An Axiomatic Characterization of Mutual Information

James Fullwood

School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, China; fullwood@sjtu.edu.cn

Abstract: We characterize mutual information as the unique map on ordered pairs of discrete random variables satisfying a set of axioms similar to those of Faddeev’s characterization of the Shannon entropy. There is a new axiom in our characterization, however, which has no analog for Shannon entropy, based on the notion of a Markov triangle, which may be thought of as a composition of communication channels for which conditional entropy acts functorially. Our proofs are coordinate-free in the sense that no logarithms appear in our calculations.

Keywords: Shannon theory; information measures; mutual information

1. Introduction

Axiomatic characterizations of information measures go back to the seminal work of Shannon [1], providing conceptual insights into their meaning as well as justification for the analytic formulae involved in their definitions. Various characterizations for Shannon entropy, relative entropy, Renyi and Tsallis entropies, von Neumann and Segal entropies, quantum relative entropy, as well as other generalized information measures have appeared in the literature [2–9], and a review of such enterprises in the classical (i.e., non-quantum) setting appears in the survey of Csiszár [10]. More recently, functorial characterizations of information measures from a categorical viewpoint have appeared in the works of Baez, Fritz, and Leinster [11,12], as well as our work with Parzygnat [13], who have proven a functorial characterization of the von Neumann entropy [14]. An axiomatic approach to entropy in the theory of biodiversity is the subject of the recent book [15] by Leinster.

In spite of the breadth of the aforementioned results, the mutual information of a pair of random variables seems to be missing from the story. While an operational characterization of mutual information in the context of algorithmic information theory appears in [16], to the best of our knowledge, an axiomatic characterization in the vein of those surveyed by Csiszár in [10] is absent in the literature. It is then the goal of the present work to introduce mutual information into the axiomatic framework.

Our main result is Theorem 1, where we prove that the mutual information $I(X, Y)$ of an ordered pair of random variables is the unique function (up to an arbitrary multiplicative factor) on pairs of random variables satisfying the following axioms:

1. **Continuity**: If $(X_n, Y_n) \to (X, Y)$, then $I(X_n, Y_n) \to I(X, Y)$.
2. **Strong Additivity**: Given a random variable $X : \Omega \to X$ with probability mass function $p : X \to [0, 1]$, and a collection of pairs of random variables $(Y^n, Z^n)$ indexed by $X$, then
   $$I\left( \bigoplus_{x \in X} p(x)(Y^n, Z^n) \right) = I(X, X) + \sum_{x \in X} p(x)I(Y^n, Z^n).$$
3. **Symmetry**: $I(X, Y) = I(Y, X)$ for every pair of random variables $(X, Y)$.
4. **Invariance Under Pullbacks**: If $\pi : \Omega' \to \Omega$ is a measure-preserving function, then for every pair of random variables $(X, Y)$ with common domain $\Omega$,
   $$I(X, Y) = I(X \circ \pi, Y \circ \pi).$$
5. **Weak Functoriality:** For every Markov triangle \( (X, Y, Z) \),

\[
\mathbb{I}(X, Z) = \mathbb{I}(X, Y) + \mathbb{I}(Y, Z) - \mathbb{I}(Y, Y).
\]

6. **Vacuity:** If \( C \) is a constant random variable, then \( \mathbb{I}(X, C) = 0 \).

The fact that mutual information satisfies Axioms 1, 3, and 6 is well known to anybody familiar with mutual information. As we work at the level of random variables as opposed to simple probability distributions (which we do for wider applicability of our results), Axiom 4 is a reflection of the fact that mutual information only depends on probabilities. For Axiom 2, we define a convex structure on pairs of random variables in such a way that the strong additivity of Shannon entropy is generalized to our context. Axiom 5 is defined in terms of the notion of a Markov triangle, a concept we define based on the notion of a “coalescable” composition of communication channels which was introduced in [13].

Intuitively, a Markov triangle may be thought of as a composition of noisy channels over which the associated conditional entropy is additive. Moreover, such axioms are sharp in the sense that if any of the axioms are removed, then mutual information may not be characterized. In particular, the joint entropy \( H(X, Y) \) satisfies Axioms 1–5, while the conditional entropy \( H(Y|X) \) satisfies all the axioms except the symmetry Axiom 3 (note that since \( H(X, X) = 0 \), Axiom 2 in the case of conditional entropy becomes convex linearity).

In the spirit of the axiomatic approach, we note that logarithms are absent from all calculations in this paper.

2. **Mutual Information**

Let \((\Omega, \Sigma, \mu)\) be a probability space, where \(\Omega\) is thought of as the set of all possible outcomes of a data generating process or experiment, \(\Sigma\) is a \(\sigma\)-algebra of measurable subsets of \(\Omega\), and \(\mu\) is a probability measure.

**Definition 1.** A finite random variable is a surjective function \(X : \Omega \to \mathcal{X}\) such that \(\mathcal{X}\) is a finite set and \(X^{-1}(x) \in \Sigma\) for all \(x \in \mathcal{X}\). In such a case, the set \(\mathcal{X}\) is often referred to as the support, or alphabet associated with \(X\). The probability mass function of \(X\) is the function \(p : \mathcal{X} \to [0, 1]\) given by

\[
p(x) = \mu\left(X^{-1}(x)\right),
\]

and the Shannon entropy of \(X\) is the non-negative real number \(H(X)\) given by

\[
H(X) = -\sum_{x \in \mathcal{X}} p(x) \log(p(x)).
\]

The collection of all finite random variables on \(\Omega\) will be denoted \(\text{FRV}(\Omega)\).

**Definition 2.** Let \((X, Y) \in \text{FRV}(\Omega) \times \text{FRV}(\Omega)\) be an ordered pair of random variables with supports \(\mathcal{X}\) and \(\mathcal{Y}\) respectively.

