Broadly speaking, an information measure is any function of one or more probability distributions. An entropy is an information measure that have units of entropy — negative logarithms of probabilities, or a linear combination of the same. (Although in many cases, tradition doth dictate that an entropic measure be referred to as an information, e.g. “mutual information” rather than “mutual entropy”.) In this brief survey we’ll explore the menagerie of entropic and information measures that have been developed to characterize classical probability distributions.

Contents

0 Notes on notation and nomenclature 2
1 Entropy 3
   Entropy ........................................... 3
   Joint entropy .................................. 3
   Marginal entropy ............................... 3
   Conditional entropy ............................ 3
   Perplexity ....................................... 4
2 Mutual information 4
   Mutual information .............................. 4
   Multivariate mutual information .......... 5
   Interaction information ...................... 5
   Conditional mutual information .......... 5
   Binding information ............................ 6
   Residual entropy ................................ 6
   Total correlation ............................... 7
   Lautum information ............................ 7
   Uncertainty coefficient ........................ 7
3 Relative entropy 7
   Relative entropy ................................ 7
   Cross entropy .................................. 7
   Burg entropy ................................... 10
   Relative joint entropy ........................ 10
   Relative conditional entropy ............... 10
   Relative mutual information ............... 10
   Relative conditional mutual information . 10
   Relative relative entropy .................... 10
   Jeffreys entropy ................................ 10
   Jensen-Shannon divergence .................. 11
   General Jensen-Shannon divergence .......... 11
   Jensen-Shannon entropy ....................... 11
   Resistor-average entropy .................... 11
4 Rényi information 11
   Rényi information ................................ 11
   Collision entropy ................................ 12
   Min-entropy ..................................... 12
   Hartley entropy ................................ 12
   Tsallis information ................................ 12
   Sharma-Mittal information .................. 12
5 Csiszár f-divergences 12
   Csiszár f-divergence .......................... 12
   Dual f-divergence ................................ 12
   Symmetric f-divergences ...................... 12
   K-divergence .................................... 14
   Fidelity ........................................... 14
   Hellinger discrimination ...................... 14
   Pearson divergence ............................. 14
   Neyman divergence .............................. 14
   LeCam discrimination .......................... 14
   Skewed K-divergence ........................... 14
   Alpha-Jensen-Shannon-entropy .............. 15
6 Chernoff divergence 15
   Chernoff divergence ........................... 15
   Chernoff coefficient ........................... 15
   Rényi divergence ................................ 15
   Alpha-divergence ................................ 15
   Cressie-Read divergence ...................... 16
   Tsallis divergence ................................ 16
   Sharma-Mittal divergence .................... 16
7 Cauchy-Schwarz divergence 16
   Cauchy-Schwarz divergence ................... 16
   Cauchy-Schwarz angle ........................... 16
In physics we often write the expectation as \( \langle f(A) \rangle \) or \( \langle f(a) \rangle \). In the last case we are using the same, potentially ambiguous, shorthand as commonly used for probabilities, i.e.

\[
P(a) \text{ for } P(A_a).
\]

Most of the information measures discussed herein could be expressed using the expectation operator, but for maximum clarity we make the averages explicit. For example the Shannon entropy (2) is the ensemble average of the specific (or point-wise) entropy \(-\ln P(A_a)\). (§9).

\[
S(A) = E \left[-\ln P(A_x)\right] = -\sum_x P(A_x) \ln P(A_x)
\]

**Collections of ensembles** Since we have used subscripts to index propositions, we’ll use superscripts to index collections of ensembles, \( A = \{A^1, A^2, \ldots, A^{|A|}\} \). We never need to take powers of ensembles or propositions, so there is little risk of ambiguity. Other useful notation includes \( P[A] \), the power set of \( A \) (the set of all subsets); \( |A| \), the set cardinality; \( \emptyset \) for the empty set; and \( A \setminus A \), the set complement (difference) of \( A \).

**Naming and notating** Information measures are given CamelCased function names, unless an unambiguous short function name is in common usage. Commas between ensembles (or propositions) denote conjunction (logical and); a colon \( : \) the mutual information (6) “between” ensembles; a double bar \( || \) the relative entropy (16) of one ensemble “relative to” another; a semicolon \( ; \) for any other comparison of ensembles; and a bar \( | \) denotes conditioning (“given”). We’ll use the operator precedence (high to low) \( :, \ldots, ;, |, ||, \cdot \) to obviate excessive bracketing [1]. Samples spaces are the same on either side of double bars \( || \) or semicolons \( ; \), but different across bars \( | \) and colons \( ; \). Measures are symmetric to interchange of the ensembles across commas and colons, i.e. \( S(A, B, C) \) is the same as \( S(C, B, A) \) and \( I(A : B : C) \) is the same as \( I(B : C : A) \).

**Information diagrams** Information diagrams (See Figs. 1, 2, 3 and 4) are a graphical display of multivariant Shannon information measures [23, 47, 72].

These are not Venn-diagrams per se, since the individual regions can have positive or negative weight. The regions of an information diagram corresponding to a particular information measure can be deduced by mapping joint distributions \( ‘A, B’ \) to the union of sets \( A \cup B \), mutual measures \( ‘A : B’ \), to the intersection of sets \( A \cap B \), and condi-
Table 1: Information measure symbology

<table>
<thead>
<tr>
<th>symbol</th>
<th>usage</th>
<th>commutative</th>
<th>precedence</th>
</tr>
</thead>
<tbody>
<tr>
<td>,</td>
<td>conjugation</td>
<td>yes</td>
<td>high</td>
</tr>
<tr>
<td>:</td>
<td>mutual</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>conditional</td>
<td>no</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>relative entropy</td>
</tr>
<tr>
<td>;</td>
<td>divergence</td>
<td>no</td>
<td>low</td>
</tr>
</tbody>
</table>

Dissimilarity An information-theoretic divergence is a measure of dissimilarity between a pair of ensembles that is non-negative and zero if (and only if) the distributions are identical. Since divergences are not symmetric to their arguments in general, we can also define the dual divergence \( d^*(A : B) = d(B : A) \). We’ll refer to a symmetric divergence as a discrimination. By distance we mean a metric distance: a measure that is non-negative; symmetric; zero if (and only if) the distributions are identical (reflective); and obeys the triangle inequality, \( d(A ; B) + d(B ; C) \geq d(A ; C) \).

