in a statistical perspective. In practice, mutual information is often to be estimated from sample data. For statistical inference to be meaningful, the estimand $MI(X,Y)$ needs to exist, i.e., $MI(X,Y) < \infty$. More specifically, in testing the hypothesis of independence between X and Y, $H_0 : \mathbf{p}_{X,Y} \in \mathscr{P}_0$, where $\mathscr{P}_0 \subset \mathscr{P}_{X,Y}$ is the subclass of all joint distributions for independent X and Y on $\mathscr{X} \times \mathscr{Y}$, and $MI(X,Y)$ needs to be finitely defined in an open neighborhood of $\mathscr{P}_0$ in $\mathscr{P}_{X,Y}$, or else the logic framework of statistical inference is not well supported. Let $\mathscr{P}_\infty$ be the subclass of $\mathscr{P}_{X,Y}$ such that $MI(X,Y) = \infty$. In general, it can be shown that $\mathscr{P}_\infty$ is dense in $\mathscr{P}_{X,Y}$ with respect to the p-norm for $p \geq 1$. Specifically, for any $\mathbf{p}_{X,Y} \in \mathscr{P}_0$, there exists a sequence of distributions $\{\mathbf{p}_{m,X,Y}\} \in \mathscr{P}_\infty$ such that $\|\mathbf{p}_{m,X,Y} - \mathbf{p}_{X,Y}\|_p \to 0$. See Example 3 below.

Example 3. *Let* $\mathbf{p}_{X,Y} = \{p_{i,j}; i = 1,2 \text{ and } j = 1,2\}$ *where* $p_{i,j} = 0.25$ *for all* (i,j) *such that* $1 \leq i \leq 2$ *and* $1 \leq j \leq 2$*. Obviously X and Y are independent under* $\mathbf{p}_{X,Y}$*, that is,* $\mathbf{p}_{X,Y} \in \mathscr{P}_0$*. Let* $\mathbf{p}_{m,X,Y}$ *be constructed based on* $\mathbf{p}_{X,Y}$ *as follows.*

Remove an arbitrarily small quantity $\varepsilon/4 > 0$ *where* $\varepsilon = 1/m$ *away from each of the four positive probabilities in* $\mathbf{p}_{X,Y}$ *so each becomes* $p_{m,i,j} = 0.25 - \varepsilon/4$ *for all* (i,j)*, such that* $1 \leq i \leq 2$ *and* $1 \leq j \leq 2$*. Extend the range of* (i,j) *to* $i \geq 3$ *and* $j \geq 3$*, and allocate the mass* ε *over the extended range according to*

$$p_{m,i,j} = \begin{cases} \dfrac{c}{i(\log i)^2} & i \geq 3, j \geq 3 \text{ and } i = j \\ 0 & i \geq 3, j \geq 3 \text{ and } i \neq j, \end{cases}$$

where c is such that $\sum_{k\geq 3} c/[k(\log k)^2] = \varepsilon$*. Under the constructed* $\{p_{m,i,j}\}$*, for any* $\varepsilon = 1/m$*, X and Y are not independent, and the corresponding mutual information is*

$$\begin{aligned}&\sum_{i\geq 1, j\geq 1} p_{m,i,j} \log\left[\frac{p_{m,i,j}}{(p_{m,i,\cdot}\, p_{m,\cdot,j})}\right] \\ &= 4(0.25 - \varepsilon/4)\log\left[\frac{0.25 - \varepsilon/4}{(0.5 - \varepsilon/2)^2}\right] - \sum_{k\geq 3} \frac{c}{k(\log k)^2} \log\frac{c}{k(\log k)^2} = \infty.\end{aligned}$$

However, noting that as $m \to \infty$*,* $\varepsilon \to 0$ *and hence* $c \to 0$*,*

$$\|\mathbf{p}_{m,X,Y} - \mathbf{p}_{X,Y}\|_2^2 = 4\varepsilon^2 + \sum_{k\geq 3}\left[\frac{c}{k(\log k)^2}\right]^2 = 4\varepsilon^2 + c^2 \sum_{k\geq 3} \frac{1}{k^2(\log k)^4} \to 0.$$

All things considered, it is therefore desirable to have a mutual information measure, say $MI_n(X,Y)$, or for that matter a family of mutual information measures indexed by a positive integer n, such that $MI_n(X,Y) < \infty$ for all distributions in $\mathscr{P}_{X,Y}$, and with an accordingly defined standardized mutual information measure $\kappa_n = \kappa_n(X,Y)$ such that the utility of Theorem 1 is preserved with κ_n in place of κ for all distributions in $\mathscr{P}_{X,Y}$.

