

---

# ENTROPY HANDBOOK

## Definitions, Theorems, M-files

---

### 1 SHANNON'S ENTROPY

*Shannon stated the following result: 'The entropy can be interpreted as a measure of uncertainty'. It has properties that agree with the intuitive notion of what a measure of uncertainty should be. The Shannon's entropy is probably the best measure of the uncertainty.*

**Notation 1:** Let  $X$  be a finite discrete random variable with alphabet  $\Sigma_X = \{x_1, x_2, \dots, x_m\}$  by  $X$ . Let  $p(x)$  be an associated *probability mass function* (distribution) defined as  $p(x) = Pr\{X = x\}$ ,  $x \in \Sigma_X$ . Then the probability mass function  $p$  may be substituted by an ordered set of respective probabilities  $p = (p_1, p_2, \dots, p_m)$  satisfying conditions

$$p_i \geq 0, \quad \sum_{i=1}^m p_i = 1, \quad i = 1, 2, \dots, m.$$

We shall also use the relation  $X \sim p(x)$ .

**Notation 2:** We shall prefer to use the random variables rather than random outcomes. To every random outcome can be assigned according to some rule a random variable by a value  $x = X(a)$  for every outcome  $a$  belonging to the alphabet of outcomes.

**Definition 1:** The *Shannon's entropy*  $H^{(s)}(X)$  (further only  $H(X)$ ) of a discrete random variable  $X$  is defined by the expression

$$\begin{aligned} H(X) &= -c \sum_{x \in \Sigma_X} p(x) \ln p(x) \\ &= -c \sum_{i=1}^m p(x_i) \ln p(x_i) \\ &= -c \sum_{i=1}^m p_i \ln p_i \end{aligned}$$

where  $c$  denotes a positive constant which may be thought of as defining the measurement units.

**Remark 1:** We may also write  $H(p)$ ,  $H_m(p)$  or  $H(p_1, p_2, \dots, p_m)$  for the above quantity. The  $\ln$  is logarithm to the base  $e$  and entropy is expressed in *nats*. We will use the convention that  $0 \ln 0 = 0$ .

**Remark 2:** The logarithm need not be only to the base  $e$ . If the base of the logarithm is  $b$ , we will denote the entropy as  $H^{(b)}(X)$ . Often we use the logarithm of the base 2. In this case, the entropy will be measured in *bits* ( $1 \text{ nat} = \ln 2 \text{ bits}$ ).

**M-function:** *Computes Shannon's entropy of a discrete random variable  $X$ .*

```
function H=et1_shn(p,c)
[m,n]=size(p);
Err=1e-6;

if abs(1-sum(p))>Err
    error('Input vector must be a probability mass function');
end;

h1=(-1)*ones(m,n);
h2=((sign(p)+h1).*h1)+p;
H=h2*log(h2)'*(-c);
```

Note that MATLAB-based log means  $\ln$ .

**Synopsis:** `H=et1_shn(p,c)`

**Visual imagination:** By MATLAB command line `et1_img('shn')` Shannon's entropy of  $(p_1, 1 - p_1)$  probability mass function with  $c = 1$  is visualized.

**Theorem 1:** We have  $H(X) \geq 0$ .

**Theorem 2:**  $H^{(b)}(X) = (\log_b a)H^{(a)}(X)$ .

This property of entropy enables us to change the base of the logarithm in the definition. Entropy can be changed from one base to another by multiplying it the appropriate factor. We shall naturally put  $c = 1$  and measure the entropy in nats. As to some conversions, see the following table for  $b = e$ .

a	2	e	8	10	16
c	0.6931	1	2.0794	2.3025	2.7725

**Theorem 3:** If  $p_k = 1$  and  $p_i = 0$  ( $1 \leq i \leq m$ ;  $i \neq k$ ) then  $H(p_1, \dots, p_m) = 0$ .

*The entropy is equal to zero, if one of the numbers  $p_1, \dots, p_m$  is unity and all the others are zero. However it is the case of certainty but not uncertainty.*

**Theorem 4:** We have  $H(p_1, \dots, p_m, 0) = H(p_1, \dots, p_m)$ .

**Theorem 5:** For any probability mass function  $p$  given by  $(p_1, p_2, \dots, p_m)$  we have

$$H(p_1, p_2, \dots, p_m) \leq H_m\left(\frac{1}{m}, \dots, \frac{1}{m}\right).$$

with equality if and only if  $X$  is uniformly distributed over  $\Sigma_X$ . At the same time we simply maximize  $H(p)$  subject to  $\sum_i p_i = 1$ .

**Corollary 1:** Equivalently to the previous theorem

$$H(X) \leq \ln |\Sigma_X|$$

where  $|\Sigma_X|$  denotes the number of elements of the alphabet  $\Sigma_X$ . *Shannon's entropy assumes its largest value in the case of the uniform probability mass function.*

**Theorem 6:** The following equality holds

$$H(p_1, \dots, p_{m-1}, p', p'') = H(p_1, \dots, p_m) + p_m H\left(\frac{p'}{p_m}, \frac{p''}{p_m}\right),$$

where  $p_m = p' + p''$ .

**Theorem 7:**  $H(p)$  is a concave function of  $p$ .

**Theorem 8:** Let  $X$  and  $Y$  be independent random variables. If  $Z = X + Y$  then  $H(Y) \leq H(Z)$  and  $H(X) \leq H(Z)$ .