- The joint distribution function of \((X, Y)\) is the function \(\theta : \mathcal{X} \times \mathcal{Y} \to [0, 1]\) given by

\[
\theta(x, y) = \mu\left(X^{-1}(x) \cap Y^{-1}(y)\right),
\]

- The joint entropy of \((X, Y)\) is the non-negative real number given by

\[
H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \theta(x, y) \log(\theta(x, y)),
\]

- The mutual information of \((X, Y)\) is the real number \(\mathbb{I}(X, Y)\) given by

\[
\mathbb{I}(X, Y) = H(X) + H(Y) - H(X, Y).
\]
Remark 1. With every pair of random variables \((X, Y)\) one may associate a probability transition matrix \(p(y|x)\) given by

\[
p(y|x) = \frac{\theta(x, y)}{p(x)},
\]

where \(p : X \rightarrow [0, 1]\) is the probability mass function of \(X\). As such, one may view \((X, Y)\) as a noisy channel \(X \sim Y\) together with the prior distribution \(p\) on its set of inputs.

We now list some well-known properties of mutual information which will be useful for our purposes (see, e.g., [17] for proofs).

Proposition 1. Mutual information satisfies the following properties.

i. \(\mathbb{I}(X, Y) \geq 0\) for all \((X, Y) \in \text{FRV}(\Omega) \times \text{FRV}(\Omega)\).

ii. \(\mathbb{I}(Y, X) = \mathbb{I}(Y, X)\) for all \((X, Y) \in \text{FRV}(\Omega) \times \text{FRV}(\Omega)\).

iii. \(\mathbb{I}(X, X) = H(X)\) for all \(X \in \text{FRV}(\Omega)\).

iv. \(\mathbb{I}(X, C) = 0\) for every constant random variable \(C \in \text{FRV}(\Omega)\).

Definition 3. The canonical product on \(\text{FRV}(\Omega)\) is the map \(\mathcal{P} : \text{FRV}(\Omega) \times \text{FRV}(\Omega) \rightarrow \text{FRV}(\Omega)\), given by \(\mathcal{P}(X, Y)(\omega) = (X(\omega), Y(\omega)) \in X \times Y\) for all \(\omega \in \Omega\).

Proposition 2. Let \((X, Y) \in \text{FRV}(\Omega) \times \text{FRV}(\Omega)\). Then the following statements hold.

i. The probability mass function of \(\mathcal{P}(X, Y)\) is the joint distribution function \(\theta(x, y)\). In particular, \(H(X, Y) = H(\mathcal{P}(X, Y))\).

ii. \(\mathbb{I}(X, \mathcal{P}(X, Y)) = H(X)\).

Proof.

i. Let \(v : X \times Y \rightarrow [0, 1]\) denote the probability mass function of \(\mathcal{P}(X, Y)\). Then for all \((x, y) \in X \times Y\) we have \(\mathcal{P}(X, Y)^{-1}(x, y) = X^{-1}(x) \cap Y^{-1}(y)\), thus

\[
v(x, y) = \mu(\mathcal{P}(X, Y)^{-1}(x, y)) = \mu(X^{-1}(x) \cap Y^{-1}(y)) = \theta(x, y),
\]

as desired.

ii. The statement follows from the fact that \(H(X, \mathcal{P}(X, Y)) = H(X)\). \(\Box\)

3. Convexity

We now generalize the notion of a convex combination of probability distributions to the setting of pairs of random variables, which will be used to extend the notion of strong additivity for Shannon entropy to mutual information.

Notation 1. We use the notation \(X \bigsqcup Y\) to denote the disjoint union of the sets \(X\) and \(Y\).

Definition 4. Let \(X\) be a finite set, and let \(p : X \rightarrow [0, 1]\) be a probability distribution on \(X\). Then \(\bigoplus_{x \in X} p(x)\) is the probability space associated with the triple \((X \times \Omega, X \times \Sigma, p \times \mu)\). Now suppose \(Y^x \in \text{FRV}(\Omega)\) is a collection of random variables indexed by \(X\), and let \(q^x : [0, 1] \rightarrow [0, 1]\) denote the probability mass function of \(Y^x\). The \(p\)-weighted convex sum \(\bigoplus_{x \in X} p(x)Y^x \in \text{FRV}(X \times \Omega)\) is the random variable given by

\[
(\bigoplus_{x \in X} p(x)Y^x)(x, \omega) = Y^x(\omega).
\]

It then follows that the probability mass function of \(\bigoplus_{x \in X} p(x)Y^x\) is a function of the form \(r : \bigsqcup_{x \in X} Y^x \rightarrow [0, 1]\), and using the fact that \(\bigsqcup_{x \in X} Y^x\) is canonically isomorphic to the set 

\[\{(x, y) \mid x \in X \text{ and } y \in Y^x\},\]
it follows that \( r \) is then given by \( r(x, y) = p(x)q'(y) \).

A reformulation of the strong additivity property for Shannon entropy in terms of the convex structure just introduced for random variables is given using the following proposition.

**Proposition 3.** Let \( \mathcal{X} \) be a finite set, let \( p : \mathcal{X} \to [0, 1] \) be a probability distribution on \( \mathcal{X} \), and suppose \( Y^X \in \text{FRV}(\Omega) \) is a collection of random variables indexed by \( \mathcal{X} \). Then

\[
H\left( \bigoplus_{x \in \mathcal{X}} p(x)Y^x \right) = H(p) + \sum_{x \in \mathcal{X}} p(x)H(Y^x),
\]

where \( H(p) \) is the Shannon entropy of the probability distribution \( p \).

