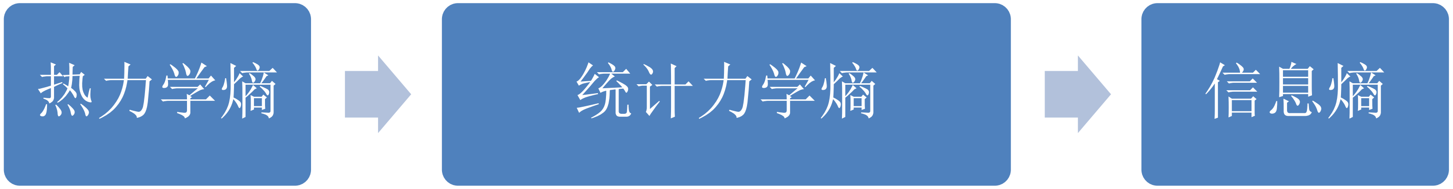
信息熵

Information entropy

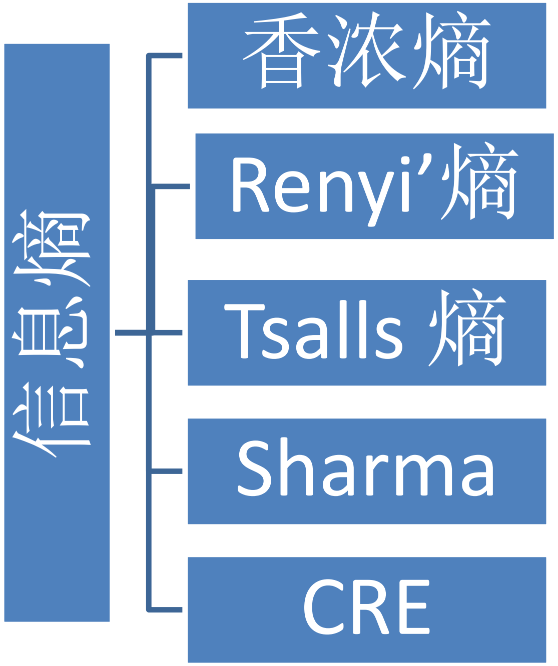
冯晨娇

2016年9月24日

熵的分类（从来源来分）



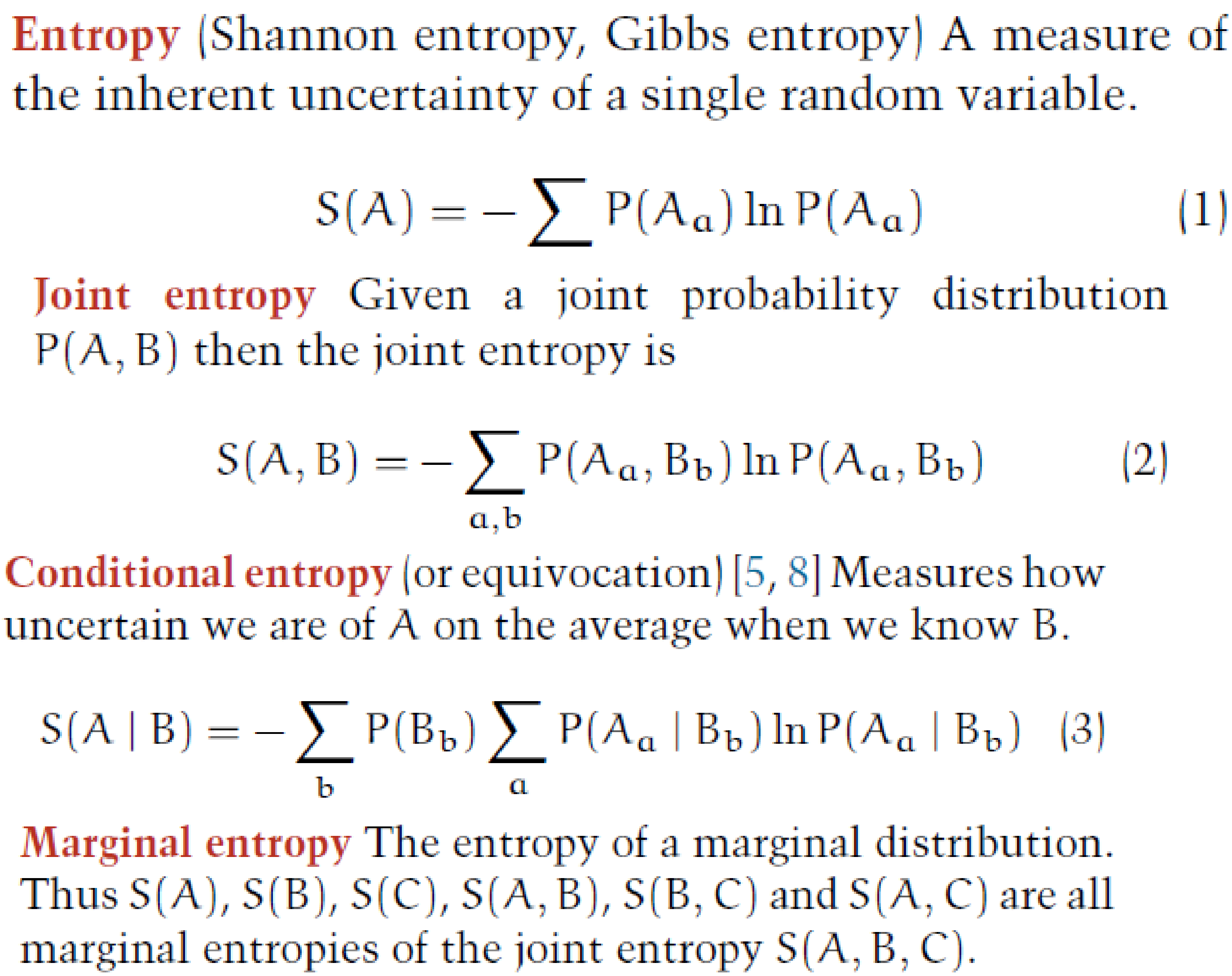
热力学熵 → 统计力学熵 → 信息熵



信息熵

- 香浓熵
- Renyi’熵
- Tsalls 熵
- Sharma
- CRE

1.Shannon entropy



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

$$S(A) = -\sum P(A_a) \ln P(A_a) \qquad (1)$$

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

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

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

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

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).

为什么信息量为$\log\frac{1}{p}$

什么是信息量

信息量是信息论中的重要概念,是指一个事件发生时所消除的不确定性的量度。例如，某初一(1)班新生第一次到学校,不知道自己的班级是哪一个教室，这个学校的24个教室都可能是初一(1)班的教室，对这个学生来说，哪个是自己班级的教室就是一个不确定性的问题。如果该生看到黑板上公布初一(1)班在二楼第1教室，就消除了这个问题的不确定性，这就是说，他得到了一定的信息量。所以，在为某个事件所消除的不确定性的量度，就是这个书件的信息量。