1 Entropy

Entropy (Shannon entropy, Gibbs entropy) A measure of the inherent uncertainty or randomness of a single random variable.

\[
S(A) := - \sum_a P(A_a) \ln P(A_a)
\]  

In information theory the entropy is typically denoted by the symbol \( H \), a notation that dates back to Boltzmann and his H-theorem [3], and adopted by Shannon [12]. The notation \( S \) is due to Clausius and the original discovery of entropy in thermodynamics [2], and adopted by Gibbs [5] for use in statistical mechanics. I tend to use \( S \) since I care about the physics of information, and the symbol \( H \) is oft needed to denote the Hamiltonian.

The units of entropy depend on the base of the logarithm: common choices are bits (“binary digits”), nats (“natural units”), or bans, for bases 2, e, or 10 respectively. Herein we’ll use natural logarithms. Note that, by convention, \( 0 \ln 0 = 0 \).

Entropy is occasionally referred to as the self-information, since entropy is equal to the mutual information between a distribution and itself, \( S(A) = I(A : A) \). This is distinct from the specific entropy (62) which is also sometimes referred to as the self-information.

Discrete entropies are non-negative and bounded.

\[ 0 \leq S(A) \leq \ln |\Omega_A| \]

Joint entropy Given a joint probability distribution \( P(A, B) \) then the joint entropy is

\[
S(A, B) := - \sum_{a,b} P(A_a, B_b) \ln P(A_a, B_b)
\]

This joint entropy can be readily generalized to any number of variables.

\[
S(A^1, A^2, \ldots, A^n) = - \sum_{a_1, a_2, \ldots, a_n} P(A_{a_1}, A_{a_2}, \ldots, A_{a_n}) \ln P(A_{a_1}, A_{a_2}, \ldots, A_{a_n})
\]

Marginal entropy The entropy of a marginal distribution. Thus \( S(A), S(B), S(C), S(A, B), S(B, C) \) and \( S(A, C) \) are all marginal entropies of the joint entropy \( S(A, B, C) \).

Conditional entropy (or equivocation) [12, 63] Measures how uncertain we are of \( A \) on the average when we know \( B \).

\[
S(A \mid B) := - \sum_b P(B_b) \sum_a P(A_a \mid B_b) \ln P(A_a \mid B_b)
\]

The conditional entropy is non-negative, since it is the expectation of non-negative entropies.

The chain rule for entropies [12, 63] expands conditional entropy as a Shannon information measure.

\[
S(A, B) = S(A \mid B) + S(B)
\]

This follows from the probability chain rule,

\[
P(A_a, B_b) = P(B_b \mid A_a)P(A_a)
\]

Subadditivity of entropy: Since entropies are always non-negative if follows that conditioning always reduces entropy, \( S(A \mid B) \leq S(A) \). This implies that entropy is subadditive: The joint entropy is less than the sum of the individual entropies (with equality only if \( A \) and \( B \) are independent).

\[
S(A, B) \leq S(A) + S(B)
\]

Bayes’ rule for probabilities is the relation \( P(A_a \mid B_b) = P(B_b \mid A_a)P(A_a)/P(B_b) \). In entropic terms the equivalent statement is (taking logarithms and averaging)

\[
S(A \mid B) = S(B \mid A) + S(A) - S(B)
\]
Table 2: Units of entropy

<table>
<thead>
<tr>
<th>Unit</th>
<th>Definition</th>
<th>Approximate Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>deciban</td>
<td>$\frac{1}{10} \log_2(10)$</td>
<td>$0.33$ bits</td>
</tr>
<tr>
<td>bit (shannon)</td>
<td>$\log_2(10)$</td>
<td>$1$ bit</td>
</tr>
<tr>
<td>nat (nit, nepit)</td>
<td>$\log_2(e)$</td>
<td>$1.44$ bits</td>
</tr>
<tr>
<td>trit</td>
<td>$\log_2(3)$</td>
<td>$1.6$ bits</td>
</tr>
<tr>
<td>quad</td>
<td>$\log_2(10)$</td>
<td>$3.32$ bits</td>
</tr>
<tr>
<td>ban (hartly)</td>
<td>$\log_2(10)$</td>
<td>$3.32$ bits</td>
</tr>
<tr>
<td>nibble (nybble)</td>
<td>$\log_2(10)$</td>
<td>$3.32$ bits</td>
</tr>
<tr>
<td>byte</td>
<td>$\log_2(10)$</td>
<td>$3.32$ bits</td>
</tr>
</tbody>
</table>

Perplexity \[^{[35]}\]

Perplexity ($A$) := $e^{S(A)}$ \hspace{1cm} (5)

The exponentiation of the entropy (with the same base).
We can also define perplexities corresponding to other entropic measures, such as the conditional perplexity
Perplexity ($A \mid B$) := $e^{S(A \mid B)}$. Partition functions in statistical mechanics are perplexities.

### 2 Mutual information

**Mutual information** (mutual entropy, transinformation) \[^{2}[12, 63]\]

$$I(A : B) := \sum_{a,b} P(A_a, B_b) \ln \frac{P(A_a, B_b)}{P(A_a)P(B_b)}$$ \hspace{1cm} (6)

Mutual information is oft notated with a semicolon, rather than a colon \[^{[12, 17]}\].

Mutual information is the reduction in uncertainty of $A$ due to the knowledge of $B$, or vice versa.

$$I(A : B) = S(A) - S(A \mid B) = S(B) - S(B \mid A) = S(A) + S(B) - S(A, B) = S(A, B) - S(A \mid B) - S(B \mid A)$$

Mutual information is non-negative.

$$0 \leq I(A : B)$$

---

\[^{2}\]I didn’t like the term Information Theory. Claude didn’t like it either. You see, the term ‘information theory’ suggests that it is a theory about information – but it’s not. It’s the transmission of information, not information. Lots of people just didn’t understand this...I coined the term ‘mutual information’ to avoid such nonsense: making the point that information is always about something. It is information provided by something, about something.” – Robert Fano \[^{[54]}\]
This also implies that entropy is **subadditive**. The sum of marginal entropy of two systems is less than, or equal to, the joint entropy.

\[ S(A) + S(B) \leq S(A, B) \]

The mutual information is zero if (and only if) A and B are **independent** (written \( A \perp B \)), such that \( P(A_a, B_b) = P(A_a)P(B_b) \). And easy proof is to note that the mutual information can be written as a relative entropy (16).