**Theorem 9:** *Markov's inequality.* Let  $p(x)$  be a probability mass function. For all  $d \geq 0$

$$Pr\{p(x) \leq d\} \log_2\left(\frac{1}{d}\right) \leq H(X).$$

**Example 1:** Let

$$X = \begin{cases} a & \text{with probability } 1/2, \\ b & \text{with probability } 1/4, \\ c & \text{with probability } 1/8, \\ d & \text{with probability } 1/8. \end{cases}$$

The entropy  $H(X)$  of  $X$  is

```
>>H=et1_shn([1/2 1/4 1/8 1/8],1)
```

```
>>H =
```

```
1.2130
```

in *nats* or

```
>>H=et1_shn([1/2 1/4 1/8 1/8],1/log(2))
```

```
>>H =
```

```
1.7500
```

in *bits* (7/4 bits).

**Remark 3:** Note that entropy is a functional of the probability mass function of  $X$ . It does not depend on the actual values taken by the random variable  $X$ , but only on the probabilities.

**Remark 4:** It is possible to define the entropy axiomatically by formulation certain properties that the entropy of a random variable must satisfy. There exist various axiomatic formulations which result in the same, similar or different definitions of entropy according to set of prior requirements. We shall also mention some further important definitions in next sections.

**Remark 5:** Definition 1 introduced for the case of random variable  $X$ , can be generalized straightforwardly to probabilistic experiment  $A$  having  $m$  possible outcomes  $\Sigma_A = \{a_1, a_2, \dots, a_m\}$  with the respective probabilities  $(p_1, p_2, \dots, p_m)$ . We use  $H(A)$  or  $H(p)$  with the same practical meaning (we refer to  $p$  as  $(p_1, p_2, \dots, p_m)$ )

$$\begin{aligned} H(A) &= -c \sum_{a \in \Sigma_A} p(a) \ln p(a) \\ &= -c \sum_{i=1}^m p(a_i) \ln p(a_i) \\ &= -c \sum_{i=1}^m p_i \ln p_i. \end{aligned}$$

**Definition 2:** *Extension of Shannon's entropy for an incomplete probability mass function.* Let us consider the incomplete probability mass function  $p$  defined by a finite sequence of non-negative numbers  $p = \{p_1, p_2, \dots, p_m\}$  such that

$$0 \leq \sum_{i=1}^m p_i < 1.$$

Shannon's entropy of such an incomplete probability mass function  $p$  is defined as

$$H_m(p) = \frac{\sum_{i=1}^m p_i \ln \frac{1}{p_i}}{\sum_{i=1}^m p_i}.$$

**Definition 3:** *Extension of Shannon's entropy for a denumerable infinite probability distribution.* Let us consider the denumerable infinite probability mass function

$$p = \{p_0, p_1, \dots\}, \quad p_i \geq 0, \quad \sum_{i=0}^{\infty} p_i = 1.$$

Shannon's entropy of such an infinite probability mass function  $p$  is defined as

$$H(p) = - \sum_{i=0}^{\infty} p_i \ln p_i$$

provided the series is convergent.

**Example 2:** Let

$$p_i = \frac{E^i}{(1 + E)^{i+1}} \quad i = 0, 1, 2, \dots$$

be an infinite probability mass function where  $E$  denotes a mathematical expectation. Then

$$H(p) = - \sum_{i=0}^{\infty} \frac{E^i}{(1 + E)^{i+1}} \ln \frac{E^i}{(1 + E)^{i+1}}$$

and after summation

$$H(E) = (1 + E) \ln(1 + E) - E \ln E$$

may be obtained. In order to visualize this dependence, call `et1_img('eee')`.

## 2 JOINT ENTROPY

*Joint entropy introduces a means for determining a measure of uncertainty of a pair of discrete random variables which are described by a joint distribution.*

**Notation 1:** Let us consider two random variables  $X$  and  $Y$  with alphabets  $\Sigma_X = \{x_1, x_2, \dots, x_m\}$  and  $\Sigma_Y = \{y_1, y_2, \dots, y_n\}$  respectively. Let  $p(x, y)$  be an associated joint probability mass function  $p(x, y) = Pr\{X = x \ \& \ Y = y\}$ ,  $x \in \Sigma_X$ ,  $y \in \Sigma_Y$ . We shall also use the relation  $(X, Y) \sim p(x, y)$ . We may substitute the probability mass function  $p$  by an ordered set of the respective probabilities  $p_{11}, p_{12}, \dots, p_{mn}$  satisfying the conditions

$$p_{ij} \geq 0, \quad \sum_{i=1}^m \sum_{j=1}^n p_{ij} = 1, \quad i = 1, 2, \dots, m \quad j = 1, 2, \dots, n.$$

**Definition 1:** The *joint entropy*  $H(X, Y)$  of a pair of discrete random variables  $(X, Y)$  with a joint distribution  $p(x, y)$  is defined by the expression

$$\begin{aligned} H(X, Y) &= - \sum_{x \in \Sigma_X} \sum_{y \in \Sigma_Y} p(x, y) \ln p(x, y) \\ &= - \sum_{i=1}^m \sum_{j=1}^n p(x_i, y_j) \ln p(x_i, y_j) \\ &= - \sum_{i=1}^m \sum_{j=1}^n p_{ij} \ln p_{ij}. \end{aligned}$$

When we use the logarithm of the base 2,  $H^{(2)}(X, Y) = (1/\ln 2)H(X, Y)$ .