**Proof.** Let \( (\tilde{x}, \omega) \in \mathcal{X} \times \Omega \). Then

\[
\left( \bigoplus_{x \in \mathcal{X}} p(x)\mathcal{P}(Y^x, Z^x) \right)(\tilde{x}, \omega) \overset{(1)}{=} \mathcal{P}(Y^\tilde{x}, Z^\tilde{x})(\omega) = (Y^\tilde{x}(\omega), Z^\tilde{x}(\omega))
\]

\[
\overset{(1)}{=} \left( \bigoplus_{x \in \mathcal{X}} p(x)Y^x(\tilde{x}, \omega), \bigoplus_{x \in \mathcal{X}} p(x)Z^x(\tilde{x}, \omega) \right)
\]

\[
= \mathcal{P}\left( \bigoplus_{x \in \mathcal{X}} p(x)Y^x, \bigoplus_{x \in \mathcal{X}} p(x)Z^x \right)(\tilde{x}, \omega),
\]

thus Equation (3) holds. \( \square \)

In light of Proposition 4, we make the following definition.

**Definition 5.** Let \( \mathcal{X} \) be a finite set, let \( p : \mathcal{X} \to [0, 1] \) be a probability distribution on \( \mathcal{X} \), and suppose \( (Y^X, Z^X) \in \text{FRV}(\mathcal{X} \times \Omega) \) is a collection of pairs of random variables indexed by \( \mathcal{X} \). The \( p \)-weighted convex sum \( \bigoplus_{x \in \mathcal{X}} p(x)(Y^x, Z^x) \in \text{FRV}(\mathcal{X} \times \Omega) \) is defined to be the ordered pair \( (\bigoplus_{x \in \mathcal{X}} p(x)Y^x, \bigoplus_{x \in \mathcal{X}} p(x)Z^x) \).

**Proposition 5** (Strong Additivity of Mutual Information). Let \( \mathcal{X} \) be a finite set, let \( p : \mathcal{X} \to [0, 1] \) be a probability distribution on \( \mathcal{X} \), and suppose \( (Y^X, Z^X) \in \text{FRV}(\mathcal{X} \times \Omega) \) is a collection of pairs of random variables indexed by \( \mathcal{X} \). Then

\[
\mathbb{I}\left( \bigoplus_{x \in \mathcal{X}} p(x|(Y^x, Z^x) \right) = H(p) + \sum_{x \in \mathcal{X}} p(x)\mathbb{I}(Y^x, Z^x),
\]

where \( H(p) \) is the Shannon entropy of the probability distribution \( p \).

**Proof.** Indeed,
\[
\begin{align*}
\mathbb{I}(\bigoplus_{x \in X} p(x)(Y^x, Z^x)) &= \mathbb{I}(\bigoplus_{x \in X} p(x)Y^x, \bigoplus_{x \in X} p(x)Z^x) \\
&= H(\bigoplus_{x \in X} p(x)Y^x) + H(\bigoplus_{x \in X} p(x)Z^x) - H(\bigoplus_{x \in X} p(x)Y^x, \bigoplus_{x \in X} p(x)Z^x) \\
&\geq 2H(p) + \sum_{x \in X} p(x)(H(Y^x) + H(Z^x)) - (H(p) + \sum_{x \in X} p(x)H(Y^x, Z^x)) \\
&= H(p) + \sum_{x \in X} p(x)\mathbb{I}(Y^x, Z^x),
\end{align*}
\]

as desired. \(\square\)

4. Continuity

**Definition 6.** Let \(X_n \in \text{FRV}(\Omega)\) be a sequence of random variables, and let \(p_n : X_n \to [0, 1]\) be the associated sequence of probability mass functions. Then \(X_n\) is said to **weakly converge** (or converge in distribution) to the random variable \(X \in \text{FRV}(\Omega)\) with probability mass function \(p : X \to [0, 1]\) if the following conditions hold.

i. There exists an \(N \in \mathbb{N}\) for which \(X_n = X\) for all \(n \geq N\).

ii. For all \(x \in X\) we have \(\lim_{n \to \infty} p_n(x) = p(x)\), i.e., \(p_n \to p\) pointwise.

In such a case, we write \(X_n \to X\). If \((X_n, Y_n) \in \text{FRV}(\Omega) \times \text{FRV}(\Omega)\) is a sequence of pairs of random variables, then \((X_n, Y_n)\) is said to **weakly converge** to \((X, Y) \in \text{FRV}(\Omega) \times \text{FRV}(\Omega)\) if \(\mathcal{P}(X_n, Y_n) \to \mathcal{P}(X, Y)\).

**Proposition 6.** Shannon entropy is continuous, i.e., if \(X_n \to X\), then

\[H(X) = \lim_{n \to \infty} H(X_n).\]

**Proof.** This result is standard, see, e.g., [3] or [11]. \(\square\)

**Proposition 7.** Mutual information is continuous, i.e., if \((X_n, Y_n) \to (X, Y)\), then

\[\mathbb{I}(X, Y) = \lim_{n \to \infty} \mathbb{I}(X_n, Y_n).\]

**Proof.** Suppose \((X_n, Y_n) \to (X, Y)\), so that \(X_n \to X, Y_n \to Y\) and \(H(X_n, Y_n) \to H(X, Y)\). We then have

\[\mathbb{I}(X, Y) = H(X) + H(Y) - H(X, Y) = \lim_{n \to \infty} (H(X_n) + H(Y_n) - H(X_n, Y_n)) = \lim_{n \to \infty} \mathbb{I}(X_n, Y_n),\]

as desired. \(\square\)

5. Markov Triangles

In this section, we define the notion of a Markov triangle, a concept based on the notion of a “coalescable” composition of communication channels which was introduced in [13]. Such a notion will be crucial for our characterization of mutual information.
Definition 7. Let \( X \in \text{FRV}(\Omega) \) be a random variable with probability mass function \( p : X \rightarrow [0,1] \), and let \( x \in X \). Then for any random variable \( Y \in \text{FRV}(\Omega) \), the conditional distribution function of \( Y \) given \( X = x \) is the function \( q^x : [0,1] \rightarrow [0,1] \) given by

\[
q^x(y) = \begin{cases} 
\frac{p(x,y)}{p(x)} & \text{if } p(x) \neq 0 \\
0 & \text{otherwise}
\end{cases}
\]

From here on, the value \( q^x(y) \) will be denoted \( q(y|x) \). The conditional entropy of \( Y \) given \( X \) is the non-negative real number \( H(Y|X) \) given by

\[
H(Y|X) = \sum_{x \in X} p(x)H(q^x),
\]

where \( H(q^x) \) is the Shannon entropy of the distribution \( q^x \) on \( Y \).

