On Measures of Entropy and Information

Tech. Note 009 v0.8
http://threeplusone.com/info

Gavin E. Crooks

2021-12-10

A brief survey of information measures for classical, discrete probability distributions.

Contents

0 Notes on notation and nomenclature

Information measures An *information measure* is, broadly speaking, any function of one or more probability distributions. We'll specifically reserve the term *entropy* for measures that have units of entropy — negative logarithms of probabilities, or a linear combination of the same. The units of entropy depend on the base of the logarithm: for instance 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, $0 \ln 0 = 0$. In many cases, tradition doth dictate that an entropic measure be referred to as an information, e.g. "mutual information" rather than "mutual entropy". But we'll eschew the not uncommon practice of using the term "entropy" to refer to generic information measures taht do not have units of entropy.

Ensembles We'll use upper case letters for *ensembles*, A, B, C, etc. An ensemble (or *probability space*) consists of a complete and mutually exclusive set of propositions, $\{A_a\}$, indexed by the *sample space* $a \in \Omega_A$. The probability of a proposition is written as $P(A_a)$, and the conditional probability as $P(A_a \mid B_b)$. For instance, A_a could represent the propositions $\{A = a\}$, where A is a random variable. Any sum is implicitly over the entire relevant sample space. When different ensembles have different samples spaces we'll use corresponding lower case letters for samples, a for A, b for B, and so on. But when ensembles share the same sample space, we'll use some other lower case letter, typically x, y, z. For instance, compare the definitions of mutual information (5) (different sample spaces) with relative entropy (15) (same sample spaces). We will not employ the common shorthand of labeling distributions by the sample alone, e.g. $P(x)$ for $P(\{A = x\})$, precisely because we need to deal with multiple distributions with the same sample space in the same expression.

Expectation [0] The expectation operator $\mathbb{E}$ denotes the ensemble average of a function.

$$\mathbb{E}\big[f(A)\big] = \sum_a P(A_a)\, f(A_a)$$

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. $P(a)$ for $P(A_a)$.

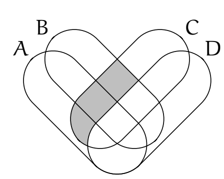
Most of the information measures discussed herein could be expressed using the expectation operator (but for maximum clarity we make the averages explicit). For example the Shannon entropy (1) is the ensemble average of the specific (or point-wise) entropy $-\ln P(A_a)$. (§9).

$$S(A) = \mathbb{E}\big[-\ln P(A_x)\big] = -\sum_x P(A_x)\, \ln P(A_x)$$

Collections of ensembles Since we have used subscripts to index propositions, we'll use superscripts to index collections of ensembles, $\mathcal{A} = \{A^1, A^2, \ldots, A^{|\mathcal{A}|}\}$. We never need to take powers of ensembles or propositions, so there is little risk of ambiguity. Other useful notation includes $\mathcal{P}(\mathcal{A})$, the power set of $\mathcal{A}$ (the set of all subsets); $|\mathcal{A}|$, the set cardinality; $\varnothing$ for the empty set; and $\mathcal{A} \setminus A$, 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 conjugation (logical and); a colon ':' the mutual information (5) "between" ensembles; a double bar '$\|$' the relative entropy (15) 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) ',', ':', '$|$', '$\|$', ';' 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* [2, 3].



$I(A, B : C \mid D)$

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 conditional '$A|B$' to set complement $A \setminus B$. For instance, the conditional mutual information $I(A, B : C \mid D)$ corresponds to the region $((A \cup B) \cap C) \setminus D$.

Table 1: Information measure symbology

symbol	usage	commutative	precedence
,	conjugation	yes	high
:	mutual	yes	:
|	conditional	no	:
||	relative entropy	no	:
;	divergence	no	low

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*.[1] 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, $d(A;B) + d(B;C) \geqslant d(A;C)$.

1 Entropy

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

$$S(A) := -\sum_{a} P(A_a) \ln P(A_a) \tag{1}$$

In information theory the entropy is typically denoted by the symbol H, a notation that dates back to Boltzmann and his H-theorem [4], and adopted by Shannon [5]. The notation S is due to Clausius and the original discovery of entropy in thermodynamics [6], and adopted by Gibbs [7] 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.

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 (61) which is also sometimes referred to as the self-information.

Discrete entropies are non-negative and bounded.

$$0 \leqslant S(A) \leqslant \ln |\Omega_A|$$

[1] The terms *disparity*, *discrimination*, and *divergence* are used essentially interchangeably as synonymies of dissimilarity. The useful distinction that a discrimination is a symmetric divergence appears to be novel, but consistent with practical usage in much of the literature.

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

$$S(A, B) := -\sum_{a,b} P(A_a, B_b) \ln P(A_a, B_b) \tag{2}$$

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

$$S(A^1, A^2, \ldots, A^n) = -\sum_{a_1,a_2,\ldots,a_n} P(A^1_{a_1}, A^2_{a_2}, \ldots, A^n_{a_n}) \ln P(A^1_{a_1}, A^2_{a_2}, \ldots, A^n_{a_n})$$

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) [5, 8] Measures how uncertain we are of A on the average when we know B.

$$S(A \mid B) := -\sum_{b} P(B_b) \sum_{a} P(A_a \mid B_b) \ln P(A_a \mid B_b) \tag{3}$$

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

The *chain rule for entropies* [5, 8] 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, $S(A \mid B) \leqslant S(A)$. 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).

$$S(A, B) \leqslant S(A) + S(B)$$

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)

$$S(A \mid B) = S(B \mid A) + S(A) - S(B) \ .$$

Perplexity [9]

$$\text{Perplexity}(A) := e^{S(A)} \tag{4}$$

The exponentiation of the entropy (with the same base). We can also define perplexities corresponding to other entropic measures, such as the conditional perplexity

Table 2: Units of entropy

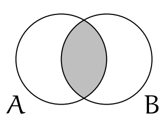
deciban	$\frac{1}{10}\log_2(10) \approx 0.33$ bits	tenth of a ban
bit (shannon)	1 bit	
nat (nit, nepit)	$\log_2(e) \approx 1.44$ bits	natural digit
trit	$\log_2(3) \approx 1.6$ bits	ternary digit
quad	2 bits	
ban (hartly)	$\log_2(10) \approx 3.32$ bits	decimal digit
nibble (nybble)	4 bits	half a byte
byte	8 bits	

$\text{Perplexity}(A \mid B) := e^{S(A|B)}$. Partition functions in statistical mechanics are perplexities.

2 Mutual information

Mutual information (mutual entropy, transinformation)[2] [5, 8]

$$I(A : B) := \sum_{a,b} P(A_a, B_b) \ln \frac{P(A_a, B_b)}{P(A_a)P(B_b)} \tag{5}$$



Mutual information is oft notated with a semicolon, rather than a colon [5, 11].

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

$$\begin{aligned} I(A : B) &= S(A) - S(A \mid B) \\ &= S(B) - S(B \mid A) \\ &= S(A) + S(B) - S(A, B) \\ &= S(A, B) - S(A \mid B) - S(B \mid A) \end{aligned}$$

Mutual information is non-negative.

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

This also implies that entropy is *subadditive*: The sum of marginal entropy of two systems is less than, or equal to, the joint entropy.

$$S(A) + S(B) \leqslant S(A, B)$$

The mutual information is zero if (and only if) A and B are *independent* (written $A \perp\!\!\!\perp B$), such that $P(A_a, B_b) =$

[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 [10]

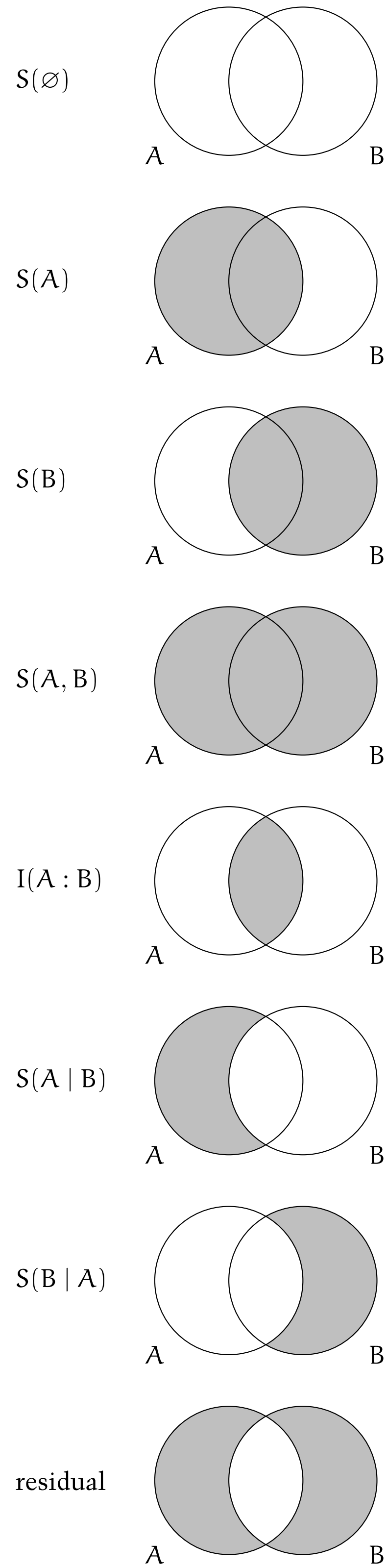


Figure 1: Two-variable Information diagrams [2, 3].