**M-function:** *Computes the joint entropy of a pair of discrete random variables with a joint distribution  $p$ .*

```
function H=et1_jnt(p)

[m,n]=size(p);
Err=1e-6;

if abs(1-sum(sum(p)))>Err
    error('Input matrix must be a joint distribution');
end;

h1=(-1)*ones(m,n);
h2=((sign(p)+h1).*h1)+p;
h3=p.*log(h2);
H=-sum(sum(h3));
```

**Synopsis:** `H=et1_jnt(p)`

**Visual imagination:** By command line `et1_img('jnt')` the joint entropy of a distribution given by respective probabilities  $p_{11} = p_1 p_2, p_{12} = p_1(1 - p_2), p_{21} = p_2(1 - p_1), p_{22} = (1 - p_1)(1 - p_2)$

is visualized.

**Theorem 1:** In the case of independence of the random variables i.e.  $p(x, y) = p(x)p(y)$ ,

$$H(X, Y) = H(X) + H(Y).$$

**Theorem 2:** *Independence bound on entropy.*  $H(X_1, X_2, \dots, X_m) \leq \sum_{i=1}^m H(X_i)$ , with the equality if and only if the random variables  $X_i$  are independent.

**Example 1:** Let  $(X, Y)$  with alphabets  $\Sigma_X = \{1, 2, 3, 4\}$  and  $\Sigma_Y = \{1, 2, 3, 4\}$  have the following joint probability mass function

X \ Y	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

The joint entropy  $H(X, Y)$  is

```
>>p=[1/8 1/16 1/32 1/32;1/16 1/8 1/32 1/32;
      1/16 1/16 1/16 1/16;1/4 0 0 0];
>>H=et1_jnt(p)
>>H =
      2.3394      % nats
>>H=1/log(2)*H
>>H =
      3.375      % bits
```

or  $27/8$  bits.

**Remark 1:** Let us consider two probabilistic experiments  $A$  and  $B$  whose possible outcomes are  $a_1, a_2, \dots, a_m$  and  $b_1, b_2, \dots, b_n$  respectively. Let us introduce a compound probabilistic experiment which consists of both of the experiments  $A$  and  $B$ . A possible outcome of this experiment will be a pair  $(a_i, b_j)$ . Let us denote by  $p_{ij}$  or equivalently by  $p(a_i, b_j)$  the probability of the outcome  $(a_i, b_j)$  of the compound probabilistic experiment. The corresponding joint entropy will be

$$\begin{aligned} H(A, B) &= - \sum_{i=1}^m \sum_{j=1}^n p(a_i, b_j) \ln p(a_i, b_j) \\ &= - \sum_{i=1}^m \sum_{j=1}^n p_{ij} \ln p_{ij}. \end{aligned}$$

### 3 CONDITIONAL ENTROPY

*Conditional entropy allows one to measure the amount of information which is contributed by one random variable about a second random variable.*

**Notation 1:** Consider two random variables  $X, Y$ ,  $(X, Y) \sim p(x, y)$  with alphabets  $\Sigma_X = \{x_1, x_2, \dots, x_m\}$  and  $\Sigma_Y = \{y_1, y_2, \dots, y_n\}$ , which are not necessarily independent from a probabilistic point of view. In other words, the relation  $p(x, y) = p(x)p(y)$  does not necessarily hold, but rather, one uses a conditional probability mass function  $p(y|x)$  defined as  $p(y|x) = Pr\{Y = y|X = x\}$ ,  $x \in \Sigma_X$ ,  $y \in \Sigma_Y$ . In this case  $p(x, y) = p(x)p(y|x)$  and  $p(x)$  means a marginal distribution. For the marginal distribution we have used the same notation as for the single probability mass function because a distinguishability single from marginal is always understood from the context.

**Notation 2:** The marginal distribution  $p(x)$ ,  $x \in \Sigma_X$  is the probability of  $x$  regardless of the occurrence of  $y$ . It is given as

$$p(x) = \sum_{y \in \Sigma_Y} p(x, y), \quad x \in \Sigma_X$$

or equivalently

$$p(x_k) = \sum_{j=1}^n p_{kj} = \sum_{j=1}^n p(x_k, y_j), \quad k = 1, 2, \dots, m.$$

Similarly the marginal distribution  $p(y)$  is probability of  $y$  regardless of the occurrence of  $x$ . It is given as

$$p(y) = \sum_{x \in \Sigma_X} p(x, y), \quad y \in \Sigma_Y.$$

We write

$$p(x|y) = \frac{p(x, y)}{p(y)}, \quad p(y|x) = \frac{p(x, y)}{p(x)}$$

for the conditional probabilities at  $p(y) > 0$  or  $p(x) > 0$  and  $x \in \Sigma_X$ ,  $y \in \Sigma_Y$ . Similarly

$$p(x_k|y_l) = \frac{p(x_k, y_l)}{p(y_l)}, \quad p(y_l|x_k) = \frac{p(x_k, y_l)}{p(x_k)}$$

for  $k = 1, 2, \dots, m$  and  $l = 1, 2, \dots, n$ .

**Definition 1:** *Entropy of  $Y$  calculated on the assumption that  $X$  occurred.*

If  $(X, Y) \sim p(x, y)$ , then the *conditional entropy*  $H(Y|X)$  is defined as

$$H(Y|X) = \sum_{x \in \Sigma_X} p(x)H(Y|X = x)$$

where  $H(Y|X = x)$  is given by the expression

$$H(Y|X = x) = - \sum_{y \in \Sigma_Y} p(y|x) \ln p(y|x).$$

**Corollary 1:** Due to Definition 1 one may observe equivalent expressions

$$H(Y|X) = - \sum_{x \in \Sigma_X} \sum_{y \in \Sigma_Y} p(x, y) \ln p(y|x) = - \sum_{x \in \Sigma_X} \sum_{y \in \Sigma_Y} p(x, y) \ln \frac{p(x, y)}{\sum_{y \in \Sigma_Y} p(x, y)}$$

Similarly we can define  $H(X|Y)$ .