The mutual information of an ensemble with itself is the entropy (which is why entropy is occasionally called the self-information).

\[ I(A : A) = S(A) \]

**Multivariate mutual information** (co-information) [17, 20, 21, 58, 61, 72]: A multivariate generalization of the mutual information. Given a collection of probability ensembles, \( A = \{ A^1, A^2, \ldots, A^{\|A\|} \} \), the multivariate mutual information is equal to an alternating signed sum of all the marginal entropies.

\[ I(A^1 : A^2 : \cdots : A^{\|A\|}) := - \sum_{\mathcal{B} \in \mathcal{P}(\mathcal{A})} (-1)^{|\mathcal{B}|} S(B^1, B^2, \ldots, B^{\|\mathcal{B}\|}) \]

Here \( \mathcal{P}(\mathcal{A}) \) is the power set of \( \mathcal{A} \) (the set of all subsets), and \( |\mathcal{A}| \) is the set cardinality. Note that there are conflicting sign conventions in the literature: the multivariate mutual information is sometimes defined with opposite sign for odd cardinalities (See interaction information (9)).

The single variable case is equal to the entropy, \( I(A) = S(A) \), the binary case is equal to the standard mutual information (7), and the ternary case is

\[ I(A : B : C) := \sum_{a,b,c} P(A_a, B_b, C_c) \ln \frac{P(A_a, B_b, C_c)P(A_a, B_b)P(B_b, C_c)}{P(A_a, B_b, C_c)P(A_a, B_b)P(B_b, C_c)} \]

For three or more variables the mutual information can be positive, negative, or zero, whereas for one or two variables the mutual information is non-negative. For zero variables the mutual information is zero, \( I(\varnothing) = 0 \).

The mutual information defines a partitioning of the total, multivariate joint entropy into single variable, binary, ternary, and higher order shared entropies.

\[
S(A) = I(A) \\
S(A, B) = I(A) + I(B) - I(A : B) \\
S(A, B, C) = I(A) + I(B) + I(A) - I(A : B) - I(A : C) + I(A : B : C)
\]

Or generally,

\[
S(A^1, A^2, \ldots, A^{\|A\|}) = - \sum_{\mathcal{B} \in \mathcal{P}(\mathcal{A})} (-1)^{|\mathcal{B}|} I(B^1 : \cdots : B^{\|\mathcal{B}\|})
\]

The triplet interaction information is the information that a pair of variable provides about the third, compared to the information that each provides separately [17, 20, 60].

\[ I(A : B : C) = I(A : B) + I(A : C) - I(A : B, C) \]

The multivariate self-information is equal the entropy for any cardinality.

\[ I(A : A : \cdots : A) = S(A) \]

**Interaction information** (synergy, mutual information) [17]: An alternative sign convention for multivariate mutual information. The interaction information is equal in magnitude to the multivariate information, but has the opposite sign for odd number of ensembles.

\[ \text{Int}(A^1 : A^2 : \cdots : A^n) := (-1)^n I(A^1 : A^2 : \cdots : A^n) \]

The sign convention used above for multivariate information generally makes more sense.

**Conditional mutual information** [21] The average mutual information between A and B given C.

\[ I(A : B \mid C) := \sum_{c} P(C_c) \sum_{a,b} P(A_a, B_b \mid C_c) \ln \frac{P(A_a, B_b \mid C_c)}{P(A_a \mid C_c)P(B_b \mid C_c)} \]

For three or more variables the mutual information can be positive, negative, or zero, whereas for one or two variables the mutual information is non-negative. For zero variables the mutual information is zero, \( I(\varnothing) = 0 \).

\[
I(A : B \mid C) = S(A \mid C) - S(A \mid B, C) \\
= S(B \mid C) - S(B \mid A, C)
\]

**Strong subadditivity:** The conditional mutual information is non-negative. If the conditional mutual information $I(A : B | C)$ is zero, then $A$ and $B$ are conditionally independent given $C$ (written $A \perp \perp B | C$). Conversely conditional independence implies that the conditional mutual information is zero.

$$A \perp \perp B | C \iff I(A : B | C) = 0$$

Conditional independence implies that $P(A_a, B_b | C_c) = P(A_a | C_c) P(B_b | C_c)$ for all $a, b, c$.

**The chain rule** for mutual information is

$$I(A : B, C) = I(A : C) + I(A : B | C)$$

The data processing inequality states that if $A$ and $C$ are conditionally independent, given $B$ (as happens when you have a Markov chain $A \rightarrow B \rightarrow C$) then

$$I(A : B) \geq I(A : C) \quad \text{given} \quad I(A : C | B) = 0$$

Proof: $I(A : B) - I(A : C) = I(A : B | C) - I(A : C | B)$, but $I(A : C | B)$ is zero, and $I(A : B | C)$ is positive.

**Binding information** (Dual total correlation) [36, 72, 75]

$$\text{Binding}(A) := S(A) - \sum_{A \in A \setminus A} S(A | A \setminus A) \quad (11)$$

Here $A \setminus A$ is the set compliment of $A$.

**Residual entropy** (erasure entropy, independent information, variation of information, shared information dis-
The total amount of information carried by correlations is mutual spelled backwards). Joint and marginal product distributions swapped. (Lautum information, multivariate constraint, redundancy, integration) [e.g.]

\[
\text{Residual}(A) := \sum_{A \in A} S(A | A \setminus A) = S(A) - \text{Binding}(A)
\]

Measures the total amount of randomness localized to individual variables.

We can express the residual entropy as a Shannon information measure if we extend the notation to allow colons ‘:’ in conditionals (i.e. on right side of a bar ‘|’) [1], e.g.

\[
\text{Residual}(A : B : C) := I(A, B, C | A, B : A, C : B, C)
\]

**Total correlation** (Multi-information, multivariate constraint, redundancy, integration) [17, 20, 37, 58]

\[
\text{TotalCorr}(A^1, A^2, \ldots, A^n) := S(A^1) + S(A^2) + \cdots + S(A^n) - S(A^1, A^2, \ldots, A^n)
\]

The total amount of information carried by correlations between the variables. Quantifies the total correlation or redundancy. Equal to the mutual information when \( n = 2 \). The independence bound on entropy states that the total correlation is non-negative.

**Lautum information** [68]:

\[
\text{Lautum}(A; B) := \sum_{a, b} P(A_a)P(B_b) \ln \frac{P(A_a)P(B_b)}{P(A_a, B_b)}
\]

Much like the mutual information, but with the roles of joint and marginal product distributions swapped. (Lautum is mutual spelled backwards).

**Uncertainty coefficient** (relative mutual information) [67] \(^3\)

\[
\text{UncertaintyCoeff}(A; B) := \frac{I(A : B)}{S(A)} = 1 - \frac{S(A | B)}{S(A)}
\]

Given B, the fraction of the information we can predict about A.

