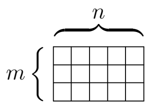
Chapter 1 Information and Uncertainty

1.1. Equal probabilities. Suppose that some situation has n possible outcomes, all of which are *equally likely*.

E.g. tossing a fair coin $(n = 2)$,
throwing a fair die $(n = 6)$,
throwing two fair dice $(n = 36)$.

Can we assign a numerical value to the *uncertainty* in this situation? If we can — call it $U(n)$ — we surely expect:

(A1) $U(n) \leqslant U(n+1)$ for all n,
(A2) $U(mn) = U(m) + U(n)$ for all m, n.



Theorem 1.1. A function $U : \mathbb{N} \to \mathbb{R}$ satisfies (A1) and (A2) if and only if

$$U(n) = C \log n \tag{1}$$

for some $C \geqslant 0$.

Proof. 'If' is obvious. To prove 'only if', assume (A1) and (A2) hold. From (A2), $U(n^r) = U(n^{r-1}) + U(n)$ and so, by induction,

$$U(n^r) = rU(n) \qquad (\forall\, n, r \in \mathbb{N}). \tag{2}$$

Thus $U(1) = rU(1)\ \forall\, r$, whence $U(1) = 0$.

Choose $C := \frac{U(2)}{\log 2}$, $\geqslant 0$ by (A1), so that (1) holds when $n = 1$ or 2. Assume $n \geqslant 3$. For any $s \in \mathbb{N}$, $\exists\ r \in \mathbb{N}$ s.t.

$$2^r \leqslant n^s \leqslant 2^{r+1}. \tag{3}$$

Then

$$U(2^r) \leqslant U(n^s) \leqslant U(2^{r+1}) \qquad \text{by (A1)}$$

and so

$$rU(2) \leqslant sU(n) \leqslant (r+1)U(2) \qquad \text{by (2).}$$

Also

$$r \log 2 \leqslant s \log n \leqslant (r+1) \log 2 \qquad \text{by (3).}$$

Thus

$$\frac{r}{s} \leqslant \frac{U(n)}{U(2)} \leqslant \frac{r+1}{s} \quad \text{and} \quad \frac{r}{s} \leqslant \frac{\log n}{\log 2} \leqslant \frac{r+1}{s},$$

and so

$$\left| \frac{U(n)}{U(2)} - \frac{\log n}{\log 2} \right| \leqslant \frac{1}{s}.$$

Since s was arbitrary, $\frac{U(n)}{U(2)} = \frac{\log n}{\log 2}$, and so $\frac{U(n)}{\log n} = \frac{U(2)}{\log 2} = C$, as required. //

Note that changing the value of C simply changes the scale of units for uncertainty. Also, changing the base of logarithms is equivalent to changing C, since $\log_a n = (\ln n)/(\ln a)$. (Proof:

$$x = \log_a n \iff n = a^x = (e^{\ln a})^x = e^{x \ln a}$$
$$\iff \ln n = x \ln a \iff x = (\ln n)/(\ln a).)$$

We *choose* to take $\log = \log_2$ and $C = 1$, and so *define* $U(n) := \log_2 n$, measured in *bits*. [Richard W. Hamming, 1915–1998; Bell labs; $\log = \log_2$?]

1.2. Unequal probabilities. Suppose now that the n outcomes have probabilities $p_1, p_2, \ldots, p_n$ $(\sum_{i=1}^n p_i = 1)$. Can we still assign a numerical value to the uncertainty? If we can — call it $H_n(p_1, \ldots, p_n)$ — we expect:

(B1) $H_n(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) = U(n) = \log_2 n$.
(B2) $H_n(p_1, \ldots, p_n)$ is a continuous function of $p_1, \ldots, p_n$.
(B3) If $p_1 + \ldots + p_r = p > 0$ and $q_1 + \ldots + q_s = q > 0$ and $p + q = 1$, then $H_{r+s}(p_1, \ldots, p_r, q_1, \ldots, q_s) =$

$$H_2(p, q) + pH_r(\tfrac{p_1}{p}, \ldots, \tfrac{p_r}{p}) + qH_s(\tfrac{q_1}{q}, \ldots, \tfrac{q_s}{q}).$$

Note that, if $\sum_{i=1}^r p_i = p$, then

$$-p \log_2 p - p \sum_{i=1}^r \frac{p_i}{p} \log_2 \frac{p_i}{p} = -\sum_{i=1}^r \left(p_i \log_2 p + p_i \log_2 \frac{p_i}{p}\right)$$
$$= -\sum_{i=1}^r p_i \log_2 p_i. \tag{4}$$