Proposition 8. Let \( \{X, Y\} \) be a pair of random variables. Then

\[
\mathbb{I}(X,Y) = \mathbb{I}(Y|X) - H(Y|X).
\]

Proof. Since \( \mathbb{I}(Y,X) = H(Y) \), the statement follows from the well-known fact that \( \mathbb{I}(X,Y) = H(Y) - H(Y|X) \), the proof of which may be found in any information theory text (e.g., [17]). \( \square \)

Definition 8. Let \( \{X, Y, Z\} \) be a triple of random variables with supports \( X, Y \) and \( Z \) respectively, and let \( q(y|x), p(z|y) \) and \( r(z|x) \) denote the associated conditional distribution functions. Then \( \{X, Y, Z\} \) is said to form a Markov triangle if there exists a function \( h : Z \times X \rightarrow Y \) such that for all \( (z, x) \in Z \times X \) we have

\[
r(z|x) = p(z|h(z,x))q(h(z,x)|x).
\]

In such a case, \( h \) is said to be a mediator function for the triple \( \{X, Y, Z\} \).

Remark 2. A Markov triangle \( \{X, Y, Z\} \) with supports \( X, Y \), and \( Z \) may be thought of as a composition of noisy channels \( X \xrightarrow{\mathbb{F}} Y \xrightarrow{\mathbb{G}} Z \) such that if \( z \in Z \) is the output of the channel \( \mathbb{G} \circ \mathbb{F} \), and one is given the information that the associated input was \( x \in X \), then the output at the intermediary stage \( Y \) was necessarily \( y = h(z, x) \) (where \( h \) is the associated mediator function). In other words, if \( \mathbb{P} \) is a symbol for general probabilities, then

\[
\mathbb{P}(x,z|y) = \mathbb{P}(x|y)\mathbb{P}(z|y),
\]

thus the Markov triangle condition says that \( X \) and \( Z \) are conditionally independent given \( Y \). As compositions of deterministic channels always satisfy this property, Markov triangles are a generalization of compositions of deterministic channels. While Markov triangles play a crucial role in our characterization of mutual information and also the characterizations of conditional entropy and information loss in [13], their broader significance in the study of information measures has yet to be determined.

Proposition 9. Suppose \( \{X, Y, Z\} \) is a Markov triangle. Then

\[
\mathbb{I}(X,Z) = \mathbb{I}(X,Y) + \mathbb{I}(Y,Z) - \mathbb{I}(Y,Y).
\]

In particular, \( \mathbb{I}(X,Z) \leq \mathbb{I}(X,Y) + \mathbb{I}(Y,Z) \).

Before giving a proof of Proposition 9, we first need the following lemma.

Lemma 1. Suppose \( \{X, Y, Z\} \) is a Markov triangle. Then

\[
H(Z|X) = H(Z|Y) + H(Y|X).
\]
Proof. The statement is simply a reformulation of Theorem 2 in [13]. 

Proof of Proposition 9. Suppose $(X, Y, Z)$ is a Markov triangle. Then

$$\Pi(X, Z) = \Pi(Z, Z) - H(Z|X) = \Pi(Y, Y) - H(Y|X)$$

as desired. 

Proposition 10. Let $X, Y \in \text{FRV}(\Omega)$ be random variables with probability mass functions $p : X \rightarrow [0, 1]$ and $q : Y \rightarrow [0, 1]$ respectively. Then the following statements hold.

i. The triple $(X, \mathcal{P}(X, Y), Y)$ is a Markov triangle.

ii. If $f : X \rightarrow X'$ is a bijection, then the triple $(X, f \circ X, Y)$ is a Markov triangle.

iii. If $g : Y \rightarrow Y'$ is a bijection, then the triple $(X, Y, g \circ Y)$ is a Markov triangle.

Proof.

i. Let $r(y|x)$ be the conditional distribution associated with $(X, Y)$, let $p(y|\tilde{x}, \tilde{y})$ be the conditional distribution associated with $(\mathcal{P}(X, Y), Y)$, and let $q(\tilde{x}, \tilde{y})|x$ be the conditional distribution function associated with $(X, \mathcal{P}(X, Y))$. Then for all $y \in Y$ and $x \in X$ we have

$$r(y|x) = \sum_{(\tilde{x}, \tilde{y}) \in X \times Y} p(y|\tilde{x}, \tilde{y})q(\tilde{x}, \tilde{y})|x = p(y|x)p(\tilde{y}|x),$$

where the second equality comes from the fact that $p(y|\tilde{x}, \tilde{y}) = 0$ unless $\tilde{y} = y$ and $q(\tilde{x}, \tilde{y})|x = 0$ unless $x = \tilde{x}$. It then follows that the function $h : Y \times X \rightarrow X \times Y$ given by $h(y, x) = (y, x)$ is a mediator function for $(X, \mathcal{P}(X, Y), Y)$, thus $(X, \mathcal{P}(X, Y), Y)$ is a Markov triangle.

ii. Let $r(y|x)$ be the conditional distribution associated with $(X, Y)$, let $p(y|x')$ be the conditional distribution associated with $(\mathcal{P}(X, f \circ X), Y)$, and let $q(x'|x)$ be the conditional distribution function associated with $(X, \mathcal{P}(X, f \circ X))$. Then for all $y \in Y$ and $x \in X$ we have

$$r(y|x) = \sum_{x' \in X'} p(y|x')q(x'|x) = p(y|f(x))q(f(x)|x),$$

where the second equality comes from the fact that $q(x'|x) = 0$ unless $x' = f(x)$. It then follows that the function $h : Y \times X \rightarrow X$ given by $h(y, x) = f(x)$ is a mediator function for $(X, f \circ X, Y)$, thus $(X, f \circ X, Y)$ is a Markov triangle.

iii. Let $r(y'|x)$ be the conditional distribution associated with $(X, g \circ Y)$, let $p(y'|y)$ be the conditional distribution associated with $(Y, g \circ Y)$, and let $q(y|x)$ be the conditional distribution associated with $(X, Y)$. Then for all $y' \in Y'$ and $x \in X$ we have

$$r(y'|x) = \sum_{y \in Y} p(y'|y)q(y|x) = p(y'|y)q^{-1}(y'),$$

where the second equality comes from the fact that $p(y'|y) = 0$ unless $y = g^{-1}(y')$. It then follows that the function $h : Y' \times X \rightarrow X' \times Y$ given by $h(y', x) = g^{-1}(y')$ is a mediator function for $(X, Y, g \circ Y)$, thus $(X, Y, g \circ Y)$ is a Markov triangle.