$P(A_a)P(B_a)$. And easy proof is to note that the mutual information can be written as a relative entropy (15).

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) [11, 12, 13, 14, 15, 3]: A multivariate generalization of the mutual information. Given a collection of probability ensembles, $\mathcal{A} = \{A^1, A^2, \cdots, A^{|\mathcal{A}|}\}$, the multivariate mutual information is equal to an alternating signed sum of all the marginal entropies.

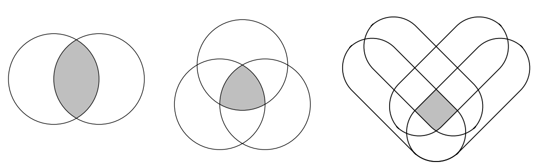
$$\begin{aligned} &I(A^1 : A^2 : \cdots : A^{|\mathcal{A}|}) \\ &\qquad := -\sum_{\mathcal{B}\in\mathcal{P}(\mathcal{A})} (-1)^{|\mathcal{B}|}\, S(B^1, B^2, \ldots, B^{|\mathcal{B}|}) \end{aligned} \tag{6}$$

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 (8)).

The single variable case is equal to the entropy, $I(A) = S(A)$, the binary case is equal to the standard mutual information (6), and the ternary case is

$$\begin{aligned} &I(A : B : C) \\ &:= \sum_{a,b,c} P(A_a, B_b, C_c) \ln \frac{P(A_a, B_b)P(A_a, C_c)P(B_b, C_c)}{P(A_a, B_b, C_c)P(A_a)P(B_b)P(C_c)} \end{aligned} \tag{7}$$

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.

$$\begin{aligned} S(A) &= I(A) \\ S(A, B) &= I(A) + I(B) - I(A : B) \\ S(A, B, C) &= I(A) + I(B) + I(A) \\ &\quad - I(A : B) - I(B : C) - I(A : C) \\ &\quad + I(A : B : C) \end{aligned}$$

Or generally,

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

The triplet interaction information is the information that a pair of variable provides about the third, compared to the information that each provides separately [11, 12, 16].

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

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

$$I(A : A : A : \cdots : A) = S(A)$$

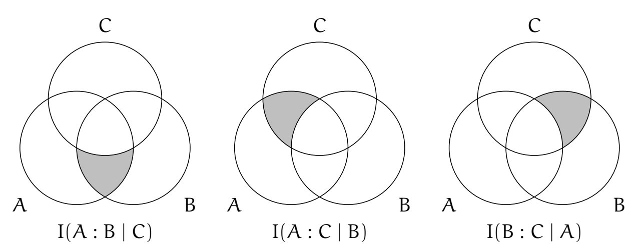
Interaction information (synergy, mutual information) [11]: An alternative sign convention for multivariate mutual information. The interaction information is equal in magnitude to the multivariate information, but has the opposite sign for odd number of ensembles.

$$\mathrm{Int}(A^1 : A^2 : \cdots : A^n) := (-1)^n I(A^1 : A^2 : \cdots : A^n) \tag{8}$$

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

Conditional mutual information [13] The average mutual information between A and B given C.

$$\begin{aligned} &I(A : B \mid C) \\ &:= \sum_{c} P(C_c) \sum_{a,b} P(A_a, B_b \mid C_c) \ln \frac{P(A_a, B_b \mid C_c)}{P(A_a \mid C_c)P(B_b \mid C_c)} \end{aligned} \tag{9}$$



$$\begin{aligned} I(A : B \mid C) &= S(A \mid C) - S(A \mid B, C) \\ &= S(B \mid C) - S(B \mid A, C) \end{aligned}$$

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

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

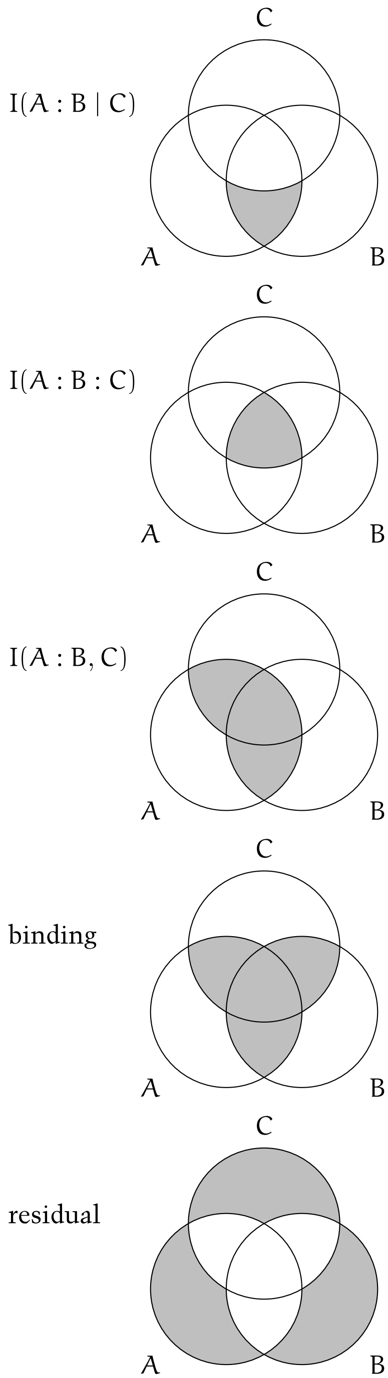
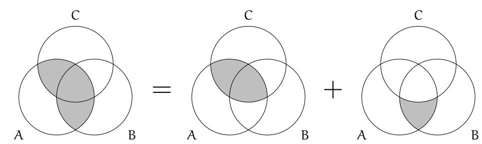


Figure 2: Three-variable information diagrams [2, 3]. Note that the central region $I(A : B : C)$ can be positive or negative (these are not Venn diagrams) and that we shade regions in proportion to their multiplicity, e.g. total correlation.

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

The *chain rule* for mutual information is

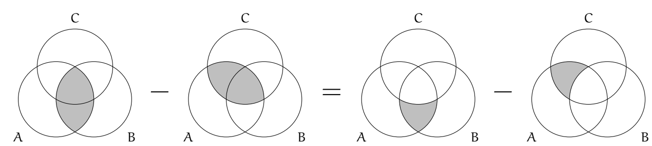
$$I(A : B, C) = I(A : C) + I(A : B \mid C)$$



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

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

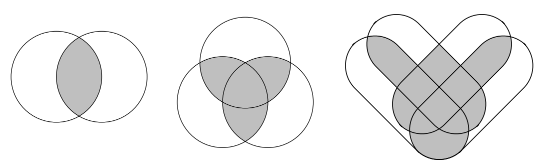
Proof: $I(A : B) - I(A : C) = I(A : B \mid C) - I(A : C \mid B)$, but $I(A : C \mid B)$ is zero, and $I(A : B \mid C)$ is positive.



Binding information (Dual total correlation) [17, 3, 18]

$$\mathrm{Binding}(\mathcal{A}) := S(\mathcal{A}) - \sum_{A \in \mathcal{A}} S(A \mid \mathcal{A} \setminus A) \tag{10}$$

Here $\mathcal{A} \setminus A$ is the set compliment of A.

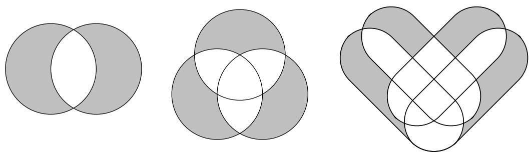


$$\begin{aligned}
\mathrm{Binding}(A : B) &= I(A : B) \\
\mathrm{Binding}(A : B : C) &= I(A, B : A, C : B, C) \\
\mathrm{Binding}(A : B : C : D) &= I(A, B, C : A, B, D : A, C, D : B, C, D)
\end{aligned}$$

Residual entropy (erasure entropy, independent information, variation of information, shared information distance) [19, 20, 21, 3]

$$\begin{aligned}
\mathrm{Residual}(\mathcal{A}) &:= \sum_{A \in \mathcal{A}} S(A \mid \mathcal{A} \setminus A) \qquad (11) \\
&= S(\mathcal{A}) - \mathrm{Binding}(\mathcal{A})
\end{aligned}$$

Measures the total amount of randomness localized to individual variables.