如何度量这个信息量

信息量是事件发生概率的单调减函数。例如，明天要下雨和明天有龙卷风，显然第二个事件信息量更大，但是它的概率是小的。

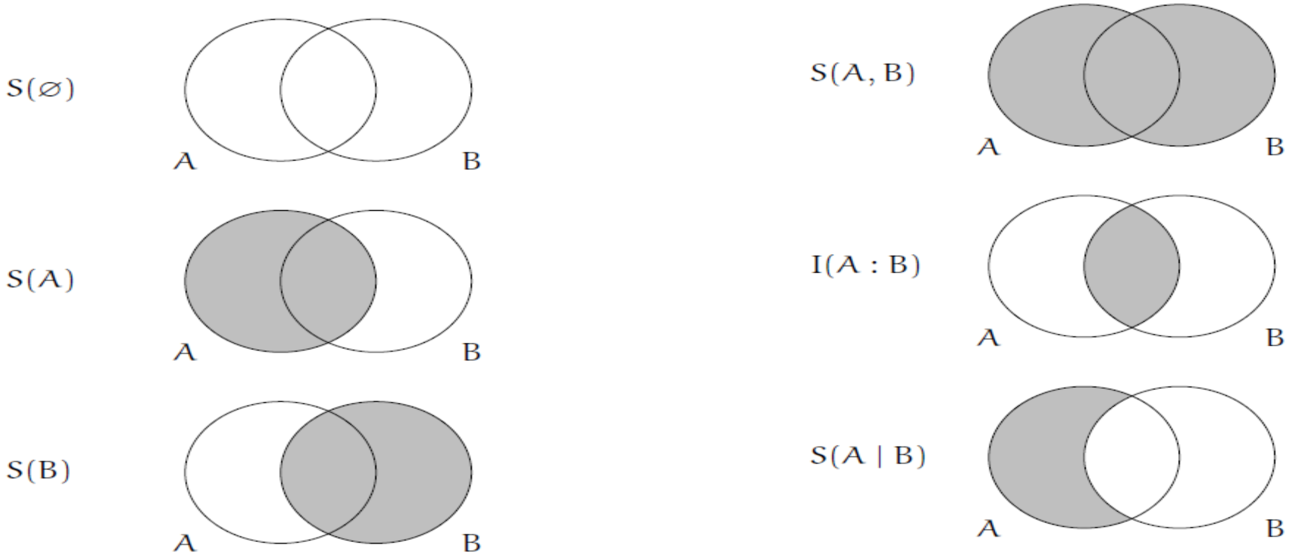
独立事件的概率是相乘的，而直觉上它们的信息量应当是相加的，为了把乘法变成加法，所以取了对数。

性质：对称性、非负性、可加性、极值性、上凸性。

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

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

图解



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

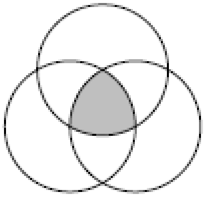
多维情况

联合熵

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

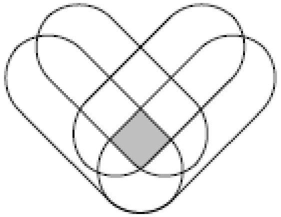
互信息：

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



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

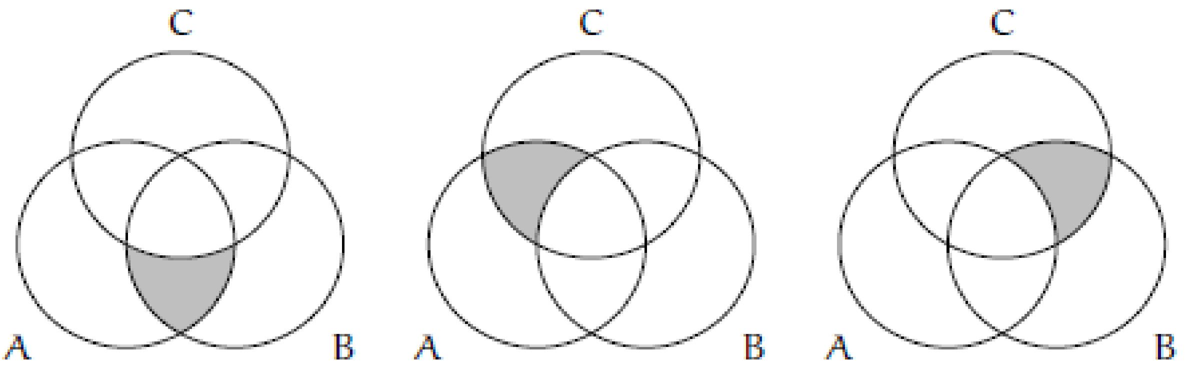
例如四维情况：



Conditional mutual information

Conditional mutual information [12] The average mutual information between A and B given C.

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



I(A : B | C) I(A : C | B) I(B : C | A)

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

注意：

1、对于三个及三个以上的变量，互信息可正，可负，也可以是零。但是对于两个变量来说，互信息是非负的。

例如：A,B是两个（0,1）分布且概率为0.5，并且独立。C为A,B的平方和，则

$$I(A:B)=0$$

$$I(A:B\,|\,C)=S(A\,|\,C)-S(A\,|\,B,C)=\frac{1}{2}\text{-}0=\frac{1}{2}$$

$$I(A:B:C)=I(A:B)-I(A:B\,|\,C)=0-\frac{1}{2}=-\frac{1}{2}$$

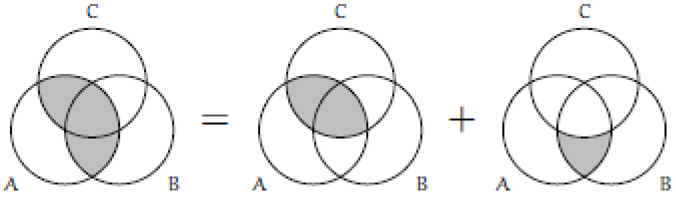
2、条件互信息是非负的。如果条件互信息为零，则A,B在给定C的条件下独立，这是个充要条件，即

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

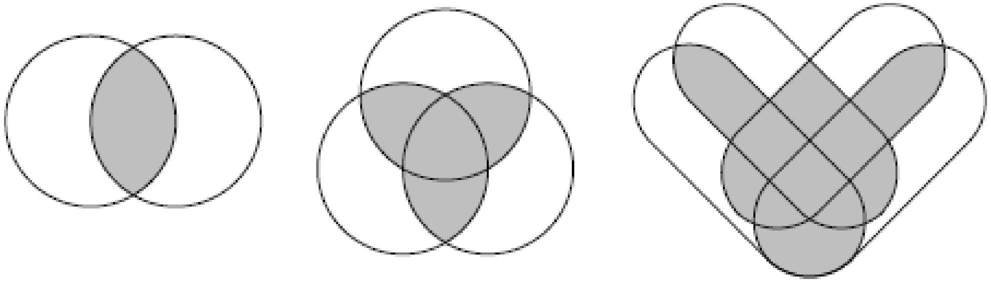
链式规则：

$$S(A,B) = S(A \mid B) + S(B)$$

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

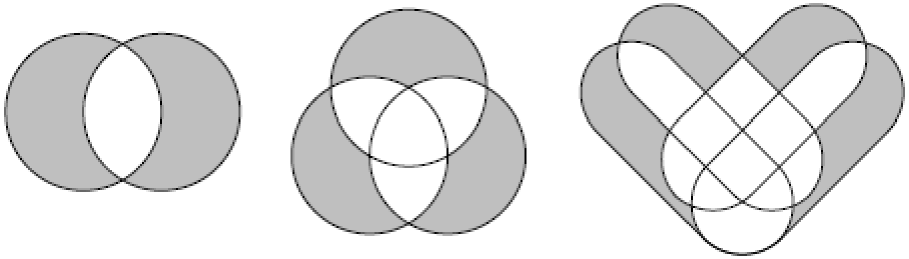


Binding information (Dual total correlation)