3. Main Results

Given $\mathscr{Z} = \{z_k; k \geq 1\}$ and $\mathbf{p} = \{p_k\}$, consider the experiment of drawing an identically and independently distributed (*iid*) sample of size n. Let C_n denote the event that all observations of the sample take on a same letter in $\mathscr{Z}$, and let C_n be referred to as the event of total collision. The conditional probability, given C_n, that the total collision occurs at letter z_k is

$$p_{n,k} = \frac{p_k^n}{\sum_{i\geq 1} p_i^n}. \tag{1}$$

It is clear that $\mathbf{p}_n = \{p_{n,k}\}$ is a probability distribution induced from $\mathbf{p} = \{p_k\}$. For each n, $p_{n,k}$ of (1) is the conditional distribution of total collision (CDOTC) with n particles.

Remark 2. *It is to be noted that, given a* $\mathbf{p}$, $\mathbf{p}_n = \{p_{n,k}; k \geq 1\}$ *of (1) is a special member of the family of the escort distributions introduced by [14]. The escort distributions are a useful tool in thermodynamics. Interested readers may refer to [15] for a concise introduction.*

Lemma 1. *For each* n, $n \geq 1$, $\mathbf{p}$ *and* $\mathbf{p}_n$ *uniquely determine each other.*

Proof. Given $\mathbf{p} = \{p_k; k \geq 1\}$, by (1), $\mathbf{p}_n = \{p_{n,k}; \geq 1\}$ is uniquely determined. Conversely, given $\mathbf{p}_n = \{p_{n,k}; \geq 1\}$, for each n and all $k \geq 1$, $p_k^n / p_1^n = p_{n,k}/p_{n,1}$ and therefore

$$p_k = p_1 \left(\frac{p_{n,k}}{p_{n,1}}\right)^{1/n}, \quad \sum_{i\geq 1} p_i = p_1 \sum_{i\geq 1} \left(\frac{p_{n,i}}{p_{n,1}}\right)^{1/n} = 1, \quad p_1 = \left[\sum_{i\geq 1} \left(\frac{p_{n,i}}{p_{n,1}}\right)^{1/n}\right]^{-1},$$

$$p_k = \left[\sum_{i\geq 1} \left(\frac{p_{n,i}}{p_{n,1}}\right)^{1/n}\right]^{-1} \left(\frac{p_{n,k}}{p_{n,1}}\right)^{1/n} = \left[\sum_{i\geq 1} \left(\frac{p_{n,i}}{p_{n,k}}\right)^{1/n}\right]^{-1} = \frac{p_{n,k}^{1/n}}{\sum_{i\geq 1} p_{n,i}^{1/n}}. \tag{2}$$

The lemma follows. □

Lemma 2. *For each* n, $n \geq 2$, *and for any* $\mathbf{p} \in \mathscr{P}$, $H_n(Z) = -\sum_{k\geq 1} p_{n,k} \ln p_{n,k} < \infty$.