### 3 Relative entropy

**Relative entropy** (Kullback-Leibler divergence\(^4\), KL-divergence, KL-distance, Kullback information, information gain, logarithmic divergence, information divergence) [14, 63]\(^5\)

\[
D(A \parallel B) := \sum_x P(A_x) \ln \frac{P(A_x)}{P(B_x)}
\]

Roughly speaking, the relative entropy measures the difference between two distributions, although it is not a metric since it is not symmetric \( D(A \parallel B) \neq D(B \parallel A) \) in general, nor does it obey the triangle inequality. Note that the two distributions must have the same sample space, and that we take as convention that \( 0 \ln 0 = 0 \).

One interpretation of relative entropy is that it represents an encoding cost [63]: if we encode messages using an optimal code for a probability distribution \( P(B_x) \) of messages \( x \), but the messages actually arrive with probabilities \( P(A_x) \), then each message requires, on average, an additional \( D(A \parallel B) \) nats to encode compared to the optimal encoding.

The mutual information (6) is the relative entropy between the joint and marginal product distributions. Let the random variables \( (\hat{A}, \hat{B}) \) be independent, but with the same marginals as \( (A, B) \), i.e. \( P(\hat{A}, \hat{B}) = P(A)P(B) \). Then

\[
I(A : B) = D(A, B \parallel \hat{A}, \hat{B})
\]

Similarly, for three or more variables, the relative entropy between the joint and marginal product distributions is the total correlation (13).

\[
\text{TotalCorr}(A^1, A^2, \ldots, A^n) = D(A^1, A^2, \ldots, A^n \parallel \hat{A}^1, \hat{A}^2, \ldots, \hat{A}^n)
\]

The Lautum information (14) is

\[
\text{Lautum}(A : B) = D(\hat{A}, \hat{B} \parallel A, B)
\]

---

\(^3\)Despite misinformation to the contrary, this uncertainty coefficient is not related to Theil’s U-statistic.

\(^4\)“Kullback-Leibler divergence” is probably the most common terminology, which is often denoted \( D_{KL} \) and verbalized as “dee-kay-ell”. I’ve chosen to use the more descriptive “relative entropy” partially so that we can more easily talk about the generalization of relative entropy to other relative Shannon measures.

\(^5\)Note that our notation for relative entropy is uncommon. Following [63], many authors instead directly supply the distributions as arguments, e.g. \( D(p(x) \parallel q(x)) \).
Figure 3: Components of four-variable information diagrams.
Figure 4: Information diagrams for four-variable mutual-information.
We can generalize any Shannon information measure to a relative Shannon information measure by combining appropriate linear combinations of relative joint entropies. The relative entropy is always greater or equal to the marginal relative entropy.

Relative mutual information \[ (1) \]
\[
D(A : B || A' : B') := \sum_{x,y} P(x,y) \ln \frac{P(x,y)}{P(x,A')P(y,B')} = S(A) + D(A || B)
\]
This follows because the relative conditional entropy

\[
D(A,B || A',B') = D(A,B || A',B') - D(B || B') \geq 0
\]

Relative conditional mutual information \[ (22) \]
\[
D(A : B | C || A' : B' | C') := \sum_{x,y} P(x,y,z) \ln \frac{P(x,y,z)}{P(z,C)P(x,A')P(y,B')} = I(A : B) - I(A : B | C)
\]
We could continue this insanity by generalizing the conditional mutual entropy to many variables.

Relative relative entropy \[ (23) \]
\[
D(|A||B) || (C||D) := \sum_{x} P(x) \ln \frac{P'(x)}{P'(x,C)}
\]

The relative entropy can be applied to relative entropy itself, leading to the recursively defined relative relative entropy. As an example, the relative mutual information can also be expressed as a relative relative entropy (just as the mutual information can be expressed as a relative entropy).

Jeffreys entropy The Jeffreys entropy (Jeffreys divergence, J-divergence or symmetrized Kullback-Leibler di-
The divergence is equal to zero only if the two distributions are identical, and therefore indistinguishable, and the sample provides about the identity of the distribution. The Jeffreys and Jensen-Shannon entropies are related by the inequalities [46]

\[ 0 \leq JS(A; B) \leq \text{Jeffreys}(A; B) \, . \]

**General Jensen-Shannon divergence** (skewed Jensen-Shannon divergence)

\[ JS_\alpha(A; B) := (1 - \alpha)D(A \parallel M) + \alpha D(B \parallel M) \, , \quad \alpha \in [0, 1] \, , \quad (26) \]

\[ P(M) = (1 - \alpha)P(A) + \alpha P(B) \, . \]

Jensen-Shannon divergence) [30, 59]

\[ JS_\alpha(A; B) = D(A \parallel B) \, , \quad \alpha = 0 \, , \quad (27) \]

\[ JS_1(A; B) = JS(A; B) \, , \quad \alpha = 1 \, , \quad (28) \]

\[ JS_0(A; B) = D(B \parallel A) \, , \quad \alpha = 0 \, , \quad (29) \]

The entropy of a mixed distribution is the average entropy of the components plus the Jensen-Shannon entropy [53]:

\[ S(M) = JS_\alpha(A^1; A^2; \ldots ; A^n) + \sum_\alpha P(\Theta_\alpha) S(A^\alpha) \, , \quad (30) \]

The multi-variate Jensen-Shannon entropy is the mutual information between the mixing ensemble \( \Theta \) and the mixed ensemble \( M \):

\[ I(\Theta : M) := \sum_{\alpha, x} P(\Theta_\alpha, M_x) \ln \frac{P(\Theta_\alpha, M_x)}{P(\Theta_\alpha)P(M_x)} \]

\[ = - \sum_x P(M_x) \ln P(M_x) \]

\[ + \sum_{\alpha, x} P(\Theta_\alpha) P(M_x | \Theta_\alpha) \ln P(M_x | \Theta_\alpha) \]

\[ = S(M) - \sum_\alpha P(\Theta_\alpha) S(A^\alpha) \]

\[ = JS_\alpha(A^1; A^2; \ldots ; A^n) \, , \quad (31) \]

Resistor-average entropy [56]:

\[ \text{ResAvg}(A; B) := \frac{1}{D(A \parallel B) + D(B \parallel A)} \, , \quad (32) \]

The harmonic mean of forward and reversed relative entropies.