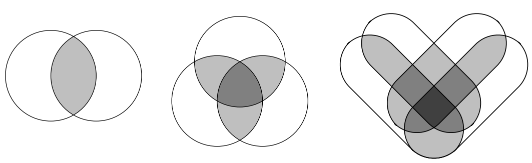
We can express the residual entropy as a Shannon information measure if we extend the notation to allow colons ':' in conditionals (i.e. on right side of a bar '|') [1], e.g.

$$\mathrm{Residual}(A : B : C) = I(A, B, C \mid A, B : A, C : B, C)$$

Total correlation (Multi-information, multivariate constraint, redundancy, integration) [11, 12, 22, 14]

$$\begin{aligned} &\mathrm{TotalCorr}(A^1, A^2, \ldots, A^n) := \\ &\quad S(A^1) + S(A^2) + \cdots + S(A^n) - S(A^1, A^2, \ldots, A^n) \end{aligned} \tag{12}$$

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 [23] :

$$\mathrm{Lautum}(A; B) := \sum_{a,b} P(A_a)P(B_b) \ln \frac{P(A_a)P(B_b)}{P(A_a, B_b)} \tag{13}$$

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) [24] [3]

$$\mathrm{UncertaintyCoeff}(A; B) := \frac{I(A : B)}{S(A)} = 1 - \frac{S(A \mid B)}{S(A)} \tag{14}$$

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) [25, 8][5]

$$D(A \parallel B) := \sum_x P(A_x) \ln \frac{P(A_x)}{P(B_x)} \tag{15}$$

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 [8]: 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 (5) is the relative entropy between the joint and marginal product distributions. Let the random variables $(\widehat{A}, \widehat{B})$ be independent, but with the same marginals as (A, B), i.e. $P(\widehat{A}, \widehat{B}) = P(A)P(B)$. Then

$$I(A : B) = D(A, B \parallel \widehat{A}, \widehat{B})$$

Similarly, for three or more variables, the relative entropy between the joint and marginal product distributions is the total correlation (12).

$$\begin{aligned} &\mathrm{TotalCorr}(A^1, A^2, \ldots, A^n) \\ &\qquad = D(A^1, A^2, \ldots, A^n \parallel \widehat{A}^1, \widehat{A}^2, \ldots, \widehat{A}^n) \end{aligned}$$

The Lautum information (13) is

$$\mathrm{Lautum}(A : B) = D(\widehat{A}, \widehat{B} \parallel A, B)$$

Cross entropy (inaccuracy) [26, 27]:

$$\begin{aligned} \mathrm{CrossEnt}(A; B) &:= -\sum_x P(A_x) \ln P(B_x) \\ &= S(A) + D(A \parallel B) \end{aligned} \tag{16}$$

The cross entropy measures the average number of bits needed to identify events that occur with probability $P(A_x)$, if a coding scheme is used that is optimal for the probability distribution $P(B_x)$.

[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 [8], many authors instead directly supply the distributions as arguments, e.g. $D(p(x) \| q(x))$.

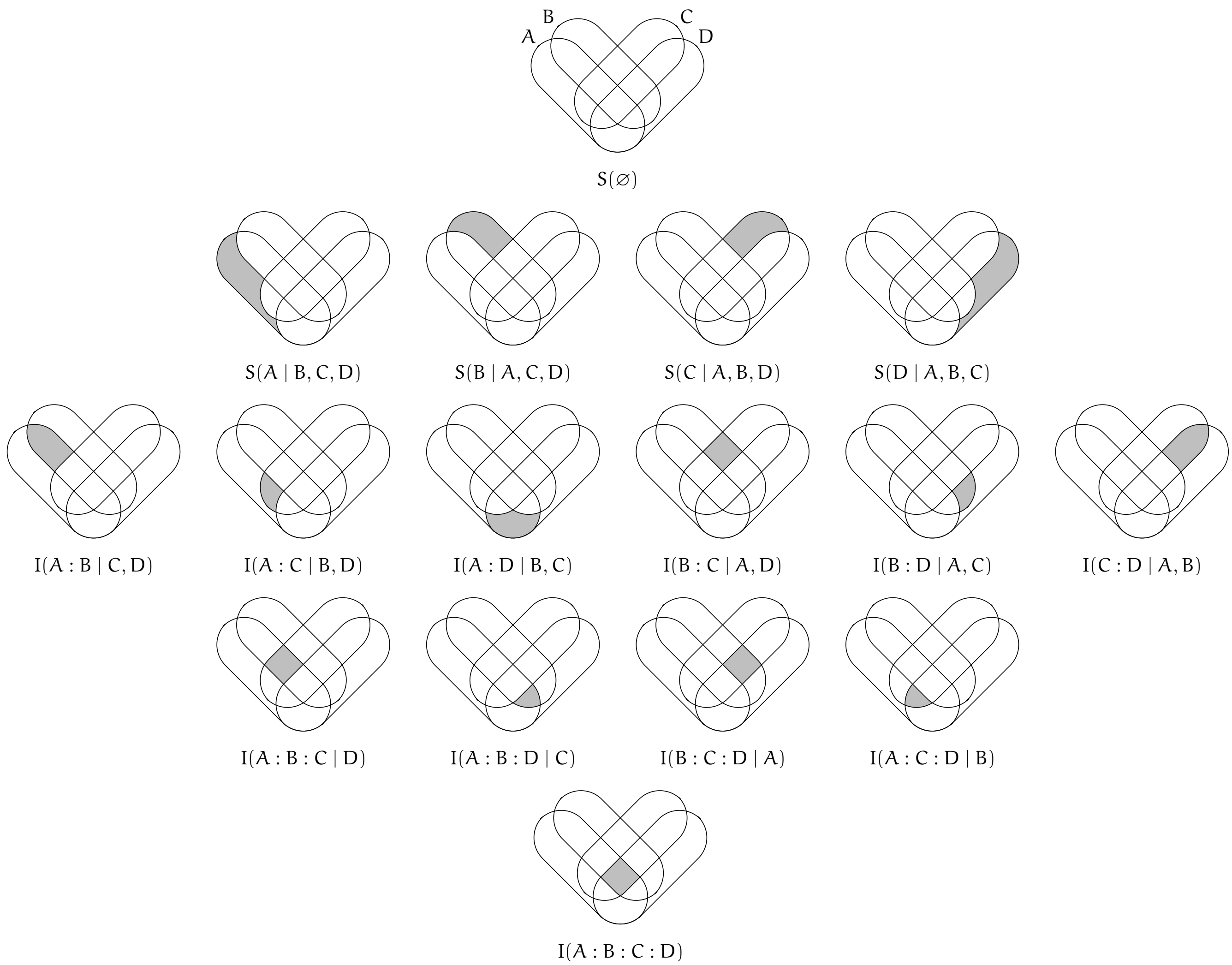


Figure 3: Components of four-variable information diagrams.

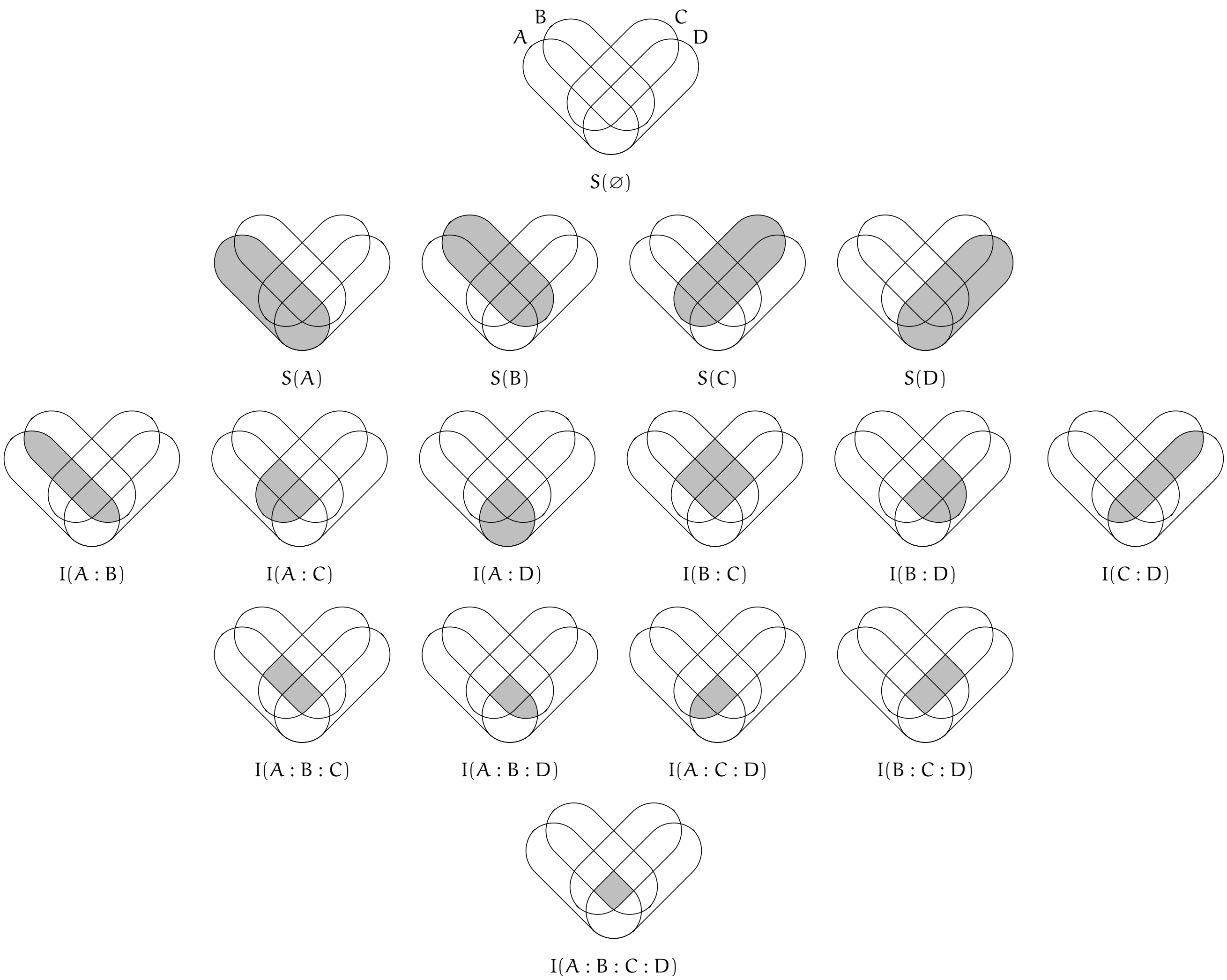


Figure 4: Information diagrams for four-variable mutual-information.