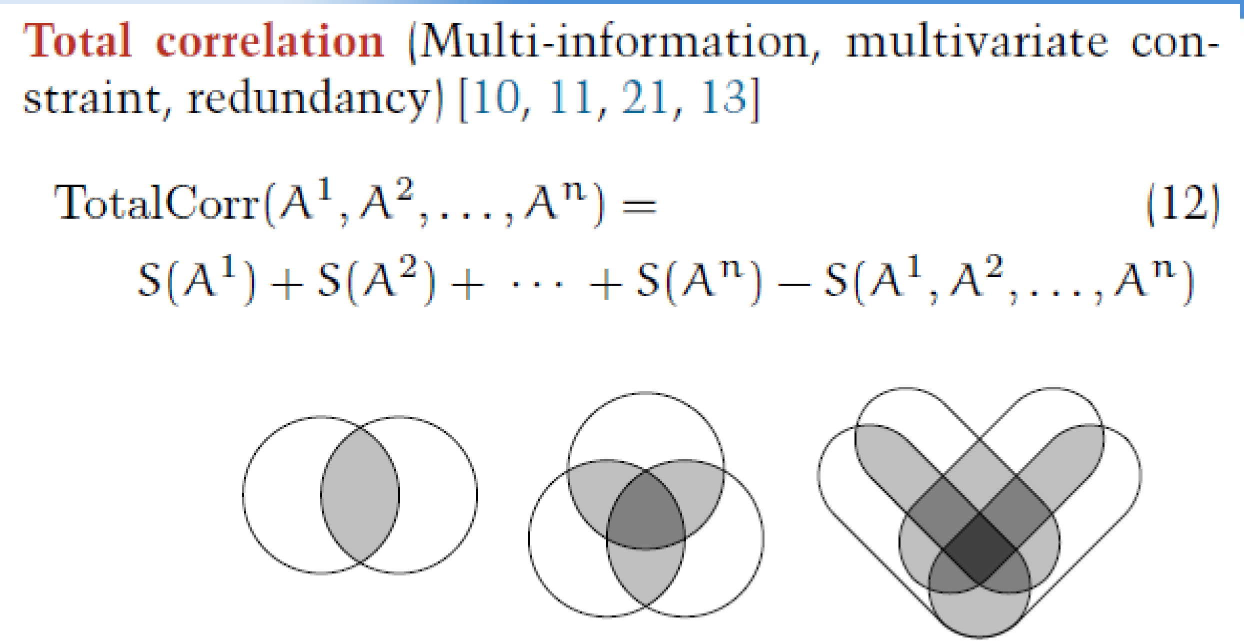


$$\text{Binding}(A:B) = I(A:B)$$

Residual entropy (erasure entropy, independent information,variation of information, shared information distance) [18, 19, 20, 3]

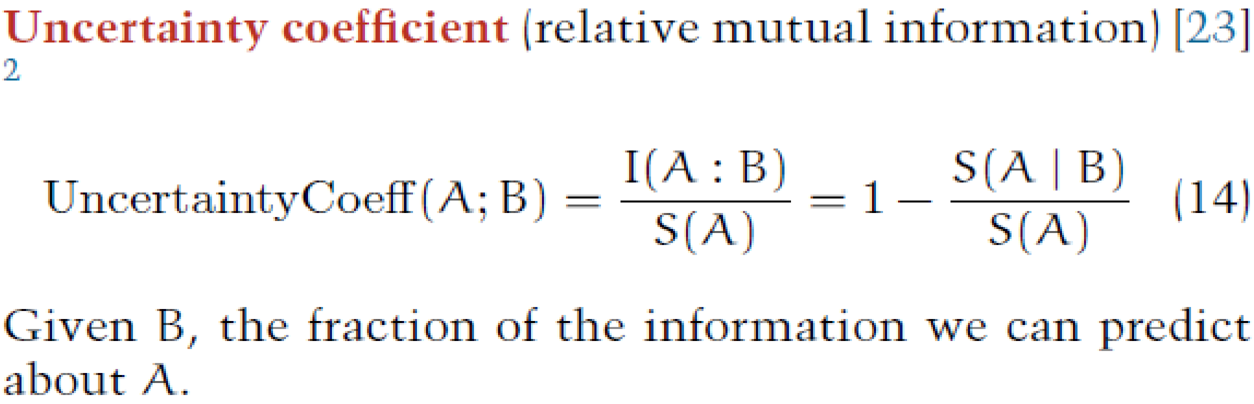


Total correlation



Total correlation (Multi-information, multivariate constraint, redundancy) [10, 11, 21, 13]

$$\text{TotalCorr}(A^1, A^2, \ldots, A^n) = \tag{12}$$
$$S(A^1) + S(A^2) + \cdots + S(A^n) - S(A^1, A^2, \ldots, A^n)$$

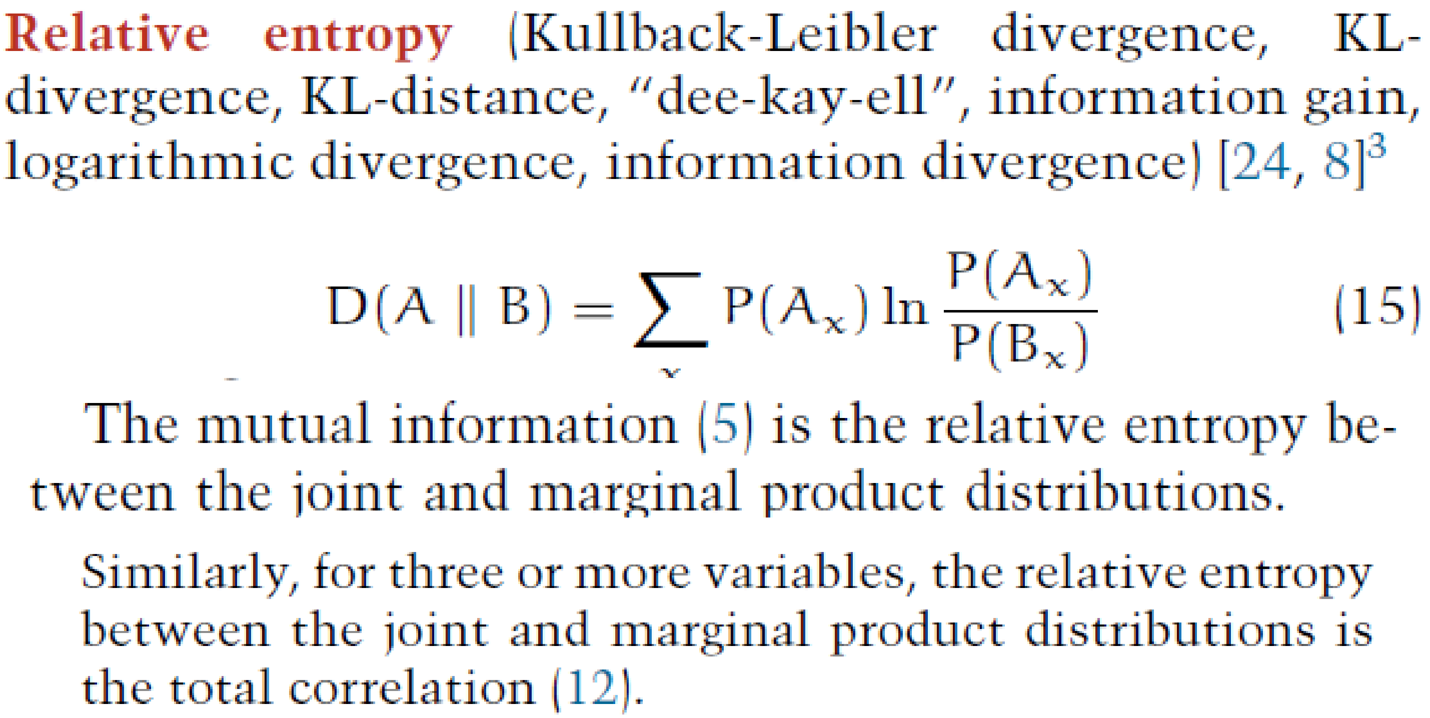


Uncertainty coefficient (relative mutual information) [23] [2]

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

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

Relative entropy



Relative entropy (Kullback-Leibler divergence, KL-divergence, KL-distance, "dee-kay-ell", information gain, logarithmic divergence, information divergence) [24, 8][3]

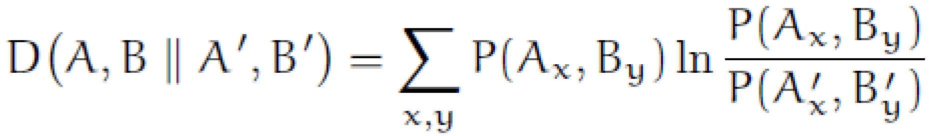
$$D(A \parallel B) = \sum_{x} P(A_x) \ln \frac{P(A_x)}{P(B_x)} \qquad (15)$$

The mutual information (5) is the relative entropy between the joint and marginal product distributions.