4 Rényi information

**Rényi information** (Rényi entropy, alpha-order entropy) [19]: A one parameter generalization of the Shannon entropy.

\[ \text{Rényi}_\alpha(A) := \frac{1}{1 - \alpha} \ln \sum_x P(A_x)^\alpha \, , \quad \alpha \neq 0 \, , \quad (33) \]

Interesting special cases of the Rényi information include the Hartley entropy (\( \alpha = 0 \)), collision entropy
(\(\alpha = 2\)), Shannon entropy (\(\alpha = 1\)), and min entropy (\(\alpha = \infty\)). See also: Rényi divergence (47).

\[
\text{Rényi}_\alpha(A) \geq \text{Rényi}_\beta(A), \quad \beta \geq \alpha
\]

**Collision entropy** (Rényi quadratic entropy, Rényi information of order 2, second order entropy) [48]

\[
\text{CollisionEntropy}(A) := -\ln \sum_x P(A_x)^2
= \text{Rényi}_2(A)
\]

A special case of the Rényi information. The negative log (62) probability that two independent samples from the distribution are the same.

**Min-entropy** [19]

\[
\text{MinEntropy}(A) := -\ln \max_x P(A_x)
\]

**Hartley entropy** (Hartley function, max-entropy, Boltzmann entropy) [7]: The logarithm of the number distinct possibilities.

\[
\text{Hartley}(A) := \ln |\Omega_A|
\]

The maximum entropy for a given cardinality. Coincides with the entropy for a uniform distribution.

**Tsallis information** [29, 31, 41] (Havrda-Charvát information, \(\alpha\) order information)

\[
\text{Tsallis}_\alpha(A) := \frac{1}{\alpha - 1} \left(1 - \sum_x P(A_x)^\alpha\right)
= \frac{1}{\alpha - 1} \left(e^{(\alpha - 1)\text{Rényi}_\alpha(A)} - 1\right)
\]

**Sharma-Mittal information** [34, 76]

\[
\text{SharmaMittal}_{\alpha,\beta}(A) := \frac{1}{\beta - 1} \left[1 - \left(\sum_x P(A_x)^\alpha\right)^{\frac{1-\beta}{\alpha}}\right]
\]

Assuming suitable limits are taken, the Sharma-Mittal information contains Shannon, Rényi and Tsallis informations as special cases.

\[
\text{SharmaMittal}_{1,1}(A) = S(A)
\]
\[
\text{SharmaMittal}_{1,1}(A) = \text{Rényi}_1(A)
\]
\[
\text{SharmaMittal}_{\alpha,\alpha}(A) = \text{Tsallis}_\alpha(A)
\]

### 5 Csizár f-divergences

**Csizár f-divergence** Many interesting divergence measures between probability distributions can be written as (or related to) an f-divergence (also know as Csizár\(^8\)), Csiszár-Morimoto, Ali-Silvey, or \(\phi\)-divergence) [27, 28, 26, 62, 70, 56].

\[
C_f(A; B) = \sum_x P(A_x) f\left(\frac{P(B_x)}{P(A_x)}\right)
\]

where the function \(f\) is convex \(\circ\) and \(f(1) = 0\). This implies \(C_f(A; B) \geq 0\) from an application of Jensen’s inequality. Examples already encountered include the relative, Jeffreys, and Jensen-Shannon entropies (see table 3). Note that the first argument to the f-divergence appears in the numerator of the ratio and is the distribution to be averaged over. The opposite convention also occurs\(^9\).

Convex functions are closed under conical combination: the function \(f(x) = c_1 f_1(x) + c_2 f_2(x) + \ldots + c_n f_n(x)\) is convex if each function of the mixture \(f_n\) is convex and each constant \(c_n \geq 0\) is non-negative. It follows that a positive linear sum of an f-divergences is also an f-divergence [74].

**Dual f-divergence** [74] The dual of an f divergence is defined by swapping the arguments.

\[
C_f(A; B) = C_f^*(B; A) = C_{f^*}(B; A)
\]

Here \(f^*\) is the **Csizár dual** [74] of a function \(f^*(x) = x f(1/x)\). For instance, if \(f(x) = -\ln(x)\) then the f divergence is the relative entropy \(C_f(A; B) = D(A \parallel B)\), with dual function \(f^*(x) = x \ln(x)\), and dual divergence the dual (or reverse) relative entropy, \(C_{f^*}(A; B) = D(B \parallel A)\).

**Symmetric f-divergences** [0] Many instances of the f-divergence are symmetric under interchange of the two distributions,

\[
C_f(A; B) = C_f(B; A)
\]

This implies that \(f(x) = x f(\frac{1}{x})\).

There are two common methods for symmetrizing an asymmetric f-divergence: the **Jeffreys symmetrization** where we average over interchanged distributions

\[
C_f(A; B) = \frac{1}{2} C_g(A; B) + \frac{1}{2} C_g(B; A)
\]
\[
f(t) = \frac{1}{2} g(t) + \frac{1}{2} g(t^{-1})
\]

and the **Jensen symmetrization** where we take the average divergence of both distributions to the average distribu-

---

\(^8\)Pronounced che-sar.

\(^9\)I, myself, have used the opposite convention on other occasions, but, on reflection, this way around makes more sense.
Table 3: Csiszár $f$-divergences (§5)

<table>
<thead>
<tr>
<th>Asymmetric $f$-divergences</th>
<th>$f(t)$</th>
<th>$f^*(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>relative entropy</td>
<td>$-\ln t$</td>
<td>$t \ln t$</td>
</tr>
<tr>
<td>K-divergence</td>
<td>$\ln \frac{2}{1 + t}$</td>
<td>$t \ln \frac{2t}{1 + t}$</td>
</tr>
<tr>
<td>Pearson divergence</td>
<td>$(t - 1)^2$</td>
<td>$(\frac{1}{\sqrt{t}} - \sqrt{t})^2$</td>
</tr>
<tr>
<td>Cressie-Read divergence</td>
<td>$\frac{t^{-\alpha} - 1}{\alpha(\alpha + 1)}$</td>
<td></td>
</tr>
<tr>
<td>Tsallis divergence</td>
<td>$\frac{t^{1-\alpha} - 1}{\alpha - 1}$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Symmetric $f$-divergences</th>
<th>$f(t)$</th>
<th>$h(t)$</th>
<th>$g(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeCam discrimination</td>
<td>$\frac{(t - 1)^2}{[t + 1]}$</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Jeffreys entropy</td>
<td>$(t - 1) \ln t$</td>
<td>$-\ln t$</td>
<td></td>
</tr>
<tr>
<td>Jensen-Shannon entropy</td>
<td>$\frac{1}{2} \ln \frac{2}{1 + t} + \frac{1}{2} t \ln \frac{2t}{1 + t}$</td>
<td>$\ln \frac{2}{1 + t}$</td>
<td>$-\ln t$</td>
</tr>
<tr>
<td>variational distance</td>
<td>$\frac{1}{2}</td>
<td>t - 1</td>
<td>$</td>
</tr>
<tr>
<td>Hellinger discrimination</td>
<td>$1 - \sqrt{t}$</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
We include the half’s in these definitions so that symmetrized f-divergences are invariant under symmetrization. We call these the Jeffreys and Jensen symmetrizations respectively [1], because a Jeffreys symmetrization of the relative entropy (16) is half the Jeffreys entropy (24), and a Jensen symmetrization gives the Jensen-Shannon entropy (25).