Burg entropy [28]:

$$\mathrm{Burg}(B) := \sum_x \ln P(B_x) \tag{17}$$

Proportional to the cross entropy with a uniform source distribution.

Relative joint entropy

$$D\big(A,B \parallel A',B'\big) := \sum_{x,y} P(A_x,B_y) \ln \frac{P(A_x,B_y)}{P(A'_x,B'_y)} \tag{18}$$

We can generalize any Shannon information measure to a *relative Shannon information measure* by combining appropriate linear combinations of relative joint entropies[6].

Monotonicity of relative entropy: The relative joint entropy is always greater or equal to the marginal relative entropy.

$$D\big(A,B \parallel A',B'\big) \geqslant D\big(B \parallel B'\big)$$

This follows because the relative conditional entropy $D(A|B \parallel A'|B') = D(A,B \parallel A',B') - D(B \parallel B')$ is non-negative.

Relative conditional entropy (conditional relative entropy, conditional Kullback-Leibler divergence) [8]

$$D\big(A \mid B \parallel A' \mid B'\big) := \sum_{x,y} P(A_x,B_y) \ln \frac{P(A_x \mid B_y)}{P(A'_x \mid B'_y)} \tag{19}$$

The relative conditional entropy is non-negative (Proof via log-sum inequality).

The *chain rule for relative entropy* is

$$D(A,B \parallel A',B') = D\big(A \mid B \parallel A' \mid B'\big) + D(B \parallel B')$$

Relative mutual information [1]

$$D(A:B \parallel A':B') := \sum_{x,y} P(A_x,B_y) \ln \frac{P(A_x,B_y)}{P(A'_x,B'_y)} \frac{P(A'_x)P(B'_y)}{P(A_x)P(B_y)} \tag{20}$$

[6] One reason for extending the standard notations to this larger class of relative Shannon information measures is that such measures turn up naturally, for instance when consider the non-equilibrium thermodynamics of strongly coupled systems [29, 30]

The relative joint entropy can be split up into various components analogously to the joint entropy.

$$\begin{aligned}
&D(A,B \parallel A',B') \\
&:= D(A \parallel A') + D(B \parallel B') + D(A:B \parallel A':B') \\
&= D\big(A \mid B \,\big\|\, A' \mid B'\big) + D(B \parallel B') \\
&= D\big(A \mid B \,\big\|\, A' \mid B'\big) + D\big(B \mid A \,\big\|\, B' \mid A'\big) \\
&\qquad - D(A:B \parallel A':B')
\end{aligned}$$

If the reference distributions are independent, the relative mutual information is equal to the mutual information between the principle distributions.

$$D(A:B \parallel A':B') = I(A:B) \text{ if } A \perp\!\!\!\perp B$$

Note that the uncertainly coefficient (14) is also sometimes called the relative mutual information.

Relative conditional mutual information [1]

$$\begin{aligned}
&D(A:B \mid C \parallel A':B' \mid C') \\
&:= \sum_z P(C_z) \sum_{x,y} P(A_x,B_y \mid C_z) \\
&\qquad \times \ln \frac{P(A_x,B_y \mid C_z)}{P(A'_x,B'_y \mid C'_z)} \frac{P(A'_x \mid C'_z)P(B'_y \mid C'_z)}{P(A_x \mid C_z)P(B_y \mid C_z)}
\end{aligned} \tag{21}$$

We could continue this insanity by generalizing the conditional mutual relative entropy to many variables.

Relative relative entropy [1]

$$D\big((A \parallel B)\big\|(C \parallel D)\big) := \sum_x P(A_x) \ln \frac{P(A_x)}{P(B_x)} \frac{P(D_x)}{P(C_x)} \tag{22}$$

The relative operation 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).

$$D(A:B \parallel A':B') = D\big((A,B \parallel \widehat{A},\widehat{B})\big\|(A',B' \parallel \widehat{A}',\widehat{B}')\big)$$

Jeffreys entropy The Jeffreys entropy (Jeffreys divergence, J-divergence or symmetrized Kullback-Leibler divergence) [31, 25] is a symmetrized relative entropy (15).

$$\begin{aligned}
&\mathrm{Jeffreys}(A;B) := D(A \parallel B) + D(A \parallel B) \\
&\quad = \sum_x P(A_x) \ln \frac{P(A_x)}{P(B_x)} + \sum_x P(B_x) \ln \frac{P(B_x)}{P(A_x)} \\
&\quad = \tfrac{1}{2} \sum_x \big(P(A_x) - P(B_x)\big) \ln \frac{P(A_x)}{P(B_x)}
\end{aligned} \tag{23}$$

This measure is symmetric and non-negative, but not a metric since it does not obey the triangle inequality. The Jeffreys entropy is a symmetric f-divergence (34), $C_f(A;B)$ with $f(t) = (t-1)\ln t$.

Note that Jeffreys entropy is sometimes defined to be half the value as originally defined by Jeffreys.

Jensen-Shannon divergence (Jensen-Shannon entropy, Jensen difference, information radius, capacitory discrimination[7]) is the mean relative entropy between two distributions and the distribution mean [33, 32]

$$\begin{aligned} \mathrm{JS}(A;B) :=& \tfrac{1}{2}\sum_x P(A_x)\ln\frac{P(A_x)}{\frac{1}{2}\big(P(A_x)+P(B_x)\big)} \\ &+ \tfrac{1}{2}\sum_x P(B_x)\ln\frac{P(B_x)}{\frac{1}{2}\big(P(A_x)+P(B_x)\big)} \\ =& \tfrac{1}{2}D(A \,\|\, M) + \tfrac{1}{2}D(B \,\|\, M)\ , \\ =& S(M) - \tfrac{1}{2}S(A) - \tfrac{1}{2}S(B)\ . \end{aligned} \tag{24}$$

where

$$P(M_x) = \tfrac{1}{2}P(A_x) + \tfrac{1}{2}P(B_x)$$

One interpretation of the Jensen-Shannon entropy is in terms of a Bayesian inference problem [34]: given a sample taken from one of two probability distributions, the Jensen-Shannon entropy is the average information the sample provides about the identity of the distribution. The divergence is equal to zero only if the two distributions are identical, and therefore indistinguishable, and reaches its maximum value of $\ln 2$ nats (i.e. 1 bit) if the two distributions do not overlap and therefore are perfectly distinguishable from a single sample.

The Jeffreys and Jensen-Shannon entropies are related by the inequalities [33]

$$0 \leqslant \mathrm{JS}(A;B) \leqslant \mathrm{Jeffreys}(A;B)\ .$$

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

$$\begin{aligned} \mathrm{JS}_\alpha(A;B) :=& (1-\alpha)D(A \,\|\, M) + \alpha D(B \,\|\, M)\ , \\ P(M) =& (1-\alpha)P(A) + \alpha P(B)\ . \end{aligned} \tag{25}$$

$$\begin{aligned} \mathrm{JS}_0(A;B) &= D(A \,\|\, B) \\ \mathrm{JS}_{\frac{1}{2}}(A;B) &= \mathrm{JS}(A;B) \\ \mathrm{JS}_1(A;B) &= D(B \,\|\, A) \end{aligned}$$

[7] Capacitory discrimination is defined as twice the Jensen-Shannon divergence by Topsøe [32]

Jensen-Shannon entropy (generalized Jensen-Shannon divergence) [35, 34]

$$\begin{aligned} \mathrm{JS}_\Theta(A^1; A^2; \cdots; A^n) :=& \sum_{\alpha=1,n} P(\Theta_\alpha)\, D\big(A^\alpha \,\|\, M\big)\ , \\ P(M_x) =& \sum_\alpha P(\Theta_\alpha)\, P(A^\alpha_x)\ . \end{aligned} \tag{26}$$

The entropy of a mixed distribution is the average entropy of the components plus the Jensen-Shannon entropy [32] :

$$S(M) = \mathrm{JS}_\Theta(A^1; A^2; \cdots; A^n) + \sum_\alpha P(\Theta_\alpha)\, S(A^\alpha)$$

The multivariate Jensen-Shannon entropy is the mutual information between the mixing ensemble Θ, and the mixed ensemble M.

$$\begin{aligned} I(\Theta : M) :=& \sum_{\alpha,x} P(\Theta_\alpha, M_x)\ln\frac{P(\Theta_\alpha, M_x)}{P(\Theta_\alpha)P(M_x)} \\ =& -\sum_x P(M_x)\ln P(M_x) \\ &+ \sum_{\alpha,x} P(\Theta_\alpha)P(M_x \mid \Theta_\alpha)\ln P(M_X \mid \Theta_\alpha) \\ =& S(M) - \sum_\alpha P(\Theta_\alpha)\, S(A^\alpha) \\ =& \mathrm{JS}_\Theta(A^1; A^2; \cdots; A^n) \end{aligned}$$

Resistor-average entropy [36]:

$$\mathrm{ResAvg}(A;B) := \frac{1}{\frac{1}{D(A \,\|\, B)} + \frac{1}{D(B \,\|\, A)}} \tag{27}$$

The harmonic mean of forward and reversed relative entropies.