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

Relative joint entropy

Relative conditional entropy



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

Relative mutual information

Relative conditional mutual information

Relative relative entropy

Jeffreys entropy

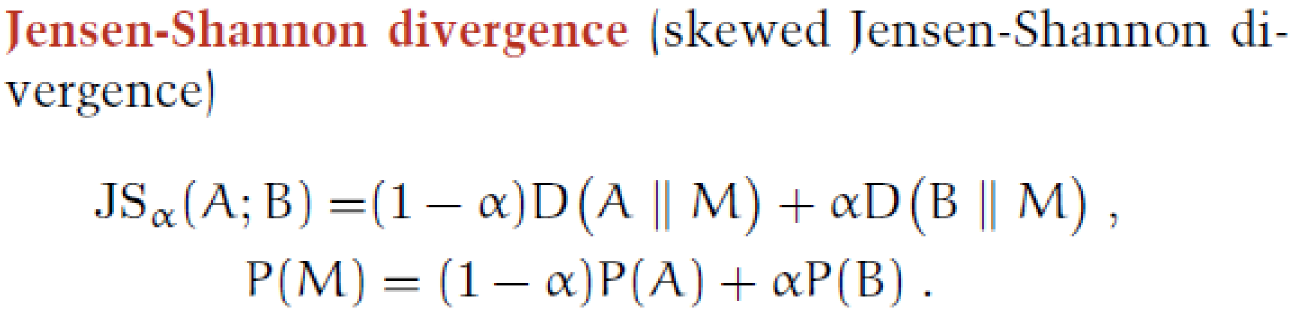
$$\begin{aligned}\mathrm{Jeffreys}(A;B) &\equiv \tfrac{1}{2}D(A \parallel B) + \tfrac{1}{2}D(B \parallel A) \\ &= \tfrac{1}{2}\sum_x P(A_x)\ln\frac{P(A_x)}{P(B_x)} + \tfrac{1}{2}\sum_x P(B_x)\ln\frac{P(B_x)}{P(A_x)} \\ &= \tfrac{1}{2}\sum \big(P(A_x) - P(B_x)\big)\ln\frac{P(A_x)}{P(B_x)}\end{aligned}$$

Jensen-Shannon entropy

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

where

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



Jensen-Shannon divergence (skewed Jensen-Shannon divergence)

$$JS_{\alpha}(A;B) = (1-\alpha)D(A \parallel M) + \alpha D(B \parallel M) ,$$
$$P(M) = (1-\alpha)P(A) + \alpha P(B) .$$

On measures of entropy and information. Gavin E.Crooks, 2016. 08. 16

连续型随机变量的shannon 熵

$$H(X) = -\int p(x)\ln p(x)dx$$

连续型随机变量的微分熵虽然具有离散熵的主要特征，但不具备非负性。

2、Tsallis entropy

自然界中存在许多用香浓熵不能完全描述的系统：长程相互作用、长程微观记忆（例如：非Markov 随机过程）、星系奇异速度、Levy反常扩散、一维耗散系统等等。

$$\begin{cases} \dfrac{dy}{dx} = y \\ y(0) = 1 \end{cases} \qquad \begin{cases} \dfrac{dy}{dx} = y^{q} \\ y(0) = 1 \end{cases}$$

解得：$y = e^{x}$　　　解得：$y = [1 + (1-q)x]^{1/(1-q)}$

其反函数为：

$$y = \ln x \qquad y = \frac{x^{1-q} - 1}{1-q}$$

定义

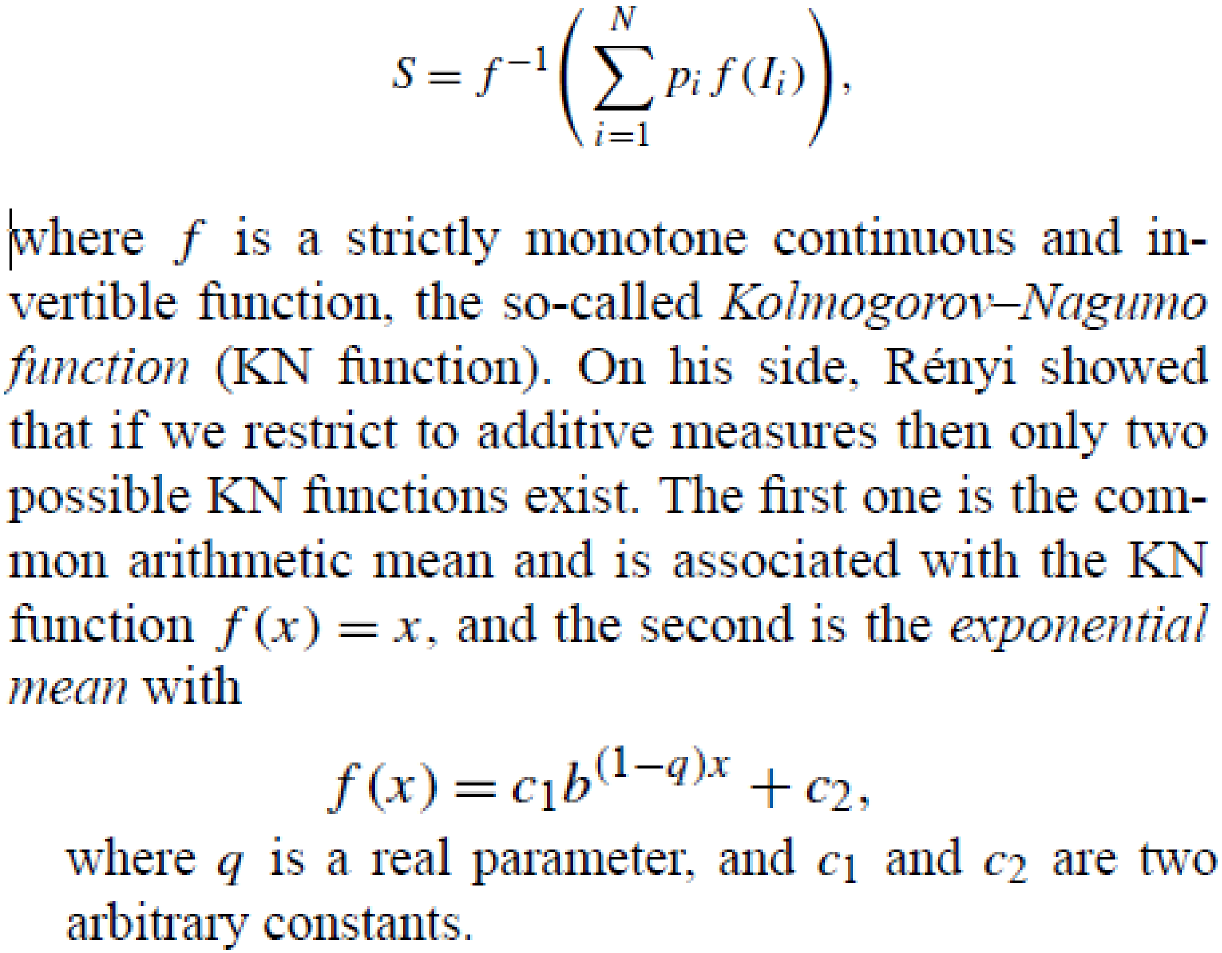
$$\begin{aligned} S_{\mathrm{T}}(P,q) &= \frac{\sum_{i=1}^{N} p_i^q - 1}{1-q} \\ &= \frac{1}{q-1} \sum_{i=1}^{N} p_i \left(1 - p_i^{q-1}\right). \end{aligned}$$

Tsallis entropy extends to a *pseudo-additive* law

$$\begin{aligned} S_{\mathrm{T}}(A \cap B) = S_{\mathrm{T}}(A) + S_{\mathrm{T}}(B|A) \\ + (1-q) S_{\mathrm{T}}(A) S_{\mathrm{T}}(B|A), \end{aligned}$$

3、Renyi's entropy

The definition of the average must be extended to the quasi-arithmetic or Quasi-linear mean defined as



$$S = f^{-1}\left(\sum_{i=1}^{N} p_i f(I_i)\right),$$

where f is a strictly monotone continuous and invertible function, the so-called *Kolmogorov–Nagumo function* (KN function). On his side, Rényi showed that if we restrict to additive measures then only two possible KN functions exist. The first one is the common arithmetic mean and is associated with the KN function $f(x) = x$, and the second is the *exponential mean* with

$$f(x) = c_1 b^{(1-q)x} + c_2,$$

where q is a real parameter, and c_1 and c_2 are two arbitrary constants.