\[ C_r(A;B) = \frac{1}{2} C_h(A;M) + \frac{1}{2} C_h(B;M) \]

\[ P(M_x) = \frac{1}{2} P(A_x) + \frac{1}{2} P(B_x) \]

\[ f(t) = \frac{1}{2} \ln \left( \frac{1}{1 + \frac{1}{2} t} \right) + \frac{1}{2} \ln \left( \frac{1}{1 + \frac{3}{2} t^{-1}} \right) \]

K-divergence [46]

\[ \text{KayDiv}(A;B) := \sum_x P(A_x) \ln \frac{P(A_x)}{\frac{1}{2} \left( P(A_x) + P(B_x) \right)} \] (37)

\[ = D(A \parallel M) + P(M_x) = \frac{1}{2} P(A_x) + \frac{1}{2} P(B_x) \]

\[ = C_r(A;B), \quad f(t) = \ln \frac{2}{1 + t} . \]

Of interest since the K-divergence is a lower bound to the relative entropy [46],

\[ \text{KayDiv}(A;B) \leq \frac{1}{2} D(A \parallel B) \]

and the Jeffreys symmetrization of the K-divergence is the Jensen-Shannon entropy.

\[ JS(A;B) = \frac{1}{2} \text{KayDiv}(A;B) + \frac{1}{2} \text{KayDiv}(B;A) \]

\[ = \frac{1}{2} D(A \parallel M) + \frac{1}{2} D(B \parallel M) \]

Fidelity (Bhattacharyya coefficient, Hellinger affinity) [9] The Bhattacharyya distance (60) and the Hellinger divergence and distance (39) are functions of fidelity. The name derives from usage in quantum information theory [49].

\[ \text{Fidelity}(A;B) := \sum_x \sqrt{P(A_x)P(B_x)} \] (38)

The range is [0, 1], with unity only if the two distributions are identical. Fidelity is not itself an f-divergence (The required function f(t) = \sqrt{t} isn’t convex), but is directly related to the Hellinger divergence (39) and Bhattacharyya distance (60).

**Hellinger discrimination** (Squared Hellinger distance, infidelity) [11]

\[ \text{HellingerDiv}(A;B) := \frac{1}{2} \sum_x \left( \sqrt{P(A_x)} - \sqrt{P(B_x)} \right)^2 \] (39)

\[ = \sum_x P(A_x) \left( 1 - \sqrt{\frac{P(B_x)}{P(A_x)}} \right) \]

\[ = C_r(A;B), \quad f(t) = (1 - \sqrt{t}) \]

\[ = 1 - \text{Fidelity}(A;B) \]

Symmetric, with range [0, 1]. A common alternative normalization omits the one-half prefactor. The name originates from that of the corresponding integral in the continuous case [6].

**Pearson divergence** (\(\chi^2\)-divergence, chi square divergence, Pearson chi square divergence, Kagan divergence, quadratic divergence, least squares) [4]

\[ \text{Pearson}(A;B) := \sum_x \frac{(P(B_x) - P(A_x))^2}{P(B_x)} \] (40)

\[ = \sum_x P(A_x) \left( \frac{P(B_x)}{P(A_x)} - 1 \right)^2 \]

\[ = C_r(A;B), \quad f(t) = (t - 1)^2 \]

**Neyman divergence** (inverse Pearson chi square divergence) [13, 39]

\[ \text{Neyman}(A;B) := \text{Pearson}(B;A) \] (41)

The dual of the Pearson divergence (arguments switched).

**LeCam discrimination** (LeCam discrimination, Vincze-LeCam divergence, triangular discrimination) [38, 40, 53]

\[ \text{LeCam}(A;B) := \sum_x \frac{(P(A_x) - P(B_x))^2}{P(A_x) + P(B_x)} \] (42)

\[ = 2C_r(A;B), \quad f(t) = \frac{(t - 1)^2}{2(t + 1)} . \]

The LeCam discrimination is a Jensen symmetrized Pearson divergence.

\[ \text{LeCam}(A;B) := \frac{1}{2} \text{Pearson}(A;M) + \frac{1}{2} \text{Pearson}(B;M) \]

\[ P(M_x) = \frac{1}{2} P(A_x) + \frac{1}{2} P(B_x) \]

The triangular discrimination is defined as twice the LeCam discrimination [0].

The range is [0, 1], with unity only if the two distributions are identical. Fidelity is not itself an f-divergence (The required function f(t) = \sqrt{t} isn’t convex), but is directly related to the Hellinger divergence (39) and Bhattacharyya distance (60).
### Skewed K-divergence [46]

\[
\text{KayDiv}_\alpha(A; B) := \sum_x P(A_x) \ln \frac{P(A_x)}{(1 - \alpha)P(A_x) + \alpha P(B_x)}
\]

\[
= D(A \parallel M),
\]

\[
P(M) = (1 - \alpha)P(A) + \alpha P(B)
\]

\[\text{Alpha-Jensen-Shannon-entropy} [46, 71] \]

\[
\text{AlphaJS}_\alpha(A; B) := \frac{1}{2} \text{KayDiv}_\alpha(A; B) + \frac{1}{2} \text{KayDiv}_\alpha(B; A)
\]

\[
\text{AlphaJS}_0(A; B) = 0
\]

\[
\text{AlphaJS}_1(A; B) = \text{JS}(A; B)
\]

\[
\text{AlphaJS}_1(A; B) = \text{Jeffreys}(A; B)
\]

The Jeffreys symmetrization of the skewed K-divergence.

### 6 Chernoff divergence

**Chernoff divergence** The Chernoff divergence [15, 51] of order \(\alpha\) is defined as

\[
\text{Chernoff}_\alpha(A; B) := -\ln \sum_x P(A_x) \left( \frac{P(A_x)}{P(B_x)} \right)^{\alpha - 1}
\]

\[
= -\ln \left[ C_f(A; B) + 1 \right], \quad f(t) = t^{1-\alpha} - 1.
\]

The Chernoff divergence is zero for \(\alpha = 1\) and \(\alpha = 0\), and reaches a maximum, the **Chernoff information** [15, 63], for some intermediate value of alpha. The Chernoff divergence is well defined for \(\alpha > 1\) if \(P(B_x) > 0\) whenever \(P(A_x) > 0\), and for \(\alpha < 0\) if \(P(A_x) > 0\) whenever \(P(B_x) > 0\), and thus defined for all \(\alpha\) if the distributions have the same support.