6. Characterization Theorem

We now state and prove our characterization theorem for mutual information.
Definition 9. Let \((\Omega, \Sigma, \mu)\) and \((\Omega', \Sigma', \mu')\) be probability spaces. A map \(\pi : \Omega' \to \Omega\) is said to be **measure-preserving** if for all \(\sigma \in \Sigma\) we have \(\pi^{-1}(\sigma) \in \Sigma'\) and
\[
\mu'(\pi^{-1}(\sigma)) = \mu(\sigma).
\]

Definition 10. Let \(F\) be a map that sends pairs of random variables to the real numbers.
- \(F\) is said to be **continuous** if
\[
F(X, Y) = \lim_{n \to \infty} F(X_n, Y_n)
\]
whenever \((X_n, Y_n) \to (X, Y)\).
- \(F\) is said to be **strongly additive** if given a random variable \(X\) with probability mass function \(p : X \to [0, 1]\), and a collection of pairs of random variables \((Y^x, Z^x)\) indexed by \(X\), then
\[
F\left(\bigoplus_{x \in X} p(x) (Y^x, Z^x)\right) = F(X, X) + \sum_{x \in X} p(x) F(Y^x, Z^x).
\]
- \(F\) is said to be **symmetric** if \(F(X, Y) = F(Y, X)\) for every pair of random variables \((X, Y)\).
- \(F\) is said to be **invariant under pullbacks** if for every pair of random variables \((X, Y) \in \text{FRV}(\Omega) \times \text{FRV}(\Omega)\) and every measure-preserving map \(\pi : \Omega' \to \Omega\) we have
\[
F(X, Y) = F(X \circ \pi, Y \circ \pi).
\]
- \(F\) is said to be **weakly functorial** if for every Markov triangle \((X, Y, Z)\) we have
\[
F(X, Z) = F(X, Y) + F(Y, Z) - F(Y, Y).
\]

Remark 3. The terminology “weakly functorial” comes from viewing \((9)\) from a category-theoretic perspective. In particular, with a pair of random variables \((X, Y)\) one may associate a noisy channel \(X \overset{f}{\to} Y\) where \(X = \text{Supp}(X)\) and \(Y = \text{Supp}(Y)\), so that a Markov triangle \((X, Y, Z)\) then corresponds to a composition \(X \overset{f}{\to} Y \overset{g}{\to} Z\) with \(Z = \text{Supp}(Z)\). If \(\text{FinPS}\) denotes the category of noisy channels and \(\text{BR}\) denotes the category with one object whose morphisms are the real numbers (with a composition corresponding to addition), then a map \(F : \text{FinPS} \to \text{BR}\) is a functor if
\[
F(g \circ f) = F(g) + F(f).
\]
Rewriting \((10)\) in terms of the pairs of random variables for which the morphisms \(f, g,\) and \(g \circ f\) are associated with, then the functoriality condition \((10)\) reads
\[
F(X, Z) = F(X, Y) + F(Y, Z),
\]
thus the condition \(F(X, Z) \leq F(X, Y) + F(Y, Z)\) is a weaker form of functoriality. For more on information measures from a category-theoretic perspective, see [11–14].

Theorem 1 (Axiomatic Characterization of Mutual Information). Let \(F\) be a map that sends pairs of random variables to the non-negative real numbers, and suppose \(F\) satisfies the following conditions.
1. \(F\) is continuous.
2. \(F\) is strongly additive.
3. \(F\) is symmetric.
4. \(F\) is weakly functorial.
5. \(F\) is invariant under pullbacks.
6. \(F(X, C) = 0\) for every constant random variable \(C\).
Then \( F \) is a non-negative multiple of mutual information. Conversely, mutual information satisfies conditions 1–6.

Before giving a proof, we first need several lemmas. The first lemma states that a map \( F \) on pairs of random variables which is continuous and invariant under pullbacks, only depends on the underlying probability mass functions of the random variables.

**Lemma 2.** Let \( F \) be a map from pairs of random variables to the real numbers, which is continuous and invariant under pullbacks, and suppose \( (X, Y) \in \text{FRV}(\Omega) \times \text{FRV}(\Omega) \) and \( (X', Y') \in \text{FRV}(\Omega') \times \text{FRV}(\Omega') \) are such that the associated joint distribution functions \( \vartheta : X \times Y \to [0, 1] \) and \( \vartheta' : X' \times Y' \to [0, 1] \) are equal. Then \( F(X, Y) = F(X', Y') \).