The exponential mean leads to *Rényi's information measure* or *Rényi's entropy*

$$S_{\mathrm{R}}(P,q) = \frac{1}{1-q}\log_b \sum_{i=1}^{N} p_i^q ,$$

with b the logarithm base (we will from now on assume the natural base, $b = e$, for Rényi's entropy either). For $q \to 1$ Rényi's measure becomes Shannon's entropy.

Therefore Shannon's information measure is an averaged information in the ordinary sense, while Rényi's measure represents an exponential mean over the same elementary information gains $\log(\frac{1}{p_i})$.

$$S_S(P) = \langle I_i \rangle_{\text{lin}} = \left\langle \log\left(\frac{1}{p_i}\right) \right\rangle_{\text{lin}},$$

where we will call the quantity

$$I_i = \log\left(\frac{1}{p_i}\right),$$

The generalized q-logarithm function

$$\log_q x = \frac{x^{1-q} - 1}{1 - q}$$

$$S_T(P,q) = \left\langle \log_q\left(\frac{1}{p_i}\right) \right\rangle_{\text{lin}} = \langle I_i \rangle_{\text{lin}}$$

The generalized q-exponential function

$$e_q^x = [1 + (1-q)x]^{\frac{1}{1-q}},$$

$$S_R(P,q) = \left\langle \log\left(\frac{1}{p_i}\right) \right\rangle_{\text{exp}}$$

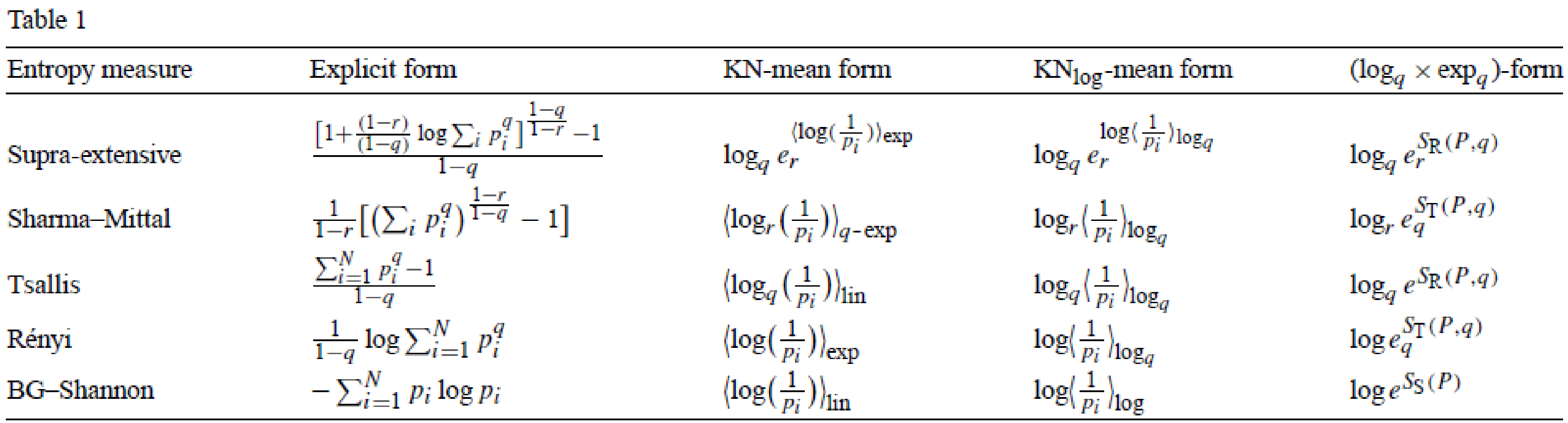
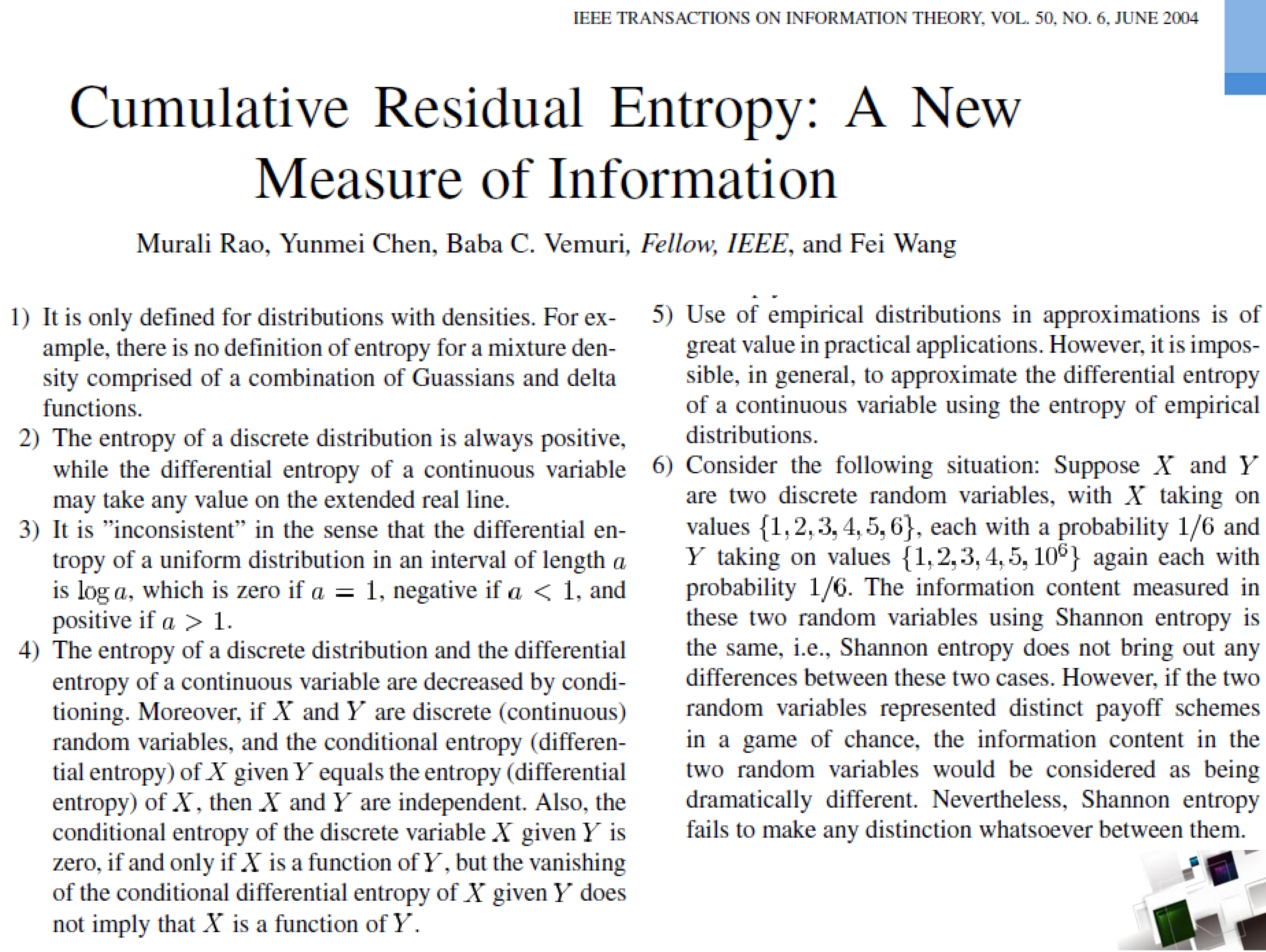