**Synopsis:** Repeatedly use of `H=et1_shn(p, c)` following from Definition 1.

**Visual imagination:** By command line `et1_img('cnd')` we can display the conditional entropy of a distribution given by respective probabilities  $p_{11} = p_1p_2, p_{12} = p_1(1 - p_2), p_{21} = p_2(1 - p_1), p_{22} = (1 - p_1)(1 - p_2)$  as  $p_1H(p_2, 1 - p_2) + (1 - p_1)H(p_2, 1 - p_2)$  or similarly as  $p_2H(p_1, 1 - p_1) + (1 - p_2)H(p_1, 1 - p_1)$ .

**Theorem 1:** We have  $H(Y|X) \geq 0$

**Theorem 2:** *Chain rule.*  $H(X, Y) = H(X) + H(Y|X)$ . Thus  $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$ .

**Theorem 3:** The following relation holds:

$$0 \leq H(Y|X) \leq H(Y)$$

with equality to  $H(Y)$  if and only if  $X$  and  $Y$  are independent (*conditioning reduces entropy*).

**Theorem 4:** The following inequality holds  $H(Y) - H(X) \leq H(Y|X)$ .

**Theorem 5:** *Information balance.*  $H(X) + H(Y|X) = H(Y) + H(X|Y)$ .  $H(Y|X) \neq H(X|Y)$  in general.

**Theorem 6:** Let  $X_1, X_2, \dots, X_m$  be drawn according to  $p(x_1, x_2, \dots, x_m)$ . Then  $H(X_1, X_2, \dots, X_m) = \sum_{i=1}^m H(X_i|X_{i-1}, \dots, X_1)$ .

**Example 1:** Let  $(X, Y)$  have the joint distribution from Example 1 of the joint entropy section. The marginal distribution of  $X$  is  $(1/2, 1/4, 1/8, 1/8)$  and the marginal distribution of  $Y$  is  $(1/4, 1/4, 1/4, 1/4)$ . Hence  $H(X) = 1.2130$  nats ( $7/4$  bits) and  $H(Y) = 1.3863$  nats (2 bits.) Also,

$$\begin{aligned} H(X|Y) &= \sum_{i=1}^4 p(Y = i)H(X|Y = i) \\ &= \frac{1}{4}H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{2}, \frac{1}{8}, \frac{1}{8}\right) + \frac{1}{4}H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) + \frac{1}{4}H(1, 0, 0, 0) \end{aligned}$$

Using MATLAB to compute  $H(X|Y)$  we have

```
h(1)=1/4*et1_shn([1/2 1/4 1/8 1/8],1);
h(2)=1/4*et1_shn([1/4 1/2 1/8 1/8],1);
h(3)=1/4*et1_shn([1/4 1/4 1/4 1/4],1);
h(4)=1/4*et1_shn([1 0 0 0],1);
H=sum(h);
```

and  $H(X|Y) = 0.9531$  nats or

```
h(1)=1/4*et1_shn([1/2 1/4 1/8 1/8],1/log(2));
h(2)=1/4*et1_shn([1/4 1/2 1/8 1/8],1/log(2));
```

$h(3)=1/4*\text{et1\_shn}([1/4\ 1/4\ 1/4\ 1/4],1/\log(2));$   
 $h(4)=1/4*\text{et1\_shn}([1\ 0\ 0\ 0],1/\log(2));$   
 $H2=\text{sum}(h);$

$H(X|Y) = 1.375$  (11/8) bits. Similarly  $H(Y|X)=13/8$  bits.

**Remark 1:** Likewise, we can define the conditional entropy of probabilistic experiments by using similar arguments. Consider two random experiments  $A, B$  with outcomes  $\Sigma_A = \{a_1, a_2, \dots, a_m\}$  and  $\Sigma_B = \{b_1, b_2, \dots, b_n\}$  with the respective probabilities  $(p_1, p_2, \dots, p_m)$  and  $(q_1, q_2, \dots, q_n)$  which are not necessarily statistically independent. The conditional entropy of random experiment  $B$  given by the random experiment  $A$  is defined as

$$H(B|A) = \sum_{i=1}^m p(a_i)H(B|A = a_i)$$

where  $H(B|A = a_i)$  is given by the expression

$$H(B|A = a_i) = - \sum_{j=1}^n p(b_j|a_i) \ln p(b_j|a_i).$$

### 3.1 MUTUAL INFORMATION

**Definition 2:** Consider two random variables  $X$  and  $Y$  with a joint probability mass function  $p(x, y)$  and marginal distributions  $p(x)$  and  $p(y)$ . The *mutual information*  $I(X, Y)$  between  $X$  and  $Y$  is defined by the expression

$$I(X, Y) = H(X) - H(X|Y).$$

**Corollary 2:** Due to Definition 2 we distinguish the measure of uncertainty from measure of information. In general *the entropy is a quantity for a measure of uncertainty*, while *the information is a difference in uncertainty that is a difference in entropies*.

**Remark 2:** Mutual information characterizes the reduction in the uncertainty of  $X$  due to the knowledge of  $Y$ . Equivalently by symmetry (using Theorem 5) it also follows that

$$I(X, Y) = H(Y) - H(Y|X)$$

and mutual information is symmetric, that is  $I(X, Y) = I(Y, X)$ .