4 Rényi information

Rényi information (Rényi entropy, alpha-order entropy) [37]: A one parameter generalization of the Shannon entropy.

$$\mathrm{Renyi}_\alpha(A) := \frac{1}{1-\alpha}\ln\sum_x P(A_x)^\alpha \tag{28}$$

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 (46).

$$\mathrm{Renyi}_\alpha(A) \geqslant \mathrm{Renyi}_\beta(A)\ , \quad \beta \geqslant \alpha$$

Collision entropy (Rényi quadratic entropy, Rényi information of order 2, second order entropy) [38]

$$\begin{aligned} \mathrm{CollisionEntropy}(A) &:= -\ln \sum_x P(A_x)^2 \\ &= \mathrm{Renyi}_2(A) \end{aligned} \tag{29}$$

A special case of the Rényi information. The negative log (61) probability that two independent samples from the distribution are the same.

Min-entropy [37]

$$\mathrm{MinEntropy}(A) := -\ln \max_x P(A_x) \tag{30}$$

Hartley entropy (Hartley function, max-entropy, Boltzmann entropy) [39]: The logarithm of the number distinct possibilities.

$$\mathrm{Hartley}(A) := \ln |\Omega_A| \tag{31}$$

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

Tsallis information [40, 41, 42] (Havrda-Charvát information, α order information)

$$\begin{aligned} \mathrm{Tsallis}_\alpha(A) &:= \frac{1}{\alpha-1}\big(1-\sum_x P(A_x)^\alpha\big) \\ &= \frac{1}{\alpha-1}\left[e^{(\alpha-1)\,\mathrm{Renyi}_\alpha(A)}-1\right] \end{aligned} \tag{32}$$

Sharma-Mittal information [43, 44]

$$\mathrm{SharmaMittal}_{\alpha,\beta}(A) := \frac{1}{\beta-1}\left[1-\left(\sum_x P(A_x)^\alpha\right)^{\frac{1-\beta}{1-\alpha}}\right] \tag{33}$$

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

$$\begin{aligned} \mathrm{SharmaMittal}_{1,1}(A) &= S(A) \\ \mathrm{SharmaMittal}_{\alpha,1}(A) &= \mathrm{Renyi}_\alpha(A) \\ \mathrm{SharmaMittal}_{\alpha,\alpha}(A) &= \mathrm{Tsallis}_\alpha(A) \end{aligned}$$

5 Csiszár f-divergences

Csiszár f-divergence Many interesting divergence measures between probability distributions can be written as (or related to) an f-divergence (also know as Csiszár[8], Csiszár-Morimoto, Ali-Silvey, or ϕ-divergence) [45, 46, 47, 48, 49, 36].

$$C_f(A;B) = \sum_x P(A_x)\, f\Big(\frac{P(B_x)}{P(A_x)}\Big)\,, \tag{34}$$

where the function f is convex $\smile$ and $f(1) = 0$. This implies $C_f(A;B) \geqslant 0$ 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 combinations: 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 $c_n \geqslant 0$ is non-negative. It follows that a positive linear sum of an f-divergences is also an f-divergence [50].

Dual f-divergence [50] The dual of an f divergence is defined by swapping the arguments.

$$C_f(A;B) = C^*_f(B;A) = C_{f^*}(B;A) \tag{35}$$

Here f^* is the *Csiszár dual* [50] 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

$$\begin{aligned} C_f(A;B) &= \tfrac{1}{2}C_g(A;B) + \tfrac{1}{2}C_g(B;A) \\ f(t) &= \tfrac{1}{2}g(t) + \tfrac{1}{2}t g(t^{-1}) \end{aligned}$$

and the *Jensen symmetrization* where we take the average divergence of both distributions to the average distribution.

$$\begin{aligned} C_f(A;B) &= \tfrac{1}{2}C_h(A;M) + \tfrac{1}{2}C_h(B;M) \\ P(M_x) &= \tfrac{1}{2}P(A_x) + \tfrac{1}{2}P(B_x) \\ f(t) &= \tfrac{1}{2}h\left(\frac{1}{1+\frac{1}{2}t}\right) + \tfrac{1}{2}h\left(\frac{1}{1+\frac{1}{2}t^{-1}}\right) \end{aligned}$$

[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)

	Asymmetric f-divergences	$f(t)$	$f^*(t)$
(15)	relative entropy	$-\ln t$	$t \ln t$
(36)	K-divergence	$\ln \dfrac{2}{(1+t)}$	$t \ln \dfrac{2t}{(1+t)}$
(39)	Pearson divergence	$(t-1)^2$	$(\frac{1}{\sqrt{t}} - \sqrt{t})^2$
(48)	Cressie-Read divergence	$\dfrac{t^{-\alpha}-1}{\alpha(\alpha+1)}$	
(49)	Tsallis divergence	$\dfrac{t^{1-\alpha}-1}{\alpha-1}$	

			Jeffreys	Jensen
	Symmetric f-divergences	$f(t)$	$h(t)$	$g(t)$
(41)	LeCam discrimination	$\dfrac{(t-1)^2}{(t+1)}$		$(t-1)^2$
(23)	Jeffreys entropy	$(t-1)\ln t$	$-\ln t$	
(24)	Jensen-Shannon entropy	$\frac{1}{2}\ln\dfrac{2}{1+t} + \frac{1}{2} t \ln \dfrac{2t}{1+t}$	$\ln \frac{2}{1+t}$	$-\ln t$
(53)	variational distance	$\frac{1}{2}\lvert t-1\rvert$	$\cdots$	
(38)	Hellinger discrimination	$1-\sqrt{t}$	$\cdots$	

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], becuase a Jeffreys symmetrization of the relative entropy (15) is half the Jeffreys entropy (23), and a Jensen symmetrization gives the Jensen-Shannon entropy (24).

K-divergence [33]

$$\begin{aligned}\mathrm{KayDiv}(A;B) &:= \sum_x P(A_x) \ln \frac{P(A_x)}{\tfrac{1}{2}\big(P(A_x)+P(B_x)\big)} \qquad (36)\\ &= D(A \parallel M)\ ,\ P(M_x) = \tfrac{1}{2}P(A_x) + \tfrac{1}{2}P(B_x)\\ &= C_f(A;B), \quad f(t) = \ln \frac{2}{(1+t)}\ .\end{aligned}$$

Of interest since the K-divergence is a lower bound to the relative entropy [33],

$$\mathrm{KayDiv}(A;B) \leqslant \tfrac{1}{2} D(A \parallel B)$$

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

$$\begin{aligned}\mathrm{JS}(A;B) &= \tfrac{1}{2}\,\mathrm{KayDiv}(A;B) + \tfrac{1}{2}\,\mathrm{KayDiv}(B;A)\\ &= \tfrac{1}{2}D(A \parallel M) + \tfrac{1}{2}D(B \parallel M)\end{aligned}$$

Fidelity (Bhattacharyya coefficient, Hellinger affinity) [51] The Bhattacharyya distance (59) and the Hellinger divergence and distance (38) are functions of fidelity. The name derives from usage in quantum information theory [52].

$$\mathrm{Fidelity}(A;B) := \sum_x \sqrt{P(A_x)P(B_x)} \qquad (37)$$

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 (38) and Bhattacharyya distance (59).

Hellinger discrimination (Squared Hellinger distance, infidelity) [53]

$$\begin{aligned}\mathrm{HellingerDiv}(A;B) &:= \tfrac{1}{2}\sum_x \Big(\sqrt{P(A_x)} - \sqrt{P(B_x)}\,\Big)^2 \qquad (38)\\ &= \sum_x P(A_x)\Big(1 - \sqrt{\frac{P(B_x)}{P(A_x)}}\,\Big)\\ &= C_f(A;B), \quad f(t) = (1-\sqrt{t})\\ &= 1 - \sum_x \sqrt{P(A_x)P(B_x)}\\ &= 1 - \mathrm{Fidelity}(A;B)\end{aligned}$$

Symmetric, with range $[0,1]$. A common alternative normalization omits the one-half prefactor. The name originates form that of the corresponding integral in the continuous case [54].

Pearson divergence (χ^2-divergence, chi square divergence, Pearson chi square divergence, Kagan divergence, quadratic divergence, least squares) [55]

$$\begin{aligned}\mathrm{Pearson}(A;B) &:= \sum_x \frac{\big(P(B_x)-P(A_x)\big)^2}{P(A_x)} \qquad (39)\\ &= \sum_x P(A_x)\Big(\frac{P(A_x)}{P(B_x)} - 1\Big)^2\\ &= C_f(A;B)\ , \quad f(t) = (t-1)^2\end{aligned}$$

Neyman divergence (inverse Pearson chi square divergence) [56, 57]

$$\mathrm{Neyman}(A;B) := \mathrm{Pearson}(B;A) \qquad (40)$$

The dual of the Pearson divergence (arguments switched).