Table 1

Entropy measure	Explicit form	KN-mean form	$\mathrm{KN}_{\log}$-mean form	$(\log_q \times \exp_q)$-form
Supra-extensive	$\frac{[1+\frac{(1-r)}{(1-q)}\log\sum_i p_i^q]^{\frac{1-q}{1-r}}-1}{1-q}$	$\log_q e_r^{\langle\log(\frac{1}{p_i})\rangle_{\exp}}$	$\log_q e_r^{\log\langle\frac{1}{p_i}\rangle_{\log_q}}$	$\log_q e_r^{S_{\mathrm{R}}(P,q)}$
Sharma–Mittal	$\frac{1}{1-r}[(\sum_i p_i^q)^{\frac{1-r}{1-q}}-1]$	$\langle\log_r(\frac{1}{p_i})\rangle_{q\text{-}\exp}$	$\log_r\langle\frac{1}{p_i}\rangle_{\log_q}$	$\log_r e_q^{S_{\mathrm{T}}(P,q)}$
Tsallis	$\frac{\sum_{i=1}^N p_i^q-1}{1-q}$	$\langle\log_q(\frac{1}{p_i})\rangle_{\mathrm{lin}}$	$\log_q\langle\frac{1}{p_i}\rangle_{\log_q}$	$\log_q e^{S_{\mathrm{R}}(P,q)}$
Rényi	$\frac{1}{1-q}\log\sum_{i=1}^N p_i^q$	$\langle\log(\frac{1}{p_i})\rangle_{\exp}$	$\log\langle\frac{1}{p_i}\rangle_{\log_q}$	$\log e_q^{S_{\mathrm{T}}(P,q)}$
BG–Shannon	$-\sum_{i=1}^N p_i\log p_i$	$\langle\log(\frac{1}{p_i})\rangle_{\mathrm{lin}}$	$\log\langle\frac{1}{p_i}\rangle_{\log}$	$\log e^{S_{\mathrm{S}}(P)}$

$$\begin{tikzcd}
\begin{array}{c}\text{Sharma-Mittal}\\ \log_r \left\langle \frac{1}{p_i} \right\rangle_{\log_q}\end{array} \arrow[dr, "r \to 1"] \arrow[drrr, "r \to q"] & & & & \begin{array}{c}\text{Supra-extensive}\\ \log_q e_r^{\log\left\langle \frac{1}{p_i} \right\rangle_{\log_q}}\end{array} \arrow[dl, "r \to 1"'] \arrow[dlll, "r \to q"'] \\
 & \begin{array}{c}\text{Rényi}\\ \log \left\langle \frac{1}{p_i} \right\rangle_{\log_q}\end{array} \arrow[dr, "q \to 1"] & & \begin{array}{c}\text{Tsallis}\\ \log_q \left\langle \frac{1}{p_i} \right\rangle_{\log_q}\end{array} \arrow[dl, "q \to 1"'] & \\
 & & \begin{array}{c}\text{Shannon (Boltzmann-Gibbs)}\\ \log \left\langle \frac{1}{p_i} \right\rangle_{\log}\end{array} & &
\end{tikzcd}$$

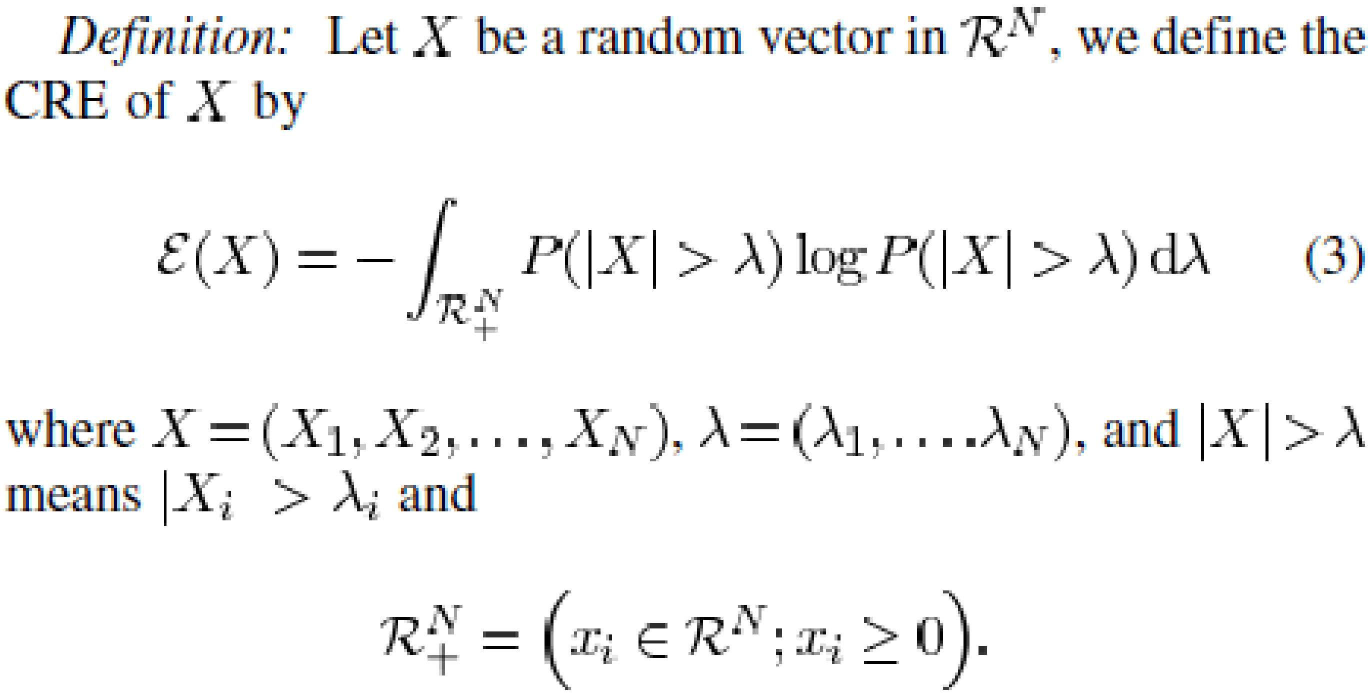


IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 50, NO. 6, JUNE 2004

Cumulative Residual Entropy: A New Measure of Information

Murali Rao, Yunmei Chen, Baba C. Vemuri, *Fellow, IEEE*, and Fei Wang

1) It is only defined for distributions with densities. For example, there is no definition of entropy for a mixture density comprised of a combination of Guassians and delta functions.
2) The entropy of a discrete distribution is always positive, while the differential entropy of a continuous variable may take any value on the extended real line.
3) It is ”inconsistent” in the sense that the differential entropy of a uniform distribution in an interval of length a is $\log a$, which is zero if $a = 1$, negative if $a < 1$, and positive if $a > 1$.
4) The entropy of a discrete distribution and the differential entropy of a continuous variable are decreased by conditioning. Moreover, if X and Y are discrete (continuous) random variables, and the conditional entropy (differential entropy) of X given Y equals the entropy (differential entropy) of X, then X and Y are independent. Also, the conditional entropy of the discrete variable X given Y is zero, if and only if X is a function of Y, but the vanishing of the conditional differential entropy of X given Y does not imply that X is a function of Y.
5) Use of empirical distributions in approximations is of great value in practical applications. However, it is impossible, in general, to approximate the differential entropy of a continuous variable using the entropy of empirical distributions.
6) Consider the following situation: Suppose X and Y are two discrete random variables, with X taking on values $\{1, 2, 3, 4, 5, 6\}$, each with a probability $1/6$ and Y taking on values $\{1, 2, 3, 4, 5, 10^6\}$ again each with probability $1/6$. The information content measured in these two random variables using Shannon entropy is the same, i.e., Shannon entropy does not bring out any differences between these two cases. However, if the two random variables represented distinct payoff schemes in a game of chance, the information content in the two random variables would be considered as being dramatically different. Nevertheless, Shannon entropy fails to make any distinction whatsoever between them.