**Theorem 7:** We have  $I(X, Y) \geq 0$  and the equality holds if and only if  $H(Y|X) = H(Y)$  or if and only if  $X$  and  $Y$  are independent.

**Theorem 8:** We have  $I(X, Y) = H(X) + H(Y) - H(X, Y)$ .  $I(X, Y) \geq 0$  with equality if and only if the random variables  $X, Y$  are independent. We also have  $I(X, X) = H(X) - H(X|X) = H(X)$  (*self-information property*).

**Corollary 3:** The amount of information which is contributed by  $X$  about  $Y$  is always positive. *There is never a loss of information.* However, in real applications may sometimes be a loss of information. For example if the received information may lead to confusion.

**Remark 3:** *Motivation for the introduction of  $I(X, Y)$ .* In all cases one has according to Theorem 3,  $0 \leq H(Y|X) \leq H(Y)$ . As a result, the difference  $H(Y) - H(Y|X)$  measures the decrease in uncertainty that the observer has about  $Y$  knowing  $X$ . This decrease in uncertainty may be considered as an amount of information about  $Y$ .

$I(X, Y)$  can be interpreted as the amount of information about  $X$  given by the performance of the  $Y$ , that is, the amount of information contained in  $Y$  about  $X$ . It is clear that it can also be viewed as the amount of information contained in  $X$  about  $Y$ .

**Theorem 9:** The mutual information  $I(X, Y)$  is a concave function of  $p(x)$  for fixed  $p(y|x)$  and a convex function of  $p(y|x)$  for fixed  $p(x)$ .

**Remark 4:** The mutual information (or transinformation) between two probability experiments  $A$  and  $B$  is  $I(B, A) = H(B) - H(B|A)$  and all the results apply without change.

**Definition 3:** *Normalized mutual information.* The normalized mutual information  $U(X, Y)$  is defined by the expression

$$\begin{aligned} U(X, Y) &= \frac{I(X, Y)}{H(X) + H(Y)} \\ &= 2 \left[ \frac{H(Y) + H(X) - H(X, Y)}{H(X) + H(Y)} \right]. \end{aligned}$$

**Theorem 10:**  $0 \leq U(X, Y) \leq 1$ .

**Theorem 11:**  $U(X, Y)$  is a description of the symmetry. If two variables are completely independent, then  $H(X, Y) = H(X) + H(Y)$ , and  $U(X, Y)$  vanishes. If the two variables are completely dependent, then  $H(X) = H(Y) = H(X, Y)$ , and  $U(X, Y)$  equals unity.

## 3.2 CONDITIONAL MUTUAL INFORMATION

**Definition 3:** Consider three random variables  $X, Y, Z$ . The *conditional mutual information* of random variables  $X$  and  $Y$  given by  $Z$  is defined by the expression

$$I(X, Y|Z) = H(X|Z) - H(X|Y, Z).$$

**Remark 5:** This quantity indicates the reduction in the uncertainty of  $X$  due to the knowledge of  $Y$  when  $Z$  is given.

**Theorem 12:** The following relation holds:

$$0 \leq I(X, Y|Z) \leq H(X|Z)$$

with equality to 0 if and only if  $X$  and  $Y$  are conditionally independent when  $Z$  is given.

**Theorem 13:** The conditional mutual information is symmetric,  $I(X, Y|Z) = I(Y, X|Z)$ .

**Theorem 14:** (*Chain rule for conditional mutual information.*) Mutual information also satisfies a chain rule

$$I(X_1, X_2, \dots, X_n, Y) = \sum_{i=1}^n I(X_i, Y|X_{i-1}, X_{i-2}, \dots, X_1).$$

## 4 RELATIVE ENTROPY

*The relative entropy introduces a measure of the distance between two distributions.*

**Notation 1:** Consider two single probability mass functions  $p(x)$  and  $q(x)$ , with respective probabilities  $(p_1, p_2, \dots, p_m)$  and  $(q_1, q_2, \dots, q_m)$ . These probability mass functions can represent two different distributions of one random variable  $X$  with  $|\Sigma_X| = m$  (a priori and a posteriori for example) or distributions of two different random variables  $X, Y$  with  $|\Sigma_X| = |\Sigma_Y| = m$ .

**Definition 1:** The *relative entropy* (or *cross-entropy*, *Kullback-Leibler distance between two probability mass functions*  $p(x)$  and  $q(x)$ ) of the probability mass function  $p$  with respect to the probability mass function  $q$  is defined by the expression

$$\begin{aligned} D(p||q) &= \sum_{x \in \Sigma_X} p(x) \ln \frac{p(x)}{q(x)} \\ &= \sum_{i=1}^m p_i \ln \frac{p_i}{q_i}. \end{aligned}$$

**Remark 1:** In Definition 1 we use the convention that  $0 \ln \frac{0}{q} = 0$  and  $p \ln \frac{p}{0} = \infty$ . Often we use logarithm of the base 2. In this case  $D^{(2)}(p||q) = (1/\ln 2)D(p||q)$ .

**M-function:** Computes the relative entropy of a pair of probability mass functions.

```
function D=et1_rel(p,q)
[m1,n1]=size(p);
[m2,n2]=size(q);
Err=1e-6;

if abs(1-sum(p))>Err
    error('1st input must be probability mass function');
end;
if abs((1-sum(q)))>Err
    error('2nd input must be probability mass function');
end;
if ((m1~=m2) | (n1~=n2))
    error('Two inputs must have the same dimension');
end;

h1=ones(m1,n1);
h2=log((p.+Err)./(q.+Err));
D=((p.*h2)*h1');
```

**Synopsis:** `D=et1_rel(p,q)`

**Visual imagination:** By command line `et1_img('rel')` the relative entropy of distributions  $p, q$  given by  $(p_1, 1 - p_1), (q_1, 1 - q_1)$  is visualized.