LeCam discrimination (LeCam discrimination, Vincze-LeCam divergence, triangular discrimination) [58, 59, 32]

$$\begin{aligned}\mathrm{LeCam}(A;B) &:= \sum_x \frac{\big(P(A_x)-P(B_x)\big)^2}{P(A_x)+P(B_x)} \qquad (41)\\ &= 2C_f(A;B), \quad f(t) = \frac{(t-1)^2}{2(t+1)}\ .\end{aligned}$$

The LeCam discrimination is a Jensen symmetrized Pearson divergence.

$$\begin{aligned}\mathrm{LeCam}(A;B) &:= \tfrac{1}{2}\,\mathrm{Pearson}\,(A;M) + \tfrac{1}{2}\,\mathrm{Pearson}\,(B;M)\\ P(M_x) &= \tfrac{1}{2}P(A_x) + \tfrac{1}{2}P(B_x)\end{aligned}$$

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

Skewed K-divergence [33]

$$\begin{aligned}\mathrm{KayDiv}_\alpha(A;B) &:= \sum_x P(A_x) \ln \frac{P(A_x)}{(1-\alpha)P(A_x) + \alpha P(B_x)} \qquad (42)\\ &= D(A \parallel M)\ ,\\ &\quad P(M) = (1-\alpha)P(A) + \alpha P(B)\end{aligned}$$

Alpha-Jensen-Shannon-entropy [33, 60]

$$\mathrm{AlphaJS}_\alpha(A;B) := \tfrac{1}{2}\,\mathrm{KayDiv}_\alpha(A;B) + \tfrac{1}{2}\,\mathrm{KayDiv}_\alpha(B;A) \tag{43}$$

$$\begin{aligned}
\mathrm{AlphaJS}_0(A;B) &= 0\\
\mathrm{AlphaJS}_{\frac{1}{2}}(A;B) &= \mathrm{JS}(A;B)\\
\mathrm{AlphaJS}_1(A;B) &= \mathrm{Jeffreys}(A;B)
\end{aligned}$$

The Jeffreys symmetrization of the skewed K-divergence.

6 Chernoff divergence

Chernoff divergence The Chernoff divergence [61, 62] of order α is defined as

$$\begin{aligned}
\mathrm{Chernoff}_\alpha(A;B) &:= -\ln \sum_x P(A_x)\left(\frac{P(A_x)}{P(B_x)}\right)^{\alpha-1} \qquad (44)\\
&= -\ln\left[C_f(A;B)+1\right], \quad f(t) = t^{1-\alpha}-1\,.
\end{aligned}$$

The Chernoff divergence is zero for $\alpha = 1$ and $\alpha = 0$, and reaches a maximum, the **Chernoff information** [61, 8], 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 α if the distributions have the same support.

The Chernoff divergence of order α is related to the Chernoff divergence of order $1-\alpha$ with the distributions interchanged [62],

$$\mathrm{Chernoff}_\alpha(A;B) = \mathrm{Chernoff}_{1-\alpha}(B;A)\,.$$

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

Chernoff coefficient (alpha divergence) [61, 62, 63]

$$\begin{aligned}
\mathrm{ChernoffCoefficient}_\alpha(A;B) &:= \sum_x P(A_x)\left(\frac{P(A_x)}{P(B_x)}\right)^{\alpha-1} \qquad (45)\\
&= \exp\big(-\mathrm{Chernoff}_\alpha(A;B)\big)\\
&= C_f(A;B)+1, \quad f(t) = t^{1-\alpha}-1
\end{aligned}$$

The exponential twist density [64] is way of mixing two distributions [65] to form a third

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

Here α is a mixing parameter between 0 and 1. The normalization constant Z_α is the Chernoff coefficient, $Z_\alpha = \mathrm{ChernoffCoefficient}_\alpha(A;B)$ [66].

Rényi divergence The Rényi divergence (or relative Rényi entropy) of order α is a one-parameter generalization of the relative entropy [37],

$$\begin{aligned}
\mathrm{Renyi}_\alpha(A;B) &:= \frac{1}{\alpha-1}\ln\sum_x P(A_x)\left(\frac{P(A_x)}{P(B_x)}\right)^{\alpha-1} \qquad (46)\\
&= \frac{1}{1-\alpha}\,\mathrm{Chernoff}_\alpha(A;B)\\
&= \frac{1}{\alpha-1}\ln\left[C_f(A;B)+1\right], \quad f(t) = t^{1-\alpha}-1\,.
\end{aligned}$$

Higher values of α give a Rényi divergence dominated by the greatest ratio between the two distributions, whereas as α 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 ∞. As previously mentioned, $\alpha = 1$ gives the relative entropy (15), and $\alpha = \frac{1}{2}$ gives the Bhattacharyya distance (59). In the limit $\alpha \to 0$, the Rényi divergence slides to the negative log probability under q that p is non-zero,

$$\begin{aligned}
\lim_{\alpha\to 0}\mathrm{Renyi}_\alpha(A;B) &= -\ln\sum_x \lim_{\alpha\to 0} P(A_x)^\alpha P(B_x)^{1-\alpha}\\
&= -\ln\sum_x P(B_x)[P(A_x)>0]\,.
\end{aligned}$$

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,

$$D_\alpha(A;B) := \frac{1}{\alpha(1-\alpha)}\left(1 - \sum_x P(A_x)^\alpha\, P(B_x)^{1-\alpha}\right) \tag{47}$$

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 entropy.

$$D_\alpha(A;B) = \begin{cases}
\mathrm{Pearson}(A;B) & \alpha = -1\\
D(B \parallel A) & \lim_{\alpha\to 0}\\
4\,\mathrm{HellingerDiv}(A;B) & \alpha = +\frac{1}{2}\\
D(A \parallel B) & \lim_{\alpha\to +1}\\
\mathrm{Neyman}(A;B) & \alpha = +2
\end{cases}$$

Another common paramaterization is [0, 0]

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

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

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

Cressie-Read divergence [57]

$$\begin{aligned}\mathrm{CressieRead}_\alpha(A;B) &:= \frac{1}{\alpha(\alpha+1)} \sum_x P(A_x)\left[\left(\frac{P(A_x)}{P(B_x)}\right)^{\alpha} - 1\right] \qquad (48)\\ &= \frac{1}{\alpha(\alpha+1)}\left[e^{\alpha\,\mathrm{Renyi}_{\alpha+1}(A;B)} - 1\right]\\ &= \frac{1}{\alpha+1}\mathrm{Tsallis}_{\alpha+1}(A;B)\\ &= C_f(A;B), \quad f(t) = \frac{t^{-\alpha}-1}{\alpha(\alpha+1)}\ .\end{aligned}$$

Tsallis divergence (relative Tsallis entropy) [67] Other closely related divergences include the relative Tsallis entropy,

$$\begin{aligned}\mathrm{Tsallis}_\alpha(A;B) &:= \frac{1}{\alpha-1}\sum_x P(A_x)\left[\left(\frac{P(A_x)}{P(B_x)}\right)^{\alpha-1} - 1\right] \qquad (49)\\ &= \frac{1}{\alpha-1}\left[e^{(\alpha-1)\,\mathrm{Renyi}_\alpha(A;B)} - 1\right]\\ &= C_f(A;B), \quad f(t) = \tfrac{t^{1-\alpha}-1}{\alpha-1}\ ,\end{aligned}$$

Sharma-Mittal divergence [43, 44]

$$\begin{aligned}&\mathrm{SharmaMittal}_{\alpha,\beta}(A;B) \qquad (50)\\ &:= \frac{1}{\beta-1}\left[1 - \left(\sum_x P(A_x)^\alpha P(B_x)^{1-\alpha}\right)^{\frac{1-\beta}{1-\alpha}}\right]\\ &= \frac{1}{1-\beta}\left(1 - \mathrm{ChernoffCoefficient}_\alpha(A;B)^{\frac{1-\beta}{1-\alpha}}\right)\\ &\alpha > 0,\ \alpha \neq 0,\ \beta \neq 0\end{aligned}$$

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 [68, 69, 44]

$$\begin{aligned}&\mathrm{CauchySchwarz}(A;B) := \qquad (51)\\ &\qquad -\ln \frac{\sum_x P(A_x)P(B_x)}{\sqrt{\left(\sum_x P(A_x)^2\right)\left(\sum_x P(B_x)^2\right)}}\end{aligned}$$

This divergence is symmetric, and takes advantage of the Cauchy-Schwarz inequality to ensure that the divergence is non-negative, being zero only for identical distributions.

Cauchy-Schwarz angle [1]

$$\begin{aligned}&\mathrm{CauchySchwarzAngle}(A;B) := \qquad (52)\\ &\qquad \arccos \frac{\sum_x P(A_x)P(B_x)}{\sqrt{\left(\sum_x P(A_x)^2\right)\left(\sum_x P(B_x)^2\right)}}\end{aligned}$$

By similar reasoning we can define an angle between distributions. This is a metric.

8 Distances

Variational distance (L_1 distance, variational divergence, Kolmogorov distance) [0, 0]

$$\begin{aligned}V(A;B) = L_1(A;B) &:= \tfrac{1}{2}\sum_x \left|P(B_x) - P(A_x)\right| \qquad (53)\\ &= C_f(A;B),\ f(t) = |t-1|\end{aligned}$$

The only f-divergence (34) which is also a metric [70].