Definition: Let X be a random vector in $\mathcal{R}^N$, we define the CRE of X by

$$\mathcal{E}(X) = -\int_{\mathcal{R}^N_+} P(|X| > \lambda) \log P(|X| > \lambda)\, \mathrm{d}\lambda \qquad (3)$$

where $X = (X_1, X_2, \ldots, X_N)$, $\lambda = (\lambda_1, \ldots . \lambda_N)$, and $|X| > \lambda$ means $|X_i \ > \lambda_i$ and

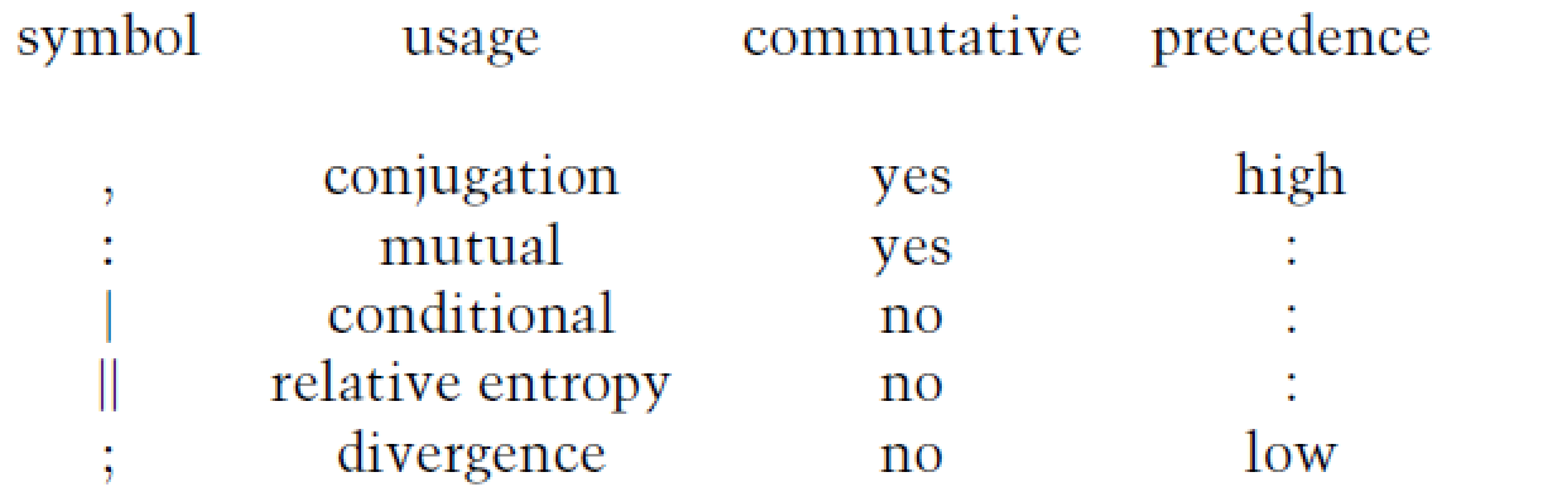
$$\mathcal{R}^N_+ = \Big(x_i \in \mathcal{R}^N; x_i \geq 0\Big).$$

Example 1: (CRE of the uniform distribution) Consider a general uniform distribution with the density function

$$p(x) = \begin{cases} \frac{1}{a}, & 0 \leq x \leq a \\ 0, & o.w. \end{cases} \tag{4}$$

Then its CRE is computed as follows:

$$\begin{aligned} \mathcal{E}(X) &= -\int_0^a P(|X \;\; > x) \log P(\; X \;\; > x) \, \mathrm{d}x \\ &= -\int_0^a \left(1 - \frac{x}{a}\right) \log \left(1 - \frac{x}{a}\right) \mathrm{d}x \\ &= \frac{1}{4} a. \end{aligned} \tag{5}$$

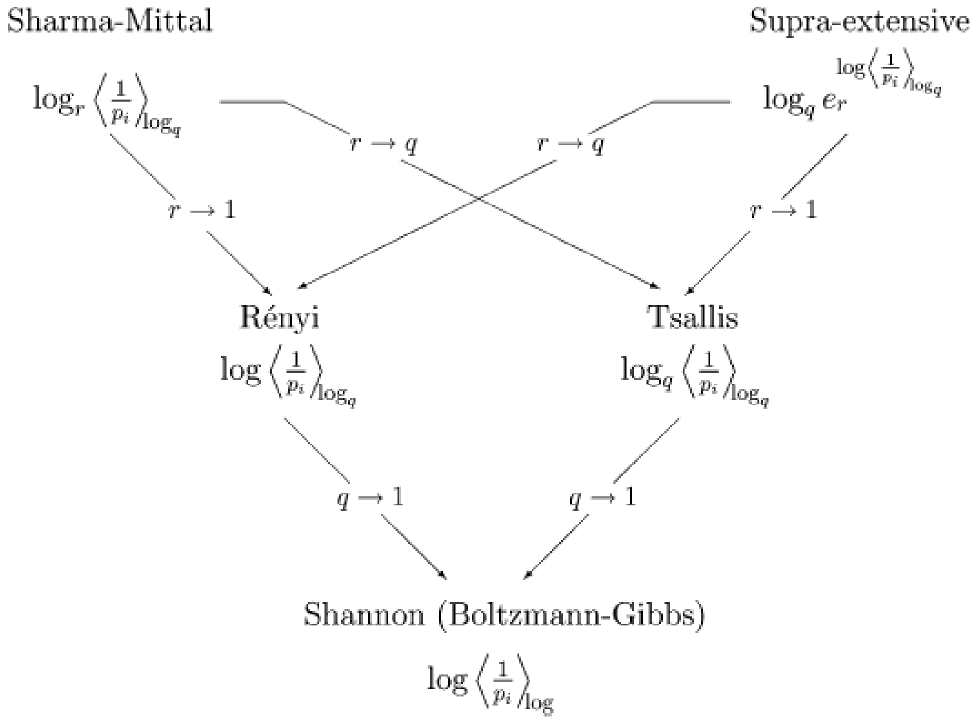


symbol	usage	commutative	precedence
,	conjugation	yes	high
:	mutual	yes	⋮
|	conditional	no	⋮
||	relative entropy	no	⋮
;	divergence	no	low