**Theorem 1:** *Information inequality.* The relative entropy  $D(p||q) \geq 0$ , with equality if

and only if  $p(x) = q(x)$  for all  $x \in \Sigma_X$  or equivalently  $p_i = q_i$ ,  $i = 1, \dots, m$ .

**Theorem 2:** The relative entropy can be written as

$$D(p||q) = - \sum_{i=1}^m p_i \ln q_i - H(p).$$

$D(p||q)$  can be thought of as a deviation from  $H(p)$ . Thus  $H(p) = \ln |\Sigma_X| - D(p||u)$  where  $u$  is the uniform distribution on  $|\Sigma_X|$ .

**Remark 2:** In general  $D(p||q) \neq D(q||p)$ .

**Theorem 3:**  $D(p||q)$  is convex in the pair  $(p, q)$ , i.e. if  $(pa, qa)$  and  $(pb, qb)$  are two pairs of probability mass functions, then

$$D(\lambda pa + (1 - \lambda)pb || \lambda qa + (1 - \lambda)qb) \leq \lambda D(pa || qa) + (1 - \lambda) D(pb || qb)$$

for all  $0 \leq \lambda \leq 1$ .

**Corollary 1:** Quantity  $D$  is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality. Nonetheless, it is often useful to think of relative entropy as a ‘distance’ between distributions.

**Theorem 4:** The mutual information  $I(X, Y)$  is the relative entropy between the joint distribution  $p(x, y)$  and the product distribution  $p(x)p(y)$

$$\begin{aligned} I(X, Y) &= D(p(x, y) || p(x)p(y)) \\ &= \sum_{x \in \Sigma_X} \sum_{y \in \Sigma_Y} \ln \frac{p(x, y)}{p(x)p(y)}. \end{aligned}$$

**Definition 2:** *Conditional relative entropy*  $D(p(y|x) || q(y|x))$  is defined as the average of the relative entropies between the conditional probability mass functions  $p(y|x)$  and  $q(y|x)$  averaged over the probability mass function  $p(x)$ , i.e.

$$D(p(y|x) || q(y|x)) = \sum_{i=1}^m p(x_i) \sum_{j=1}^m p(y_j|x_i) \ln \frac{p(y_j|x_i)}{q(y_j|x_i)}.$$

**Theorem 5:** *Chain rule for relative entropy.* We have

$$D(p(x, y) || q(x, y)) = D(p(x) || q(x)) + D(p(y|x) || q(y|x)).$$

**Theorem 6:** We have  $D(p(y|x) || q(y|x)) \geq 0$  with equality if and only if  $p(y|x) = q(y|x)$  for all  $y$  and  $x$  with  $p(x) > 0$ .

## 5 WEIGHTED ENTROPY

In order to distinguish the values  $x_1, x_2, \dots, x_m$  of random variable  $X$  according to their importance with respect to a given qualitative characteristic, we can ascribe to each value  $x_k$  a weight  $w_k \geq 0$  directly proportional to its importance.

**Definition 1:** The *weighted entropy* is defined by the expression

$$H_w(X) = H_w(w_1, \dots, w_m; p_1, \dots, p_m) = - \sum_{i=1}^m w_i p_i \ln p_i.$$

**M-function:** Computes the weighted entropy of a discrete random variable  $X$ .

```
function HW=et1_wen(w,p)

[m,n]=size(p);
[mw,nw]=size(w);

Err=1e-6;

if n~=nw
    error('Two inputs must have the same length');
end;

if abs(1-sum(p))>Err
    error('2nd input must be a probability mass function');
end;

h1=(-1)*ones(m,n);
h2=((sign(p)+h1).*h1)+p;
h3=p.*w;
HW=h3*log(h2)'*(-1);
```

**Synopsis:** HW=et1\_wen(w,p)

**Visual imagination:** By MATLAB command line `et1_img('wen')` weighted entropy of  $(p_1, 1 - p_1)$  probability mass function is visualized. The weight  $w_1$  attains values 0, 0.3, 0.7, 1 at  $w_2 = 1$ .

**Theorem 1:** We have  $H_w(w_1, \dots, w_m; p_1, \dots, p_m) \geq 0$ .

**Theorem 2:** If  $w_1 = \dots = w_m = w$ , then

$$H_w(w_1, \dots, w_m; p_1, \dots, p_m) = -w \sum_{i=1}^m p_i \ln p_i = H^{(s)}(p_1, \dots, p_m)$$

with  $c = w$ .

**Theorem 3:** The following equality holds:

$$\begin{aligned} & H_w(w_1, \dots, w_{m-1}, w', w''; p_1, \dots, p_{m-1}, p', p'') = \\ & = H_w(w_1, \dots, w_m; p_1, \dots, p_m) + p_m H_w(w', w''; \frac{p'}{p_m}, \frac{p''}{p_m}) \end{aligned}$$

where

$$w_m = \frac{p'w' + p''w''}{p' + p''}, \quad p_m = p' + p''.$$

**Example 1:** Let

$$X = \begin{cases} a & \text{with probability } 1/2 \text{ and weight } 0.7, \\ b & \text{with probability } 1/4 \text{ and weight } 0.1, \\ c & \text{with probability } 1/8 \text{ and weight } 0.1, \\ d & \text{with probability } 1/8 \text{ and weight } 0.1. \end{cases}$$

The weighted entropy  $H_w(X)$  of  $X$  is

```
>>HW=et1_wen([0.7 0.1 0.1 0.1],[1/2 1/4 1/8 1/8])
>>HW =
```

0.3292

in *nats*.