**Proof.** Let \( \pi : \Omega \times \Omega' \to \Omega \) and \( \pi' : \Omega \times \Omega' \to \Omega' \) be the natural projections. Since both the natural projections are measure-preserving, we have \( F(X, Y) = F(X \circ \pi, Y \circ \pi) = F(X', Y') = F(X' \circ \pi', Y' \circ \pi') \), and moreover, from the assumption that \( \vartheta = \vartheta' \) it follows that the joint distribution functions associated with \( (X \circ \pi, Y \circ \pi) \) and \( (X' \circ \pi', Y' \circ \pi') \) are equal. It then follows that if \( (X_n, Y_n) \) is the constant sequence given by \( X_n = X' \circ \pi' \) and \( Y_n = Y' \circ \pi' \) for all \( n \in \mathbb{N} \), then \( (X_n, Y_n) \to (X \circ \pi, Y \circ \pi) \) (since \( \mathcal{P}(X_n, Y_n) \to \mathcal{P}(X \circ \pi, Y \circ \pi) \)). We then have

\[
F(X, Y) = F(X \circ \pi, Y \circ \pi) = \lim_{n \to \infty} F(X_n, Y_n) = F(X' \circ \pi', Y' \circ \pi') = F(X', Y'),
\]

as desired. \( \square \)

**Lemma 3.** Let \( X \) be a random variable with probability mass function \( p : X \to [0, 1] \), let \( f : X \to Y \) be a bijection, and suppose \( C \) is a constant random variable. Then the following statements hold.

i. The triples \( (X, f \circ X, C) \) and \( (f \circ X, X, C) \) are both Markov triangles.

ii. Let \( F \) be a map that sends pairs of random variables to real numbers, and suppose \( F \) is symmetric and weakly functorial. Then

\[
F(X, C) - F(f \circ X, C) + F(f \circ X, f \circ X) = F(f \circ X, C) - F(X, C) + F(X, X). \tag{11}
\]

**Proof.**

i. The statement follows from item ii of Proposition 10.

ii. By item 3, the triples \( (X, f \circ X, C) \) and \( (f \circ X, X, C) \) are both Markov triangles, thus the weak functoriality of \( F \) yields

\[
F(X, C) = F(X, f \circ X) + F(f \circ X, C) - F(f \circ X, f \circ X), \tag{12}
\]

and

\[
F(f \circ X, C) = F(f \circ X, X) + F(X, C) - F(X, X). \tag{13}
\]

And since \( F \) is symmetric \( F(X, f \circ X) = F(f \circ X, X) \), thus Equations (12) and (13) imply Equation (11), as desired. \( \square \)

The next lemma is Baez, Fritz, and Leinster’s reformulation of Faddeev’s characterization of Shannon entropy [3], which they use in their characterization of the information loss associated with a deterministic mapping [11]. This lemma will allow us to relate \( F(X, X) \) to the Shannon entropy \( H(X) \).

**Lemma 4.** Let \( S \) be a map that sends finite probability distributions to the non-negative real numbers, and suppose \( S \) satisfies the following conditions.

i. \( S \) is continuous, i.e., if \( p_n : X \to [0, 1] \) is a convergent sequence of probability distributions on a finite set \( X \) (i.e., if \( \lim_{n \to \infty} p_n(x) \) exists for all \( x \in X \)), then

\[
S\left( \lim_{n \to \infty} p_n \right) = \lim_{n \to \infty} S(p_n).
\]
Entropy

Let \( F \) be a map that sends pairs of random variables to the non-negative real numbers satisfying conditions 1–6 of Theorem 1, and let \( \mathcal{E} \) be the map on random variables given by

\[
\mathcal{E}(X) = F(X, X).
\]

Then \( \mathcal{E} \) is a non-negative multiple of Shannon entropy.

**Proof.** Let \( \phi \) be the map that takes a random variable to its probability mass function, let \( \sigma \) be a section (so that \( \phi \circ \sigma \) is the identity), and let \( \mathcal{S} = \mathcal{E} \circ \sigma \). Since \( F \) is invariant under pullbacks (condition 5 of Theorem 1) Lemma 2 holds, thus the map \( \mathcal{S} \) is independent of the choice of a section \( \phi \), and as such, it follows that \( \mathcal{E} = \mathcal{S} \circ \sigma \). We now show that \( \mathcal{S} \) satisfies items i–iv of Lemma 4, which then implies \( \mathcal{E}(X) \) is a non-negative multiple of the Shannon entropy \( H(X) \).

i. Let \( p_n : X \to [0, 1] \) be a sequence of probability distributions on a finite set \( X \), and suppose \( \lim_{n \to \infty} p_n = p \). It then follows that \( X_n = \sigma(p_n) \) weakly converges to \( X = \sigma(p) \), thus

\[
\mathcal{S}\left( \lim_{n \to \infty} p_n \right) = \mathcal{S}(p) = (\mathcal{E} \circ \sigma)(p) = \mathcal{E}(X) = F(X, X) = \lim_{n \to \infty} F(X_n, X_n)
\]

where the fifth equality follows from the continuity assumption on \( F \) (condition 1 of Theorem 1).

ii. Let \( 1 : \{\ast\} \to [0, 1] \) be a point mass distribution, so that \( \sigma(1) = C \) with \( C \) a constant random variable. Then \( \mathcal{S}(1) = \mathcal{E}(\sigma(1)) = \mathcal{E}(C) = F(C, C) = 0 \), where the last equality follows from condition 6 of Theorem 1, i.e., that \( F(X, C) = 0 \) for every constant random variable \( C \).

iii. Let \( X \) be a random variable with probability mass function \( p : X \to [0, 1] \), and suppose \( f : X \to Y \) is a bijection. Since \( F \) is symmetric and weakly functorial (conditions 3 and 4 of Theorem 1), the hypotheses of item ii Lemma 3 are satisfied, so that Equation (11) holds, i.e., for any constant random variable \( C \) we have

\[
F(X, C) = F(f \circ X, C) + F(f \circ X, f \circ X) = F(f \circ X, C) - F(X, C) + F(X, X).
\]