Proof. Write $\eta_n = \sum_{k\geq 1} p_k^n$. Noting $0 < \eta_n \leq 1$, $0 \leq -p \ln p \leq 1/e$ and therefore $-p \log p \leq 1/(e \log 2)$ for all $p \in [0,1]$,

$$\begin{aligned} H_n(Z) &= -\sum_{k\geq 1} p_{n,k} \log p_{n,k} = -\sum_{k\geq 1} \frac{p_k^n}{\sum_{i\geq 1} p_i^n} \log \frac{p_k^n}{\sum_{i\geq 1} p_i^n} \\ &= -\frac{n}{\eta_n} \sum_{k\geq 1} p_k^n \log p_k + \log \eta_n \leq \left(\frac{n}{e \log 2}\right)\left(\frac{\eta_{n-1}}{\eta_n}\right) + \log \eta_n < \infty. \end{aligned}$$

The lemma follows. □

On the joint alphabet $\mathscr{X} \times \mathscr{Y} = \{(x_i, y_j)\}$ with distribution $\mathbf{p}_{X,Y} = \{p_{i,j}\}$, consider the associated CDOTC for an n and all pairs (i,j) such that $i \geq 1$ and $j \geq 1$,

$$p_{n,i,j} = \frac{p_{i,j}^n}{\sum_{s\geq 1, t\geq 1} p_{s,t}^n}. \tag{3}$$

Let $\mathbf{p}_{n,X,Y} = \{p_{n,i,j}; i \geq 1, j \geq 1\}$. It is to be noted that $\mathbf{p}_{n,X,Y} \in \mathscr{P}_{X,Y}$. The two marginal distributions of (3) are $\mathbf{p}_{n,X} = \{p_{n,i,\cdot}\}$ and $\mathbf{p}_{n,Y} = \{p_{n,\cdot,j}\}$, respectively, where

$$p_{n,i,\cdot} = \sum_{j\geq 1} p_{n,i,j} = \sum_{j\geq 1} \left(\frac{p_{i,j}^n}{\sum_{s\geq 1,t\geq 1} p_{s,t}^n} \right) = \frac{\sum_{j\geq 1} p_{i,j}^n}{\sum_{s\geq 1,t\geq 1} p_{s,t}^n}, \tag{4}$$

$$p_{n,\cdot,j} = \sum_{i\geq 1} p_{n,i,j} = \sum_{i\geq 1} \left(\frac{p_{i,j}^n}{\sum_{s\geq 1,t\geq 1} p_{s,t}^n} \right) = \frac{\sum_{i\geq 1} p_{i,j}^n}{\sum_{s\geq 1,t\geq 1} p_{s,t}^n}. \tag{5}$$

Lemma 3. $\mathbf{p}_{X,Y} = \{p_{i,j}\} = \{p_{i,\cdot} \times p_{\cdot,j}\}$ *if and only if* $\mathbf{p}_{n,X,Y} = \{p_{n,i,j}\} = \{p_{n,i,\cdot} \times p_{n,\cdot,j}\}$.

Proof. For each positive integer n, if $p_{i,j} = p_{i,\cdot} \times p_{\cdot,j}$ for all pairs (i,j), $i \geq 1$ and $j \geq 1$, then

$$p_{n,i,j} = \frac{p_{i,j}^n}{\sum_{s\geq 1,t\geq 1} p_{s,t}^n} = \frac{p_{i,\cdot}^n . p_{\cdot,j}^n}{\sum_{s\geq 1,t\geq 1} p_{s,\cdot}^n . p_{\cdot,t}^n} = \left(\frac{p_{i,\cdot}^n}{\sum_{s\geq 1} p_{s,\cdot}^n} \right) \left(\frac{p_{\cdot,j}^n}{\sum_{t\geq 1} p_{\cdot,t}^n} \right)$$

where the two factors of the last expression above are respectively $\mathrm{P}(X_1 = \cdots = X_n = x_i | C_n)$ and $\mathrm{P}(Y_1 = \cdots = Y_n = y_j | C_n)$, (X_r, Y_r), $r = 1, \cdots, n$, are letter values of the n observations in the sample.