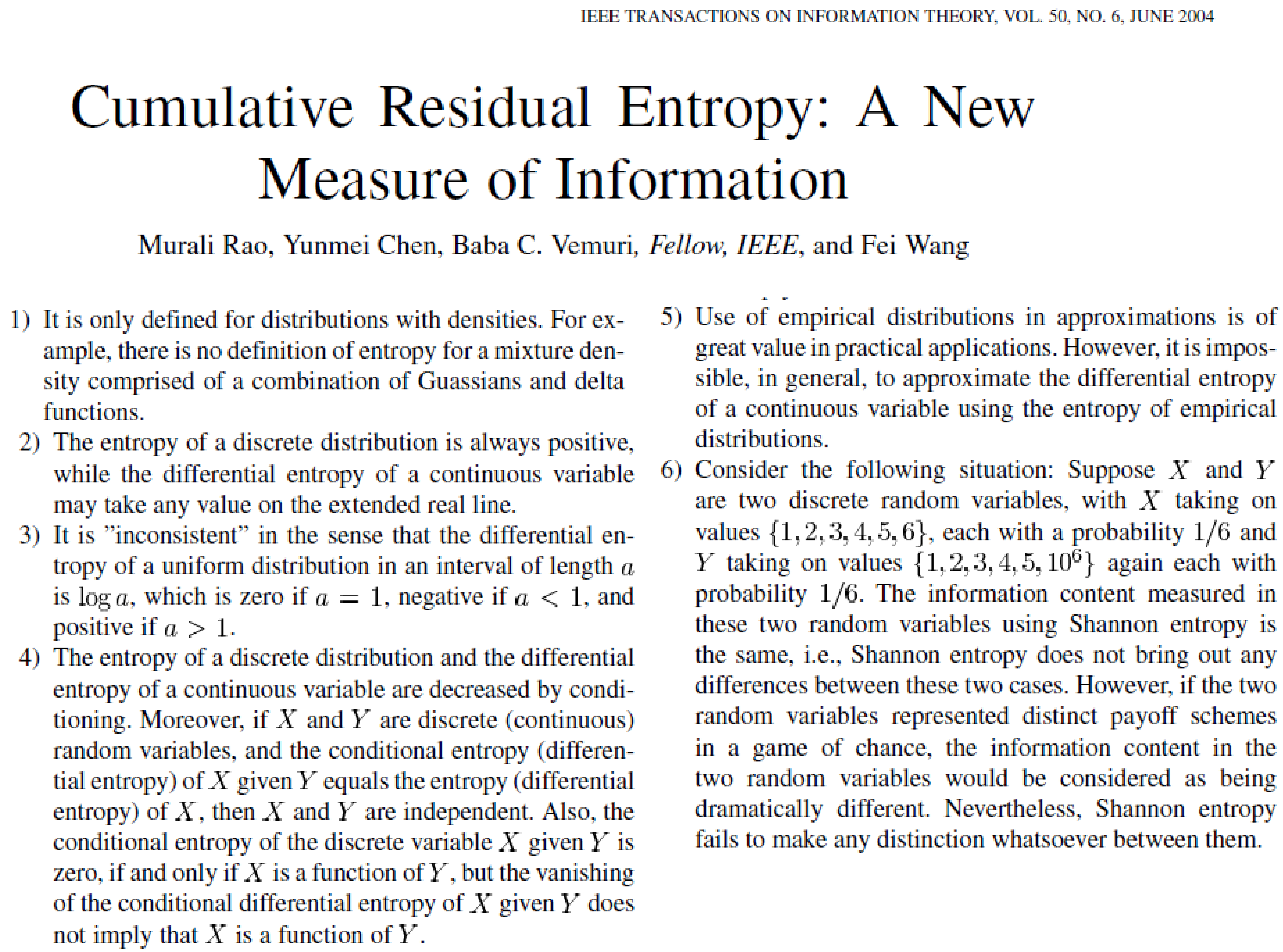
symbol	usage	commutative	precedence
,	conjugation	yes	high
:	mutual	yes	:
|	conditional	no	:
||	relative entropy	no	:
;	divergence	no	low

Table 1

Entropy measure	Explicit form	KN-mean form	$\text{KN}_{\log}$-mean form	$(\log_q \times \exp_q)$-form
Supra-extensive	$\frac{[1+\frac{(1-r)}{(1-q)}\log\sum_i p_i^q]^{\frac{1-q}{1-r}}-1}{1-q}$	$\log_q e_r^{\langle\log(\frac{1}{p_i})\rangle_{\exp}}$	$\log_q e_r^{\log\langle\frac{1}{p_i}\rangle_{\log_q}}$	$\log_q e_r^{S_{\mathrm{R}}(P,q)}$
Sharma–Mittal	$\frac{1}{1-r}[(\sum_i p_i^q)^{\frac{1-r}{1-q}}-1]$	$\langle\log_r(\frac{1}{p_i})\rangle_{q\text{-}\exp}$	$\log_r\langle\frac{1}{p_i}\rangle_{\log_q}$	$\log_r e_q^{S_{\mathrm{T}}(P,q)}$
Tsallis	$\frac{\sum_{i=1}^N p_i^q-1}{1-q}$	$\langle\log_q(\frac{1}{p_i})\rangle_{\mathrm{lin}}$	$\log_q\langle\frac{1}{p_i}\rangle_{\log_q}$	$\log_q e^{S_{\mathrm{R}}(P,q)}$
Rényi	$\frac{1}{1-q}\log\sum_{i=1}^N p_i^q$	$\langle\log(\frac{1}{p_i})\rangle_{\exp}$	$\log\langle\frac{1}{p_i}\rangle_{\log_q}$	$\log e_q^{S_{\mathrm{T}}(P,q)}$
BG–Shannon	$-\sum_{i=1}^N p_i\log p_i$	$\langle\log(\frac{1}{p_i})\rangle_{\mathrm{lin}}$	$\log\langle\frac{1}{p_i}\rangle_{\log}$	$\log e^{S_{\mathrm{S}}(P)}$



Sharma-Mittal
$\log_r\left\langle\frac{1}{p_i}\right\rangle_{\log_q}$
Supra-extensive
$\log_q e_r^{\log\left\langle\frac{1}{p_i}\right\rangle_{\log_q}}$
$r \to q$
$r \to q$
$r \to 1$
$r \to 1$
Rényi
$\log\left\langle\frac{1}{p_i}\right\rangle_{\log_q}$
Tsallis
$\log_q\left\langle\frac{1}{p_i}\right\rangle_{\log_q}$
$q \to 1$
$q \to 1$
Shannon (Boltzmann-Gibbs)
$\log\left\langle\frac{1}{p_i}\right\rangle_{\log}$



Cumulative Residual Entropy: A New Measure of Information

Murali Rao, Yunmei Chen, Baba C. Vemuri, *Fellow, IEEE*, and Fei Wang

1) It is only defined for distributions with densities. For example, there is no definition of entropy for a mixture density comprised of a combination of Guassians and delta functions.
2) The entropy of a discrete distribution is always positive, while the differential entropy of a continuous variable may take any value on the extended real line.
3) It is "inconsistent" in the sense that the differential entropy of a uniform distribution in an interval of length a is $\log a$, which is zero if $a = 1$, negative if $a < 1$, and positive if $a > 1$.
4) The entropy of a discrete distribution and the differential entropy of a continuous variable are decreased by conditioning. Moreover, if X and Y are discrete (continuous) random variables, and the conditional entropy (differential entropy) of X given Y equals the entropy (differential entropy) of X, then X and Y are independent. Also, the conditional entropy of the discrete variable X given Y is zero, if and only if X is a function of Y, but the vanishing of the conditional differential entropy of X given Y does not imply that X is a function of Y.
5) Use of empirical distributions in approximations is of great value in practical applications. However, it is impossible, in general, to approximate the differential entropy of a continuous variable using the entropy of empirical distributions.
6) Consider the following situation: Suppose X and Y are two discrete random variables, with X taking on values $\{1, 2, 3, 4, 5, 6\}$, each with a probability $1/6$ and Y taking on values $\{1, 2, 3, 4, 5, 10^6\}$ again each with probability $1/6$. The information content measured in these two random variables using Shannon entropy is the same, i.e., Shannon entropy does not bring out any differences between these two cases. However, if the two random variables represented distinct payoff schemes in a game of chance, the information content in the two random variables would be considered as being dramatically different. Nevertheless, Shannon entropy fails to make any distinction whatsoever between them.

Definition: Let X be a random vector in $\mathcal{R}^N$, we define the CRE of X by

$$\mathcal{E}(X) = -\int_{\mathcal{R}_+^N} P(|X| > \lambda) \log P(|X| > \lambda)\, \mathrm{d}\lambda \qquad (3)$$

where $X = (X_1, X_2, \ldots, X_N)$, $\lambda = (\lambda_1, \ldots .\lambda_N)$, and $|X| > \lambda$ means $|X_i \;\; > \lambda_i$ and

$$\mathcal{R}_+^N = \left(x_i \in \mathcal{R}^N; x_i \geq 0\right).$$

Example 1: (CRE of the uniform distribution) Consider a general uniform distribution with the density function

$$p(x) = \begin{cases} \frac{1}{a}, & 0 \leq x \leq a \\ 0, & o.w. \end{cases} \qquad (4)$$

Then its CRE is computed as follows:

$$\begin{aligned} \mathcal{E}(X) &= -\int_0^a P(|X \;\; > x) \log P(\; X \;\; > x)\, \mathrm{d}x \\ &= -\int_0^a \left(1 - \frac{x}{a}\right) \log\left(1 - \frac{x}{a}\right) \mathrm{d}x \\ &= \frac{1}{4} a. \end{aligned} \qquad (5)$$