The Chernoff divergence of order \(\alpha\) is related to the Chernoff divergence of order \(1 - \alpha\) with the distributions interchanged [51],

\[
\text{Chernoff}_\alpha(A; B) = \text{Chernoff}_{1-\alpha}(B; A).
\]

This relation always holds for \(\alpha \in [0, 1]\), and for all \(\alpha\) when the distributions have the same support.

**Chernoff coefficient (alpha divergence) [15, 51, 52]**

\[
\text{ChernoffCoefficient}_\alpha(A; B) := \sum_x P(A_x) \left( \frac{P(A_x)}{P(B_x)} \right)^{\alpha - 1}
\]

\[
= \exp(- \text{Chernoff}_\alpha(A; B))
\]

\[
= C_f(A; B) + 1, \quad f(t) = t^{1-\alpha} - 1
\]

The exponential twist density [42] is way of mixing two distributions [55] to form a third

\[
P(C_x) = \frac{1}{Z_\alpha} P(A_x)^\alpha P(B_x)^{1-\alpha}
\]

Here \(\alpha\) is a mixing parameter between 0 and 1. The normalization constant \(Z_\alpha\) is the Chernoff coefficient, \(Z_\alpha = \text{ChernoffCoefficient}_\alpha(A; B)\) [79].

**Rényi divergence** The Rényi divergence (or relative Rényi entropy) of order \(\alpha\) is a one-parameter generalization of the relative entropy [19],

\[
\text{Rényi}_\alpha(A; B) := \frac{1}{\alpha - 1} \ln \sum_x P(A_x) \left( \frac{P(A_x)}{P(B_x)} \right)^{\alpha - 1}
\]

\[
= \frac{1}{1 - \alpha} \text{Chernoff}_\alpha(A; B)
\]

\[
= \frac{1}{\alpha - 1} \ln \left[ C_f(A; B) + 1 \right], \quad f(t) = t^{1-\alpha} - 1.
\]

Higher values of \(\alpha\) give a Rényi divergence dominated by the greatest ratio between the two distributions, whereas as \(\alpha\) approaches zero the Rényi entropy weighs all possibilities more equally, regardless of their dissimilarities. We recover the relative entropy in the limit of \(\alpha \to 1\).

Interesting special cases of the Rényi divergence occur for \(\alpha = 0, \frac{1}{2}, 1\) and \(\infty\). As previously mentioned, \(\alpha = 1\) gives the relative entropy (16), and \(\alpha = \frac{1}{2}\) gives the Bhattacharyya distance (60). In the limit \(\alpha \to 0\), the Rényi divergence slides to the negative log probability under \(p\) that \(p\) is non-zero,

\[
\lim \limits_{\alpha \to 0} \text{Rényi}_\alpha(A; B) = -\ln \sum_x \lim \limits_{\alpha \to 0} P(A_x)^\alpha P(B_x)^{1-\alpha}
\]

\[
= -\ln \sum_x P(B_x)P(A_x) > 0
\]

Here we have used the Iverson bracket, \([a]\), which evaluates to 1 if the condition inside the bracket is true, and 0 otherwise. If the two distributions have the same support then in the \(\alpha \to 0\) Rényi divergence is zero.

**Alpha-divergence [0, 0]** In the most widespread parameterization,

\[
D_\alpha(A; B) := \frac{1}{\alpha(1 - \alpha)} \left( 1 - \sum_x P(A_x)^\alpha P(B_x)^{1-\alpha} \right)
\]

The alpha-divergence is self-dual,

\[
D_\alpha(A; B) = D_{1-\alpha}(B; A)
\]

Special cases include the Neyman and Pearson divergences, the Hellinger discrimination, and the relative en-
The Sharma-Mittal divergence encompasses Cressie-Read divergence ($\beta = 1 - \alpha (\alpha + 1)$), Rényi divergence ($\beta \to 1$), Tsallis divergence ($\beta \to 0$), and the relative entropy ($\beta, \alpha \to 1$).

7 Cauchy-Schwarz divergence

Cauchy-Schwarz divergence [64, 73, 76]

Cauchy-Schwarz divergence

$$D_{\alpha'}(A; B) := \frac{4}{1 - \alpha'^2} \left(1 - \sum_x P(A_x) \frac{x - 1}{2} P(B_x) \frac{x - 1}{2}\right)$$

where $\alpha' = 1 - 2\alpha$. This parameterization has the advantage that the duality corresponds to negating the parameter.

$$D_{+\alpha'}(A; B) = D_{-\alpha'}(B; A)$$

Cressie-Read divergence [39]

CressieRead$\alpha$(A; B) := \frac{1}{\alpha(\alpha + 1)} \sum_x P(A_x) \left[\left(\frac{P(A_x)}{P(B_x)}\right)^{\alpha} - 1\right]$

$$\alpha = 1$$

$$= \frac{1}{\alpha + 1} [e^{\alpha \text{Renyi}_{\alpha + 1}(A; B)} - 1]$$

$$= \frac{1}{\alpha + 1} \text{Tsallis}_{\alpha + 1}(A; B)$$

$$= \text{Cf}(A; B), \quad f(t) = \frac{t^{-\alpha} - 1}{\alpha(\alpha + 1)}.$$}

Tsallis divergence (relative Tsallis entropy) [50] Other closely related divergences include the relative Tsallis entropy,

$$\text{Tsallis}_{\alpha}(A; B) := \frac{1}{\alpha - 1} \sum_x P(A_x) \left[\left(\frac{P(A_x)}{P(B_x)}\right)^{\alpha - 1} - 1\right]$$

$$= \frac{1}{\alpha - 1} [e^{(\alpha - 1) \text{Renyi}_{\alpha}(A; B)} - 1]$$

$$= \text{Cf}(A; B), \quad f(t) = \frac{t^{1-\alpha} - 1}{\alpha - 1},$$

Sharma-Mittal divergence [34, 76]