Conversely, if $p_{n,i,j} = p_{n,i}^* \times p_{n,j}^*$ where $p_{n,i}^* \geq 0$ depends only on n and i and $p_{n,j}^* \geq 0$ only depends on n and j, then by (2),

$$p_{i,j} = \frac{p_{n,i,j}^{1/n}}{\sum_{s\geq 1,t\geq 1} p_{n,s,t}^{1/n}} = \frac{(p_{n,i}^*)^{1/n} (p_{n,j}^*)^{1/n}}{\sum_{s\geq 1} (p_{n,s}^*)^{1/n} \sum_{t\geq 1} (p_{n,t}^*)^{1/n}}$$
$$= \left(\frac{(p_{n,i}^*)^{1/n}}{\sum_{s\geq 1} (p_{n,s}^*)^{1/n}} \right) \times \left(\frac{(p_{n,j}^*)^{1/n}}{\sum_{t\geq 1} (p_{n,t}^*)^{1/n}} \right).$$

The lemma immediately follows the factorization theorem. □

For each $n \in \mathbb{N}$, let $H_n(X,Y)$, $H_n(X)$ and $H_n(Y)$ be Shannon's entropies defined with the joint CDOTC, $\{p_{n,i,j}; i \geq 1\}$ as in (3), and the marginal distributions $\{p_{n,i,\cdot}; i \geq 1\}$ and $\{p_{n,\cdot,j}; j \geq 1\}$ as in (4) and (5), respectively. Let

$$MI_n = MI_n(X,Y) = H_n(X) + H_n(Y) - H_n(X,Y). \tag{6}$$

Theorem 2. *For every* $n \geq 2$ *and any* $\mathbf{p}_{X,Y} \in \mathscr{P}_{X,Y}$,

1. $0 \leq MI_n(X,Y) < \infty$,
2. $MI_n(X,Y) = 0$ *if and only X and Y are independent.*

Proof. In Part 1, $MI_n \geq 0$, since MI_n is a mutual information and $MI_n < \infty$ by Lemma 2. Part 2 follows Lemma 3 and the fact that MI_n is a mutual information. □

Let

$$\kappa_n = \kappa_n(X,Y) = \frac{H_n(X) + H_n(Y) - H_n(X,Y)}{H_n(X,Y)} \tag{7}$$

be referred to as the nth order standardized mutual information, and write $\mathbb{I}_S = \{\kappa_n; n \geq 1\}$. Let (X^*, Y^*) be a pair of random elements on $\mathscr{X} \times \mathscr{Y}$ according to the induced joint distribution $\mathbf{p}_{n,X,Y}$ with index value $n \geq 1$.

Lemma 4. *X and Y have a one-to-one correspondence if and only if* X^* *and* Y^* *have one.*

Proof. If X and Y have a one-to-one correspondence, then for each i, there is a unique j_i such that $p_{i,j_i} > 0$ and $p_{i,j} = 0$ for all other j, $j \neq j_i$. By (3), $p_{n,i,j_i} > 0$ and $p_{n,i,j} = 0$ for all other j, $j \neq j_i$. That is, X^* and Y^* have a one-to-one correspondence.

Conversely, if X^* and Y^* have a one-to-one correspondence, then for each i, there is a unique j_i such that $p_{n,i,j_i} > 0$ and $p_{n,i,j} = 0$ for all other j, $j \neq j_i$. On the other hand, by (2),

$$p_{i,j} = \frac{p_{n,i,j}^{1/n}}{\sum_{s \geq 1, t \geq 1} p_{n,s,t}^{1/n}},$$

it follows that $p_{i,j_i} > 0$ and $p_{i,j} = 0$ for all other j, $j \neq j_i$. That is, X and Y have a one-to-one correspondence. □

Corollary 1. *For every $n \geq 2$ and any $\mathbf{p}_{X,Y} \in \mathscr{P}_{X,Y}$,*

1. $0 \leq \kappa_n(X,Y) \leq 1$,
2. $\kappa_n(X,Y) = 0$ *if and only if X and Y are independent, and*
3. $\kappa_n(X,Y) = 1$ *if and only if X and Y are one-to-one corresponded.*

Proof. By Lemma 3, X and Y are independent if and only if X^* and Y^* are. By Lemma 4, X and Y are one-to-one corresponded if and only if X^* and Y^* are. The statement of Corollary 1 follows directly from Theorem 1. □

Theorem 2 and Corollary 1 together fill the void in $\mathscr{P}_{X,Y}$ left behind by *MI*.