And since \( F(X, C) = F(f \circ X, C) = 0 \) by condition 6 of Theorem 1, it follows that \( F(X, X) = F(f \circ X, f \circ X) \). Now let \( q : Y \to [0, 1] \) be the probability mass function of \( f \circ X \), so that \( q = p \circ f^{-1} \). We then have

\[
\mathcal{S}(p) = \mathcal{E}(X) = F(X, X) = F(f \circ X, f \circ X) = \mathcal{E}(f \circ X) = \mathcal{S}(q) = \mathcal{S}(p \circ f^{-1}),
\]

thus \( \mathcal{S} \) satisfies item iii of Faddeev’s Theorem.
iv. Let \( X \) be a random variable with probability mass function \( p : \mathcal{X} \to [0, 1] \), \( Y^x \) a collection of random variables indexed by \( \mathcal{X} \), and let \( q^x : Y^x \to [0, 1] \) be the associated probability mass functions for all \( x \in \mathcal{X} \). Then \( \bigoplus_{x \in \mathcal{X}} p(x)Y^x \) has probability mass function \( \bigoplus_{x \in \mathcal{X}} p(x)q^x \), thus

\[
S\left( \bigoplus_{x \in \mathcal{X}} p(x)q^x \right) = \delta'\left( \bigoplus_{x \in \mathcal{X}} p(x)Y^x \right) = F\left( \bigoplus_{x \in \mathcal{X}} p(x)Y^x, \bigoplus_{x \in \mathcal{X}} p(x)Y^x \right) = F\left( \bigoplus_{x \in \mathcal{X}} p(x)(Y^x, Y^x) \right) = F(X,X) + \sum_{x \in \mathcal{X}} p(x)F(Y^x, Y^x) = \delta'(X) + \sum_{x \in \mathcal{X}} p(x)\delta'(Y^x) = S(p) + \sum_{x \in \mathcal{X}} p(x)S(q^x),
\]

where the fourth equality follows from the strong additivity of \( F \), i.e., condition 2 of Theorem 1. It then follows that \( S \) satisfies item iv of Faddeev’s Theorem, as desired. \( \Box \)

The next lemma is the analog of property iii of Lemma 4 for information measures on pairs of random variables.

**Lemma 6.** Let \( X, Y \in \text{FRV}(\Omega) \) be random variables with probability mass functions \( p : \mathcal{X} \to [0, 1] \) and \( q : \mathcal{Y} \to [0, 1] \) respectively, and suppose \( F \) is a map on pairs of random variables to the real numbers which is symmetric, weakly functorial, and \( F(X, C) = 0 \) for every constant random variable \( C \). If \( f : \mathcal{X} \to \mathcal{X}' \) and \( g : \mathcal{Y} \to \mathcal{Y}' \) are bijections, then

\[
F(X,Y) = F(f \circ X, g \circ Y). \tag{14}
\]

**Proof.** Since \( f \) is a bijection, \( (X, f \circ X, X) \) is a Markov triangle by item ii of Proposition 10, thus

\[
F(X,X) = F(f \circ X, f \circ X) + F(f \circ X, X) - F(f \circ X, f \circ X). \tag{15}
\]

From the proof of Lemma 5 it follows that if \( F \) is weakly functorial, symmetric and \( F(X, C) = 0 \) for every constant random variable \( C \), \( F(f \circ X, f \circ X) = F(X,X) \). Moreover by the symmetry of \( F \) we have \( F(X, f \circ X) = F(f \circ X, X) \), thus Equation (15) implies \( F(X,X) = F(X, f \circ X) \).

Now consider the triples \( (f \circ X, X, g \circ Y) \) and \( (X, Y, g \circ Y) \), which are both Markov triangles by items ii and iii of Proposition 10. The weakly functorial assumption on \( F \) then yields

\[
F(f \circ X, g \circ Y) = F(f \circ X, X) + F(X, g \circ Y) - F(X, X)
= F(f \circ X, X) + F(X, Y) + F(Y, g \circ Y) - F(Y, Y) - F(X, X),
\]

and since \( F(f \circ X, X) = F(X,X) \) and \( F(Y, g \circ Y) = F(Y, Y) \), it follows that \( F(X,Y) = F(f \circ X, g \circ Y) \), as desired. \( \Box \)

The next lemma, together with the fact that \( (X, \mathcal{P}(X,Y), Y) \) is a Markov triangle (by Proposition 10) is the crux of the proof, as we will soon see.

**Lemma 7.** Let \( F \) be a map from pairs of random variables to the real numbers satisfying conditions 1–6 of Theorem 1, and let \((X, Y)\) be a pair of random variables. Then \( F(X, \mathcal{P}(X,Y)) = F(X,X) \) and \( F(\mathcal{P}(X,Y), Y) = F(Y,Y) \).
Proof. Let \( p : \mathcal{X} \to [0, 1] \) and \( q : \mathcal{Y} \to [0, 1] \) be the probability mass functions of \( X \) and \( Y \) respectively, and for all \( x \in \mathcal{X} \), let \( Y^x \) be a random variable with probability mass function \( q^x : [0, 1] \to [0, 1] \) given by \( q^x(y) = q(y|x) \), so that \( q^x \) is the conditional distribution of \( Y \) given \( X = x \). By pulling back to larger sample spaces if necessary, we can assume without loss of generality that each \( Y^x \in \text{FRV}(\Omega) \) for some fixed \( \Omega \). We also let \( C^x \in \text{FRV}(\Omega) \) be the constant random variable supported on \( \{x\} \) for all \( x \in \mathcal{X} \), we let \( f : \bigoplus_{x \in \mathcal{X}} p(x) C^x \) and \( g : \bigoplus_{x \in \mathcal{X}} p(x) Y^x \) be the canonical bijections, and we let \( \pi : \mathcal{X} \times \Omega \to \Omega \) be the natural projection. It then follows that \( f \circ \bigoplus_{x \in \mathcal{X}} p(x) C^x \) and \( X \circ \pi \) both have probability mass function \( p : \mathcal{X} \to [0, 1] \), and also, that \( g \circ \bigoplus_{x \in \mathcal{X}} p(x) Y^x \) and \( \mathcal{P}(X, Y) \circ \pi \) both have probability mass function equal to the joint distribution function \( \vartheta : \mathcal{X} \times \mathcal{Y} \to [0, 1] \) associated with \( (X, Y) \), thus Lemma 2 yields

\[
F(X \circ \pi, \mathcal{P}(X, Y) \circ \pi) = F\left( f \circ \bigoplus_{x \in \mathcal{X}} p(x) C^x, g \circ \bigoplus_{x \in \mathcal{X}} p(x) Y^x \right). \tag{16}
\]