**Example 2:** Let us consider the weighted entropy and put

$$w_k = -\frac{p_k}{\ln p_k} \quad k = 1, 2, \dots, m.$$

In this case, we obtain the following expression for the weighted entropy

$$H_w(w_1, \dots, w_m; p_1, \dots, p_m) = \sum_{k=1}^m p_k^2.$$

This quantity is denoted by the term *uncertainty energy*. It relates to Renyi's entropy of the order 2 (a logarithmic relation).

## 6 FURTHER ENTROPIES

*Since Shannon designed a measure of uncertainty by the entropy, an explosion of further measures (further entropies) has begun. It is given by a lot of prior requirements of what a measure of uncertainty should satisfy.*

### Renyi's Entropy

*Renyi's entropy brings an extension of Shannon's entropy to the finite discrete random variable  $X$  with an incomplete probability mass function*

$$p_i \geq 0 \quad (i = 1, \dots, m), \quad \sum_{i=1}^m p_i \leq 1.$$

**Definition 1:** The *Renyi's entropy* of order  $c$  ( $c > 0$ ,  $c \neq 1$ ) is defined by the expression

$$H_c^{(r)}(X) = \frac{1}{1-c} \ln \frac{\sum_{i=1}^m p_i^c}{\sum_{i=1}^m p_i}$$

with the convention  $0^c = 0$  for all  $c$ .

**M-function:** *Computes the Renyi's entropy of order  $c$  of a discrete random variable  $X$ .*

```
function HRc=et1_ren(p,c)
[m,n]=size(p);

if c==1
    error('Input constant must be different from one');
end;

h1=sum(p.^c);
h2=sum(p);
HRc=log(h1/h2)/(1-c);
```

**Synopsis:** HRc=et1\_ren(p,c)

**Visual imagination:** By MATLAB command line `et1_img('ren')` Renyi's entropy of  $(p_1, 1 - p_1)$  probability mass function is visualized for some selected values of  $c$ .

**Theorem 1:** Consider a complete probability mass function associated with the finite random variable  $X$ . If the order  $c$  is such that  $0 < c < 1$  then  $H_c^r(X) > H(X)$ . If the order  $c$  is such that  $c > 1$  then  $H_c^r(X) < H(X)$ . As  $c$  tends to unity,  $H_c^r(X)$  reduces to  $H(X)$ .

**Example 1:** Let

$$X = \begin{cases} a & \text{with probability } 1/2, \\ b & \text{with probability } 1/4, \\ c & \text{with probability } 1/8, \\ d & \text{with probability } 1/8. \end{cases}$$

The Renyi's entropy  $H_c^{(r)}(X)$  of order  $c = 0.7$  of  $X$  is

```
>>HRc=et1_ren([1/2 1/4 1/8 1/8],0.7)
```

```
>>HRc =
```

1.2638

or for  $c = 1.7$

```
>>HRc=et1_ren([1/2 1/4 1/8 1/8],1.7)
```

```
>>HRc =
```

1.1067.

Shannon's entropy  $H$  is 1.2130.

**Remark 1:** Several extensions of Renyi's entropy have been proposed (J.N. Kapur, R.S. Varma, M. Behara and P. Nath). For example Kapur (the best known extension obviously) generalised Renyi's entropy to give an entropy of order  $\alpha$  and type  $\beta$

$$H_{\alpha\beta}^{(k)}(X) = \frac{1}{1-\alpha} \ln \frac{\sum_{i=1}^m p_i^{\alpha+\beta-1}}{\sum_{i=1}^m p_i^\beta}$$

for  $\alpha \neq 1$ ,  $\beta > 0$ ,  $\alpha + \beta - 1 > 0$ . This reduces to Renyi's measure when  $\beta = 1$ , to Shannon's measure when  $\beta = 1$ ,  $\alpha$  tends to unity and to  $\ln m$  (so called Hartley's measure) when  $\beta = 1$  and  $\alpha = 0$ .

## Havrda and Charvat's Entropy

*Havrda and Charvat's entropy brings an additional measure of entropy of the finite discrete random variable  $X$  with a complete probability mass function*

$$p_i \geq 0 \quad (i = 1, \dots, m), \quad \sum_{i=1}^m p_i = 1.$$

**Definition 1:** The *Havrda and Charvat entropy* of order  $c$  ( $c > 0$ ,  $c \neq 1$ ) is defined by the expression

$$H_c^{(hch)}(X) = \frac{\sum_{i=1}^m p_i^c - 1}{2^{1-c} - 1}.$$

**M-function:** *Computes the Havrda and Charvat's entropy of order  $c$  of a discrete random variable  $X$ .*

```

function HHChc=et1_hch(p,c)
[m,n]=size(p);
Err=1e-6;

if abs(1-sum(p))>Err
    error('Input vector must be a probability mass function');
end;

if c==1
    error('Input constant must be different from one');
end;

h1=sum(p.^c)-1;
h2=2^(1-c)-1;
HHChc=h1/h2;

```

**Synopsis:** HHChc=et1\_hch(p,c)

**Visual imagination:** By MATLAB command line `et1_img('hch')` Havrda and Charvat's entropy of  $(p_1, 1 - p_1)$  probability mass function is visualized for some selected values of  $c$ .