4. Concluding Remarks

The main results of this article may be summarized as follows. A family of generalized mutual information indexed by a positive integer n is proposed. The member corresponding to $n = 1$ is Shannon's mutual information for a given joint distribution, $\mathbf{p}_{X,Y}$. The other members of the family correspond to other integer values of n. They are also Shannon's information defined, not with $\mathbf{p}_{X,Y}$, but with induced distributions based on the given distribution $\mathbf{p}_{X,Y}$. These induced distributions are called conditional distributions of total collision (CDOTC), which collectively is a special subset of a more general family called the escort distributions, which is often studied in extensive thermodynamics. The main motivation of the generalized mutual information is to resolve the issue of the fact that the standard mutual information is not finitely defined for all distributions of a countable joint alphabet $\mathscr{P}_{X,Y} = \{\text{all probability distributions on } \mathscr{X} \times \mathscr{Y}\}$, which leads to the issue of mutual information's utility only realized on a fraction of $\mathscr{P}$.

On a more specific and finer level, the following facts are established.

1. There is a one-to-one correspondence between each CDOTC and the given distribution $\mathbf{p}$ on a countable alphabet, and hence each CDOTC is a characteristic representation of the original distribution $\mathbf{p}$. One of the implications of this fact is that understanding the underlying $\mathbf{p}$ is equivalent to understanding one of its CDOTC. It can be shown that the CDOTC with an order greater than 1 is much easier to estimate than $\mathbf{p}$ with sparse data.
2. Each generalized mutual information is guaranteed to be finite. This result essentially guarantees the validity of statistically testing the null hypothesis of independence of two discrete random elements, as it guarantees the existence of (generalized) mutual information anywhere in the alternative space of dependent join distributions.
3. It is shown that a particular form of standardized mutual information κ, defined with any CDOTC of any order greater than 1, preserves the zero-to-one scale with independence on one end and total dependence on the other, which is enjoyed by Shannon's entropy only when it is finite.

In short, the family of conditional distributions of total collision embeds the underlying probability distribution $\mathbf{p}$ as a special member, and the family of generalized mutual information embeds Shannon's mutual information as a special member. Consequently, the stochastic association on joint alphabets can be measured by not only one index but by a host of indices, which collectively offer a much extended space to study stochastic dependence in information theory.

Funding: This research received no external funding

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

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656.
2. Zhang, Z. *Statistical Implications of Turing's Formula*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2017.
3. Csiszár, I. Axiomatic characterizations of information measures. *Entropy* **2008**, *10*, 261–273; doi:10.3390/e10030261.
4. Amigó, J.M.; Balogh, S.G.; Hernández, S. A brief review of generalized entropies. *Entropy* **2018**, *20*, 813.
5. Khinchin, A.I. *Mathematical Foundations of Information Theory*; Dover: New York, NY, USA, 1957.
6. Chakrabarti, C.G.; Chakrabarty, I. Shannon entropy: Axiomatic characterization and application. *Int. J. Math. Math. Sci.* **2005**, *17*, 2847–2854, doi:10.1155/IJMMS.2005.2847.
7. Rényi, A. On measures of information and entropy. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.
8. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.* **1988**, *52*, 479–487.
9. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
10. Kvalseth, T.O. Entropy and correlation: Some comments. *IEEE Trans. Syst. Man Cybern.* **1987**, *17*, 517–519.
11. Strehl, A.; Ghosh, J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.
12. Yao, Y.Y. Information-theoretical measures for knowledge discovery and data mining. In *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, 1st ed.; Karmeshu, Ed.; Springer: Berlin, Germany, 2003; pp. 115–136.
13. Vinh, N.X.; Epps, J.; Bailey, J. Information theoretical measures for clustering comparison: Variants, properties, normalization and correlation for chance. *J. Mach. Learn. Res.* **2018**, *11*, 2837–2854.
14. Beck, C.; Schlögl, F. *Thermodynamics of Chaotic Systems: An Introduction*; Cambridge University Press: Cambridge, UK, 1993.
15. Matsuzoe, H. A sequence of escort distributions and generalizations of expectations on q-exponential family. *Entropy* **2017**, *19*, 7, doi:10.3390/e19010007.