SharmaMittal$\alpha, \beta$(A; B)

$$:= \frac{1}{\beta - 1} \left[1 - \left(\sum_x P(A_x)^\alpha P(B_x)^{1-\alpha}\right)^{\frac{1}{1-\beta}}\right]$$

$$= \frac{1}{1 - \beta} \left(1 - \text{ChernoffCoefficient}_{\alpha}(A; B)^{\frac{1}{1-\beta}}\right)$$

$$\alpha > 0, \alpha \neq 0, \beta \neq 0$$

The only f-divergence (35) which is also a metric [66]. Pinsker’s inequality:

$$D(A \parallel B) \geq \frac{1}{2} V(A; B)^2$$

8 Distances

Variational distance (L1 distance, variational divergence, Kolmogorov distance) [0, 0]

$$V(A; B) = L_1(A; B) := \frac{1}{2} \sum_x |P(B_x) - P(A_x)|$$

$$= \text{Cf}(A; B), \quad f(t) = |t - 1|$$

The only $f$-divergence which is also a metric. Pinsker’s inequality:

$$D(A \parallel B) \geq \frac{1}{2} V(A; B)^2$$

Total variational distance [0] The largest possible difference between the probabilities that the two distributions can assign to the same event. Equal to twice the Variational distance

Euclidian distance (L2 distance)

$$L_2(A; B) := \sqrt{\sum_x |P(B_x) - P(A_x)|^2}$$

$$\alpha > 0, \alpha \neq 0, \beta \neq 0$$
It is sometimes useful to treat a probability distribution as a vector in a Euclidean vector space, and therefore consider Euclidean distances between probability distributions.

**Minkowski distance**

\[ L_p(A; B) := \left( \sum_x |P(B_x) - P(A_x)|^p \right)^{\frac{1}{p}} \quad (56) \]

A metric distance provided that \( p \geq 1 \).

**Chebyshev distance**

\[ L_\infty(A; B) := \max_x |P(B_x) - P(A_x)| \quad (57) \]

**LeCam distance** The square root of the LeCam discrimination (42) [45].

\[ \text{LeCamDist}(A; B) := \sqrt{\frac{1}{2} \text{LeCam}(A; B)} \quad (58) \]

**Hellinger distance** The square root of the Hellinger divergence.

\[ \text{Hellinger}(A; B) := \sqrt{\text{HellingerDiv}(A; B)} \quad (59) \]

**Jensen-Shannon distance** is the square root of the Jensen-Shannon divergence, and is a metric between probability distributions [59, 57].

**Bhattacharyya distance** [8] The Chernoff divergence of order \( \alpha = \frac{1}{2} \). The negative logarithm of the Bhattacharyya coefficient (fidelity).

\[ \text{Bhattacharyya}(A; B) := -\ln \sum_x \sqrt{P(A_x)P(B_x)} \quad (60) \]

\[ = \text{Chernoff}_2(A; B) \]

\[ = -\ln \text{Fidelity}(A; B) . \]

**Bhattacharyya angle** (statistical angle) [0] The inverse cosine of the Bhattacharyya coefficient (fidelity).

\[ \text{Bhattacharyya}(A; B) := \arccos \sum_x \sqrt{P(A_x)P(B_x)} \quad (61) \]

\[ = \arccos \text{Fidelity}(A; B) . \]

9 **Specific information**

A specific (or point-wise, or local) entropy is the entropy associated with a single event, as opposed to the average entropy over the entire ensemble [21].

A common convention is to use lower cased function names for specific information measures; \( s \) for \( S \), \( i \) for \( I \). However, since the expectation of a single proposition is equal to the value associated with the event, we can also write \( S(A_x) := s(A_x) \), and \( I(A_x : B_y) := i(A_x : B_y) \). With this notation we can express point-wise measures corresponding to all the ensemble measures defined previously, without having to create a host of new notation. We can also write multivariate specific informations averaged over only one ensemble, not both, e.g. \( I(A : B_x) \) [18].

**Specific entropy** (information content, self-information, score, surprise, surprisal) [16, 24] is the negative logarithm of a probability.

\[ s(A_x) := -\ln P(A_x) \quad (62) \]

The expectation of the specific entropy is the Shannon entropy, the average point-wise entropy of the ensemble (2).

\[ S(A) = \mathbb{E}[s(A)] = \sum_x P(A_x) s(A_x) \]

**Local conditional entropy** [0]

\[ S(A \mid B_b) := -\sum_a P(A_a \mid B_b) \ln P(A_a \mid B_b) \]

\[ = \mathbb{E}_A [-\ln P(A \mid B)] \quad (63) \]

The entropy of \( A \) conditioned on a particular proposition from ensemble \( B \). A partially localized specific information.

**Specific mutual information** (point-wise mutual information, local mutual information) [18, 21, 43]

\[ i(A_x : B_y) := \frac{P(A_x, B_y)}{P(A_x)P(B_y)} \quad (65) \]

Similarly, the expectation of the specific mutual information is the mutual information of the ensemble (6).

\[ I(A : B) = \mathbb{E}[i(A : B)] \]

**Back Matter**

**Acknowledgments**

Useful sources and pedagogical expositions include Cover and Thomas’s classic Elements of Information Theory [63], David MacKay’s Information Theory, Inference, and Learning Algorithms [81], the inestimable (for some value of truthiness) Wikipedia, and various other reviews [65, 72]. Notation and nomenclature was adapted from Good
This review is inevitably incomplete, inaccurate and otherwise imperfect — *caveat emptor*.

**Version history**

0.9 (2024-03-02) Fix error in definition of Pearson divergence. Miscellaneous minor improvements.

0.8 (2021-12-10) Added Cauchy-Schwarz divergence and angle; Bhattacharyya angle; perplexity; conditional perplexity. Miscellaneous improvements.

0.7 (2018-09-22) Added units of entropy. Corrected typos and citations. (Kudos: Glenn Davis, Susanna Still)

0.6 (2017-05-07) Added alpha-divergence (48) and Burg entropy (18). Corrected typos. (Kudos: Glenn Davis)

0.5 (2016-08-16) Extensive miscellaneous improvements.


0.3 (2016-04-21) Added residual entropy (12) (and combined with independent information, which is essentially another name for the same thing), and uncertainty coefficient (15). Added information diagrams. Miscellaneous minor improvements. Fixed sign error in lautum information (14) (Kudos: Ryan James)

0.2 (2015-06-26) Improved presentation and fixed miscellaneous minor errors.

0.1 (2015-01-22) Initial release: over 50 information measures described.

**References**

(Recursive citations mark neologisms and other innovations [1].)

[0] [citation needed]. (pages 12, 14, 15, 15, 16, 16, 16, 16, 17, and 17).


Copyright © 2015-2024 Gavin E. Crooks

http://threeplusone.com/info
typeset on 2024-03-06 with XeTeX version 0.999995
fonts: Trump Mediaeval (text), Euler (math) 2 7 1 8 2 8 1 8 3