**Theorem 1:** Consider the random variable  $X$  associated with a complete probability mass function. We have

$$H_c^{(r)}(X) = \frac{1}{1-c} \ln[(2^{1-c} - 1)H_c^{(hch)}(X)] + 1.$$

**Example 1:** Let

$$X = \begin{cases} a & \text{with probability } 1/2, \\ b & \text{with probability } 1/4, \\ c & \text{with probability } 1/8, \\ d & \text{with probability } 1/8. \end{cases}$$

The Havrda and Charvat's entropy  $H_c^{(hch)}(X)$  of order  $c = 0.7$  of  $X$  is

```

>>HHChc=et1_hch([1/2 1/4 1/8 1/8],0.7)
>>HHChc =

```

1.9945

or for  $c = 1.7$

```

>>HHChc=et1_hch([1/2 1/4 1/8 1/8],1.7)
>>HHChc =

```

1.4025.

Shannon's entropy  $H$  is 1.2130.

**Remark 1:** A number of other similar entropies have been proposed (M. Belis and S. Guiasu, J.N. Kapur gave about 12). For example M. Belis and S. Guiasu introduced a 'utility distribution'  $w = (w_1, w_2, \dots, w_m)$  where each  $w_i > 0$  is the utility of a value with probability  $p_i$ . Then entropy with preference is given by

$$H_w^{(bg)}(X) = \frac{\sum_{i=1}^m w_i p_i (p_i^c - 1)}{2^{1-c} - 1}.$$

## Quadratic Entropy

*Quadratic entropy was first used in theoretical physics, by Fermi for example. The term 'quadratic entropy' was introduced by I. Vajda. Consider the finite discrete random variable  $X$  with a complete probability mass function*

$$p_i \geq 0 \quad (i = 1, \dots, m), \quad \sum_{i=1}^m p_i = 1.$$

**Definition 1:** The *quadratic entropy* is defined by the expression

$$H_q(X) = \sum_{i=1}^m p_i(1 - p_i).$$

**M-function:** *Computes the quadratic entropy of a discrete random variable  $X$ .*

```
function HQ=et1_qdt(p)
[m,n]=size(p);
Err=1e-6;

if abs((1-sum(p)))>Err
    error('Input vector must be a probability mass function');
end;

h1=ones(m,n);
HQ=(p.*(h1-p))*h1';
```

**Synopsis:** HQ=et1\_qdt(p)

**Visual imagination:** By MATLAB command line `et1_img('qdt')` quadratic entropy of  $(p_1, 1 - p_1)$  probability mass function is visualized.

**Example 1:** Let

$$X = \begin{cases} a & \text{with probability } 1/2, \\ b & \text{with probability } 1/4, \\ c & \text{with probability } 1/8, \\ d & \text{with probability } 1/8. \end{cases}$$

The quadratic entropy  $H_q(X)$  of  $X$  is

```
>>HQ=et1_qdt([1/2 1/4 1/8 1/8])
>>HQ =
```

0.6563

Shannon's entropy  $H$  is 1.2130.

## Bayesian Entropy

*Shannon's measure of uncertainty is maximum when all the members of a probability mass function are equally likely. However; on the basis of intuition or experience, one may have reasons to believe that the a priori probability mass function is given by*

$$p_1 = \alpha_1, \quad p_2 = \alpha_2, \quad \dots, \quad p_m = \alpha_m; \quad \sum_{i=1}^m \alpha_i = 1$$

*then we define another measure of entropy which we call Bayesian entropy.*

**Definition 1:** The *Bayesian entropy* is defined by the expression

$$H^{(b)}(X) = - \sum_{i=1}^m p_i \ln \frac{p_i}{\alpha_i \alpha_{\min}}$$

where  $\alpha_{\min} = \min(\alpha_1, \alpha_2, \dots, \alpha_m)$  and none of the  $\alpha$ 's is zero.

**Theorem 1:** We have

$$H^b(X) = D(p||\alpha) - \ln\left(\frac{1}{\alpha_{\min}}\right),$$

where  $D(p||\alpha)$  is Kullback Leibler distance of the probability mass function  $p = (p_1, p_2, \dots, p_m)$  from the a priori probability mass function  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)$  of the associated random variable  $X$ .

**Corollary 1:** Due to the Information inequality (Theorem 1 in Section 4) the Bayesian entropy is maximum when  $p_i = \alpha_i$  for all  $i$  and is minimum when the  $X$ -value with the minimum a priori probability mass is certain to occur. Thus maximizing (minimizing) Bayesian entropy is equivalent to minimizing (maximizing) the Kullback Leibler distance between  $p$  and  $\alpha$ .

**Corollary 2:** We may have a computational recipe for Bayesian entropy as (we use  $a \equiv \alpha$ )  
>>HB=-et1\_rel(p,a)-log(1/min(a)).

**Corollary 3:** *Shannon's measure is a special case of this when the a priori probability*

*mass function is the uniform distribution.*

### *Used Literature:*

- [1 ] Guiasu S. (1976): *Information Theory with Applications*. McGraw-Hill New-York.
- [2 ] Cover T.M., Thomas J.A. (1991): *Elements of Information Theory*. John Wiley & Sons.
- [3 ] Jumarie G. (1990): *Relative Information*. Springer-Verlag Berlin Heidelberg.
- [4 ] Kapur J.N. (1989): *Maximum-Entropy Models in Science and Engineering*. John Wiley & Sons.