We then have

\[
F(X, \mathcal{P}(X, Y)) \stackrel{(8)}{=} F(X \circ \pi, \mathcal{P}(X, Y) \circ \pi) \stackrel{(16)}{=} F\left( f \circ \bigoplus_{x \in \mathcal{X}} p(x) C^x, g \circ \bigoplus_{x \in \mathcal{X}} p(x) Y^x \right) \stackrel{(14)}{=} F\left( \bigoplus_{x \in \mathcal{X}} p(x) C^x, \bigoplus_{x \in \mathcal{X}} p(x) Y^x \right) \stackrel{(9)}{=} F\left( \bigoplus_{x \in \mathcal{X}} p(x)(C^x, Y^x) \right) \stackrel{(7)}{=} F(X, X) + \sum_{x \in \mathcal{X}} p(x)F(C^x, Y^x) = F(X, X),
\]

where the last equality follows from the fact that \( F(C, X) = 0 \) for every constant random variable \( C \) (since \( F \) is symmetric and \( F(X, C) = 0 \) for every constant random variable \( C \)).

As for \( F(\mathcal{P}(X, Y), Y) \), first note that \( F(Y, \mathcal{P}(X, Y)) = F(Y, Y) \) by what we have just proved. We then have

\[
F(\mathcal{P}(X, Y), Y) = F(Y, \mathcal{P}(X, Y)) = F(Y, \mathcal{P}(X, Y)) = F(Y, Y),
\]

where the first and second equalities follow from symmetry and invariance under pullbacks. \( \square \)

Proof of Theorem 1. Suppose \( F \) is a map from pairs of random variables to the non-negative real numbers satisfying conditions 1–6 of Theorem 1. According to Lemma 5, there exists a constant \( c \geq 0 \) such that \( F(X, X) = ch(X) \) for all random variables \( X \). Now let \( (X, Y) \) be an arbitrary pair of random variables. According to Proposition 10, the triple \( (X, \mathcal{P}(X, Y), Y) \) is a Markov triangle, thus

\[
F(X, Y) \stackrel{(9)}{=} F(X, \mathcal{P}(X, Y)) + F(\mathcal{P}(X, Y), Y) - F(\mathcal{P}(X, Y), \mathcal{P}(X, Y)) = F(X, X) + F(Y, Y) - ch(\mathcal{P}(X, Y)) = ch(X) + ch(Y) - ch(X, Y) = cI(X, Y),
\]

where the second equality follows from Lemma 7 and Lemma 5, and the third equality follows from Lemma 5 and item 2 of Proposition 2, thus \( F \) is a non-negative multiple of mutual information.

Conversely, mutual information satisfies condition 1 by Proposition 7, condition 2 by Proposition 5, condition 3 by item 1 of Proposition 1, condition 4 by Proposition 9, condi-
tion 5 by the fact that mutual information only depends on probabilities, and condition 6 by item 1 of Proposition 1.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** We thank Arthur J. Parzygnat for many useful discussions.

**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. [CrossRef]
2. Aczél, J.; Daróczy, Z. On measures of information and their characterizations. In *Mathematics in Science and Engineering*; Academic Press: New York, NY, USA; London, UK, 1975; Volume 115, p. xii + 234.
3. Faddeev, D.K. On the concept of entropy of a finite probabilistic scheme. *Uspehi Mat. Nauk (N.S.)* **1956**, *11*, 227–231.
4. Furuichi, S. On uniqueness theorems for Tsallis entropy and Tsallis relative entropy. *IEEE Trans. Inform. Theory* **2005**, *51*, 3638–3645. [CrossRef]
5. Rényi, A. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*; University California Press: Berkeley, CA, USA, 1961; pp. 547–561.
6. Petz, D. Characterization of the relative entropy of states of matrix algebras. *Acta Math. Hung.* **1992**, *59*, 449–455. [CrossRef]
7. Ohya, M.; Petz, D. Quantum Entropy and Its Use. In *Texts and Monographs in Physics*; Springer: Berlin/Heidelberg, Germany, 1993; p. viii + 335. [CrossRef]
8. Leinster, T. A short characterization of relative entropy. *J. Math. Phys.* **2019**, *60*, 023302. [CrossRef]
9. Ebanks, B.R.; Kannappan, P.; Sahoo, P.K.; Sander, W. Characterizations of sum form information measures on open domains. *Aequationes Math.* **1997**, *54*, 1–30. [CrossRef]
10. Csiszár, I. Axiomatic characterizations of information measures. *Entropy* **2008**, *10*, 261–273. [CrossRef]
11. Baez, J.C.; Fritz, T.; Leinster, T. A characterization of entropy in terms of information loss. *Entropy* **2011**, *13*, 1945–1957. [CrossRef]
12. Baez, J.C.; Fritz, T. A Bayesian characterization of relative entropy. *Theory Appl. Categ.* **2014**, *29*, 422–457.
13. Fullwood, J.; Parzygnat, A.J. The Information Loss of a Stochastic Map. *Entropy* **2021**, *23*, 1021. [CrossRef] [PubMed]
14. Parzygnat, A.J. A functorial characterization of von Neumann entropy. *Cah. Topol. Géom. Différ. Catég.* **2022**, *63*, 89–128.
15. Leinster, T. *Entropy and Diversity: The Axiomatic Approach*; Cambridge University Press: Cambridge, UK, 2021; p. 442. [CrossRef]
16. Romashchenko, A.; Zimand, M. An operational characterization of mutual information in algorithmic information theory. *J. ACM* **2019**, *66*, 42. [CrossRef]
17. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley Series in Telecommunications and Signal Processing; Wiley: Hoboken, NJ, USA, 2006. [CrossRef]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.