Pinsker's inequality:

$$D(A \parallel B) \geqslant \tfrac{1}{2} V(A;B)^2$$

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 (L_2 distance)

$$L_2(A;B) := \sqrt{\sum_x \left|P(B_x) - P(A_x)\right|^2} \qquad (54)$$

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

$$L_p(A;B) := \left(\sum_x \left| P(B_x) - P(A_x) \right|^p \right)^{\frac{1}{p}} \tag{55}$$

A metric distance provided that $p \geqslant 1$.

Chebyshev distance

$$L_\infty(A;B) := \max \left| P(B_x) - P(A_x) \right| \tag{56}$$

LeCam distance The square root of the LeCam discrimination (41) [71].

$$\text{LeCamDist}(A;B) := \sqrt{\tfrac{1}{2}\,\text{LeCam}(A;B)} \tag{57}$$

Hellinger distance The square root of the Hellinger divergence.

$$\text{Hellinger}(A;B) := \sqrt{\text{HellingerDiv}(A;B)} \tag{58}$$

Jensen-Shannon distance is the square root of the Jensen-Shannon divergence, and is a metric between probability distributions [34, 72].

Bhattacharyya distance [73] The Chernoff divergence of order $\alpha = \frac{1}{2}$. The negative logarithm of the Bhattacharyya coefficient (fidelity).

$$\begin{aligned} \text{Bhattacharyya}(A;B) &:= -\ln \sum_x \sqrt{P(A_x)P(B_x)} \qquad (59) \\ &= \text{Chernoff}_{\frac{1}{2}}(A;B) \\ &= -\ln \text{Fidelity}(A;B)\ . \end{aligned}$$

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

$$\begin{aligned} \text{Bhattacharyya}(A;B) &:= \arccos \sum_x \sqrt{P(A_x)P(B_x)} \qquad (60) \\ &= \arccos \text{Fidelity}(A;B)\ . \end{aligned}$$

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 [13].

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. $I(A : B_x)$ [26].

Specific entropy (information content, self-information, score, surprise, surprisal) [74, 75] is the negative logarithm of a probability.

$$s(A_x) := -\ln P(A_x) \tag{61}$$

The expectation of the specific entropy is the Shannon entropy, the average point-wise entropy of the ensemble (1).

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

Local conditional entropy [0]

$$\begin{aligned} S(A \mid B_b) &:= -\sum_a P(A_a \mid B_b) \ln P(A_a \mid B_b) \qquad (62) \\ &= \mathbb{E}_A\left[-\ln P(A \mid B)\right] \qquad (63) \end{aligned}$$

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) [26, 13, 76]

$$i(A_x : B_y) := \ln \frac{P(A_x, B_y)}{P(A_x)P(B_y)} \tag{64}$$

Similarly, the expectation of the specific mutual information is the mutual information of the ensemble (5).

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

Back Matter

Acknowledgments

Useful sources and pedagogical expositions include Cover and Thomas's classic *Elements of Information Theory* [8], the inestimable (for some value of truthiness) Wikipedia, and various other reviews [77, 3]. Notation and nomenclature was adapted from Good (1956) [26]. The title of this essay was shamelessly lifted from Rényi's paper on a similar subject [37].

This review is inevitably incomplete, inaccurate and otherwise imperfect — *caveat emptor*.

Version history

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 (47) and Burg entropy (17). Corrected typos. (Kudos: Glenn Davis)
0.5 (2016-08-16) Extensive miscellaneous improvements.
0.4 (2016-06-22) Added Jensen-Shannon divergence; alpha Jensen-Shannon entropy; skewed K-Divergence; dual divergence; and total variational distance. Reordered operator precedence. Expanded and refined collection of relative Shannon entropies. Extensive miscellaneous improvements.
0.3 (2016-04-21) Added residual entropy (11) (and combined with independent information, which is essentially another name for the same thing), and uncertainty coefficient (14). Added information diagrams. Miscellaneous minor improvements. Fixed sign error in lautum information (13) (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 2, 12, 14, 15, 15, 16, 16, 16, 16, 16, 17, and 17).

[1] Gavin E. Crooks. On measures of entropy and information. Technical report, (2017). Tech. Note 009 v0.7 http://threeplusone.com/info. (pages 2, 7, 10, 10, 10, 14, 16, and 18).

[2] Raymond W. Yeung. A new outlook on Shannon's information measures. *IEEE Trans. Inf. Theory*, 37(3):466–474 (1991). doi:10.1109/18.79902. (pages 2, 4, and 6).

[3] Ryan G. James, Christopher J. Ellison, and James P. Crutchfield. Anatomy of a bit: Information in a time series observation. *Chaos*, 21(3):037109 (2011). doi:10.1063/1.3637494. (pages 2, 4, 5, 6, 6, 6, and 17).

[4] Ludwig Boltzmann. Weitere studien über das Wärmegleichgewicht unter Gasmolekülen (Further studies on the thermal equilibrium of gas molecules). *Sitzungsberichte Akad. Wiss., Vienna, part II*, 66:275–370 (1872). (page 3).

[5] Claude E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423, 623–656 (1948). doi:10.1002/j.1538-7305.1948.tb01338.x. (pages 3, 3, 3, 4, and 4).

[6] Rudolf Clausius. Ueber verschiedene für die anwendung bequeme formen der hauptgleichungen der mechanischen wärmetheorie. *Annalen der Physik und Chemie*, 201(7):353–400 (1865). doi:10.1002/andp.18652010702. (page 3).

[7] J. Willard Gibbs. *Elementary principles in statistical mechanics*. Charles Scribner's Sons, New York (1902). (page 3).

[8] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley, New York, 2nd edition (2006). (pages 3, 3, 4, 7, 7, 7, 10, 15, and 17).

[9] Frederick Jelinek. Continuous speech recognition by statistical methods. *Proc. IEEE*, 64(4):532–556 (1976). doi:10.1109/PROC.1976.10159. (page 3).

[10] O. Aftab, P. Cheung, A. Kim, S. Thakknar, and N. Yeddanapudi. Information theory and the digital revolution (2001). http://web.mit.edu/6.933/www/Fall2001/Shannon2.pdf. (page 4).

[11] William J. McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116 (1954). doi:10.1007/BF02289159. (pages 4, 5, 5, 5, and 7).

[12] Satosi Watanabe. Information theoretical analysis of multivariate correlation. *IBM J. Res. Develop.*, 4(1):66–82 (1960). doi:10.1147/rd.41.0066. (pages 5, 5, and 7).

[13] Robert Fano. *Transmission of Information: A Statistical Theory of Communications*. MIT Press, Cambridge, MA (1961). (pages 5, 5, 17, and 17).

[14] Anthony J. Bell. The co-information lattice. In *Proc. 4th Int. Symp. Independent Component Analysis and Blind Source Separation* (2003). (pages 5 and 7).

[15] Yaneer Bar-Yam. Multiscale complexity / entropy. *Adv. in Complex Sys.*, 7:47–63 (2004). doi:10.1142/S0219525904000068. (page 5).

[16] Elad Schneidman, Susanne Still, Michael J. Berry, II, and William Bialek. Network information and connected correlations. *Phys. Rev. Lett.*, 91:238701 (2003). doi:10.1103/PhysRevLett.91.238701. (page 5).

[17] Te Sun Han. Nonnegative entropy measures of multivariate symmetric correlations. *Inf. Control*, 36:133–156 (1978). doi:10.1016/S0019-9958(78)90275-9. (page 6).

[18] Samer A. Abdallah and Mark D. Plumbley. A measure of statistical complexity based on predictive information with application to finite spin systems. *Phys. Lett. A*, 376(4):275–281 (2012). doi:10.1016/j.physleta.2011.10.066. (page 6).

[19] Phipps Arabie and Scott A. Boorman. Multidimensional scaling of measures of distance between partitions. *J. Math. Psy.*, 10(148-203) (1973). doi:10.1016/0022-2496(73)90012-6. (page 6).

[20] James P. Crutchfield. Information and its metric. In L. Lam and H. C. Morris, editors, *Nonlinear Structures in Physical Systems – Pattern Formation, Chaos and Waves*, pages 119–130. Springer-Verlag, New York (1990). (page 6).

[21] Sergio Verdú and Tsachy Weissman. The information lost in erasures. *IEEE Trans. Inf. Theory*, 54:5030–5058 (2008). doi:10.1109/TIT.2008.929968. (page 6).

[22] Te Sun Han. Multiple mutual information and multiple interactions in frequency data. *Inf. Control*, 46:26–45 (1980). doi:10.1016/S0019-9958(80)90478-7. (page 7).

[23] Daniel P. Palomar and Sergio Verdú. Lautum information. *IEEE Trans. Inf. Theory*, 54(3):964–975 (2008). (page 7).

[24] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, New York, 3rd edition (2007). (page 7).

[25] Solomon Kullback and Richard A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86 (1951). doi:10.1214/aoms/1177729694. (pages 7 and 10).

[26] Irving J. Good. Some terminology and notation in information theory. *Proc. Inst. Elec. Eng. C*, 103(3):200–204 (1956). doi:10.1049/pi-c.1956.0024. (pages 7, 17, 17, and 17).

[27] David F. Kerridge. Inaccuracy and inference. *J. Roy. Statist. Soc. Sec. B*, 23:184–194 (1961). http://www.jstor.org/stable/2983856. (page 7).

[28] John Parker Burg. The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics*, 37:375–376 (1972). doi:10.1190/1.1440265. (page 10).

[29] Giovanni Diana and Massimiliano Esposito. Mutual entropy production in bipartite systems. *J. Stat. Mech.: Theor. Exp.*, (4):P04010 (2014). doi:10.1088/1742-5468/2014/04/P04010. (page 10).

[30] Gavin E. Crooks and Susanne Still. Marginal and conditional second laws of thermodynamics. *Europhys. Lett.*, 125(4):40005 (2019). doi:10.1209/0295-5075/125/40005. arXiv:1611.04628. (page 10).

[31] Harold Jeffreys. *Theory of probability*. Clarendon Press, Oxford, 2nd edition (1948). (page 10).

[32] Flemming Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inf. Theory*, 46(4):1602–1609 (2000). doi:10.1109/18.850703. (pages 11, 11, 11, and 14).

[33] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory*, 37:145–151 (1991). doi:10.1109/18.61115. (pages 11, 11, 14, 14, 14, and 15).

[34] Dominik M. Endres and Johannes E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49(7):1858–1860 (2003). doi:10.1109/TIT.2003.813506. (pages 11, 11, and 17).

[35] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons (1968). (page 11).

[36] Don H. Johnson and Sinan Sinanović. Symmetrizing the Kullback-Leibler distance. Technical report, Rice Univ., Houston, TX (2001). http://www.ece.rice.edu/~dhj/resistor.pdf. (pages 11 and 12).

[37] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematics, statistics and probability*, volume 1, pages 547–561 (1960). http://projecteuclid.org/euclid.bsmsp/1200512181. (pages 11, 12, 15, and 17).

[38] Charles H. Bennett, François Bessette, Gilles Brassard, Louis Salvail, and John Smolin. Experimental quantum cryptography. *J. Cryptology*, 5:3–28 (1992). doi:10.1007/BF00191318. (page 12).

[39] Ralph V. L. Hartley. Transmission of information. *Bell Syst. Tech. J.*, 7:535–563 (1928). doi:10.1002/j.1538-7305.1928.tb01236.x. (page 12).

[40] Jan Havrda and František Charvát. Quantification method of classification processes. Concept of structural α-entropy. *Kybernetika*, 3:30–35 (1967). http://dml.cz/dmlcz/125526. (page 12).

[41] Zoltán Daróczy. Generalized information functions. *Inf. Control*, 16(1):36–51 (1970). doi:10.1016/S0019-9958(70)80040-7. (page 12).

[42] Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *J. Stat. Phys.*, 52:479–487 (1988). doi:10.1007/BF01016429. (page 12).

[43] Bhu Dev Sharma and D. P. Mittal. New non-additive measures of entropy for discrete probability distributions. *J. Math. Sci. (Soc. Math. Sci., Delhi, India)*, 10(28-40) (1975). (pages 12 and 16).

[44] Frank Nielsen and Richard Nock. A closed-form expression for the Sharma-Mittal entropy of exponential families. *J. Phys. A*, 45(3):032003 (2012). doi:10.1088/1751-8113/45/3/032003. (pages 12, 16, and 16).

[45] S. M. Ali and Samuel D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. B*, 28(1):131–142 (1966). (page 12).

[46] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar.*, 2:299–318 (1967). (page 12).

[47] Tetsuzo Morimoto. Markov processes and the H-theorem. *J. Phys. Soc. Jpn*, 18:328–331 (1963). (page 12).

[48] Inder J. Taneja. Refinement inequalities among symmetric divergence measure. *Aust. J. Math. Anal. Appl.*, 2(1):8 (2005). (page 12).

[49] Alexander N. Gorban, Pavel A. Gorban, and George Judge. Entropy: The Markov ordering approach. *Entropy*, 12:1145–1193 (2010). doi:10.3390/e12051145. (page 12).

[50] Mark D. Reid and Robert C. Williamson. Information, divergence and risk for binary experiments. *J. Mach. Learn. Res.*, 12:731–817 (2011). http://jmlr.csail.mit.edu/papers/v12/reid11a.html. (page 12, 12, and 12).

[51] Anil K. Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā*, 7:401–406 (1946). (page 14).

[52] Richard Jozsa. Fidelity for mixed quantum states. *J. Mod. Opt.*, 41:2315–2323 (1994). doi:10.1080/09500349414552171. (page 14).

[53] Shizuo Kakutani. On equivalence of infinite product measures. *Ann. Math.*, 49(1):214–224 (1948). https://www.jstor.org/stable/1969123. (page 14).

[54] Ernst Hellinger. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angew. Math.*, 136:210–271 (1909). doi:10.1515/crll.1909.136.210. (page 14).

[55] Karl Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag. Ser. 5*, 50(302):157–175 (1900). doi:10.1080/14786440009463897. (page 14).

[56] Jerzy Neyman. Contribution to the theory of the χ^2 test. In *Proceedings of the [first] Berkeley symposium on mathematics, statistics and probability*, pages 239–273. Univ. of Calif. Press (1949). `http://projecteuclid.org/euclid.bsmsp/1166219208`. (page 14).

[57] Noel Cressie and Timothy R. C. Read. Multinomial goodness of fit. *J. Roy. Statist. Soc. B*, 46:440–464 (1984). `http://www.jstor.org/stable/2345686`. (pages 14 and 16).

[58] Istán Vincze. On the concept and measure of information contained in an observation. In J. Gani and V. F. Rohatgi, editors, *Contributions to Probability: A Collection of Papers Dedicated to Eugene Lukacs*, pages 207–214. Academic Press, New York (1981). (page 14).

[59] Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, Berlin (1986). (page 14).

[60] Frank Nielsen. A family of statistical symmetric divergences based on Jensen's inequality. arXiv:1009.4004. (page 15).

[61] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23(4):493–507 (1952). doi:10.1214/aoms/1177729330. (page 15, 15, and 15).

[62] Yoshihide Kakizawa, Robert H. Shumway, and Masanobu Taniguchi. Discrimination and clustering for multivariate time series. *J. Amer. Statist. Assoc.*, 93(441):328–340 (1998). (page 15, 15, and 15).

[63] Shun'ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Society (2000). (page 15).

[64] James A. Bucklew. *Large deviation techniques in decision, simulation, and estimation,*. Wiley (1990). (page 15).

[65] Anand G. Dabak and Don H. Johnson. Relations between Kullback-Leibler distance and Fisher information (2001). `http://www.ece.rice.edu/~dhj/distance.pdf`. (page 15).

[66] John C. Baez. Rényi entropy and free energy. arXiv:1102.2098. (page 15).

[67] Lisa Borland, Angel R. Plastino, and Constantino Tsallis. Information gain within nonextensive thermostatistics. *J. Math. Phys.*, 39(12):6490–6501 (1998). doi:10.1063/1.532660. Erratum: 40 2196 (1999). (page 16).

[68] Robert Jenssen, Jose C. Principe, Deniz Erdogmus, and Torbjørn Eltoft. The cauchy–schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. *J. Franklin Institute*, 343(6):614–629 (2006). doi:10.1016/j.jfranklin.2006.03.018. (page 16).

[69] Kittipat Kampa, Erion Hasanbelliu, and Jose C. Principe. Closed-form cauchy-schwarz pdf divergence for mixture of gaussians. In *The 2011 International Joint Conference on Neural Networks*, pages 2578–2585 (2011). doi:10.1109/IJCNN.2011.6033555. (page 16).

[70] Mohammadali Khosravifard, Dariush Fooladivanda, and T. Aaron Gulliver. Confliction of the convexity and metric properties in f-divergences. *IEICE-Tran. Fund. Elec. Comm. Comp. Sci.*, E90-A:1848–1853 (2007). doi:10.1093/ietfec/e90-a.9.1848. (page 16).

[71] P. Kafka, Ferdinand Österreicher, and I. Vincze. On powers of f-divergences defining a distance. *Studia Sci. Math. Hungar.*, 26:415–422 (1991). (page 17).

[72] Ferdinand Österreicher and Igor Vajda. A new class of metric divergences on probability spaces and and its statistical applications. *Ann. Inst. Statist. Math.*, 55(3):639–653 (2003). doi:10.1007/BF02517812. (page 17).

[73] Anil K. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35(1):99–109 (1943). (page 17).

[74] E. W. Samson. Fundamental natural concepts of information theory. *ETC: A Review of General Semantics*, 10:283 (1953). (page 17).

[75] Myron Tribus. *Thermostatics and thermodynamics*. D. Van Nostrand Company (1961). (page 17).

[76] Kenneth W. Church and Patrick Hanks. Word association norms, mutual information and lexicography. *Comp. Ling.*, 16(1):22–29 (1990). (page 17).

[77] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *Int. J. Math. Mod. Meth. Appl. Sci.*, 1(4):300–307 (2007). `http://www.naun.org/main/NAUN/ijmmas/mmmas-49.pdf`. (page 17).

`http://threeplusone.com/info`

typeset on 2022-10-12 with XeTeX version 0.99999
fonts: Trump Mediaeval (text), Euler (math)
2 7 1 8 2 8 1 8 3