Theorem 1.2. A set of functions H_n $(n = 1, 2, \ldots)$ satisfies (B1), (B2) and (B3) if and only if [$H_n : [0,1]^n \to \mathbb{R}$? No!]

$$H_n(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (\geqslant 0).$$

Proof. 'If': (B1) holds because $-\sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n} = \frac{n}{n} \log_2 n = \log_2 n = U(n)$. (B2) is obvious. And (B3) holds because

$$\begin{aligned} -p\log_2 p - q\log_2 q - p\sum_{i=1}^{r} \tfrac{p_i}{p} \log_2 \tfrac{p_i}{p} - q\sum_{j=1}^{s} \tfrac{q_j}{q} \log_2 \tfrac{q_j}{q} \\ = -\sum_{i=1}^{r} p_i \log_2 p_i - \sum_{j=1}^{s} q_j \log_2 q_j \qquad \text{by (4).} \end{aligned}$$

'Only if': We prove the result by induction on n. It is true if $n = 1$ since $H_1(1) = \log_2 1 = 0$ by (B1). Suppose next that $n = 2$, and suppose first that p_1, p_2 are rational, say $p_1 = p = \frac{r}{t}$, $p_2 = q = \frac{s}{t}$ where $r + s = t$. By (B3),

$$H_t(\tfrac{1}{t}, \ldots, \tfrac{1}{t}) = H_2(\tfrac{r}{t}, \tfrac{s}{t}) + \tfrac{r}{t} H_r(\tfrac{1}{r}, \ldots, \tfrac{1}{r}) + \tfrac{s}{t} H_s(\tfrac{1}{s}, \ldots, \tfrac{1}{s})$$

and so, by (B1),

$$H_2(\tfrac{r}{t}, \tfrac{s}{t}) = \log_2 t - \tfrac{r}{t}\log_2 r - \tfrac{s}{t}\log_2 s = -\tfrac{r}{t}\log_2 \tfrac{r}{t} - \tfrac{s}{t}\log_2 \tfrac{s}{t}$$

since $\log_2 t = (\frac{r}{t} + \frac{s}{t}) \log_2 t$. Thus

$$H_2(p_1, p_2) = -p_1 \log_2 p_1 - p_2 \log_2 p_2$$

whenever p_1 and p_2 are rational. By continuity (B2), this holds even if p_1 and p_2 are irrational.

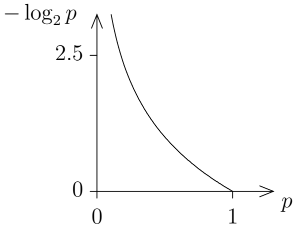
Finally, let $n \geqslant 3$ and apply (B3) with $r = n - 1$, $s = 1$, $p = \sum_{i=1}^{n-1} p_i$ and $q = p_n$. Then

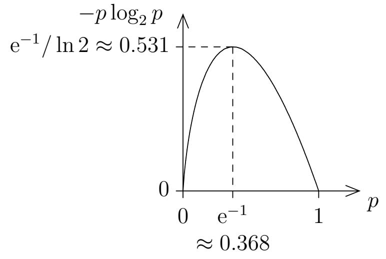
$$\begin{aligned} H_n(p_1, \ldots, p_n) &= H_2(p, q) + pH_{n-1}(\tfrac{p_1}{p}, \ldots, \tfrac{p_{n-1}}{p}) + qH_1(1) \\ &= -p\log_2 p - q\log_2 q - p\sum_{i=1}^{n-1} \tfrac{p_i}{p}\log_2 \tfrac{p_i}{p} + 0 \\ &\qquad \text{by induction} \end{aligned}$$

$$\begin{aligned} &= -\sum_{i=1}^{n-1} p_i \log_2 p_i - q\log_2 q \qquad \text{by (4)} \\ &= -\sum_{i=1}^{n} p_i \log_2 p_i \end{aligned}$$

as required. //

Note that $-p\log_2 p \to 0$ as $p \to 0+$ (because it is $\frac{\log_2 \frac{1}{p}}{\frac{1}{p}} \to 0$ as $\frac{1}{p} \to \infty$). So we shall interpret $-0\log_2 0$ as 0 and feel free to write $-p\log_2 p$ even when p may be 0.





1.3. Definitions and properties. If X is a random variable that takes n values with probabilities $p_1, \ldots, p_n$ $(\sum_{i=1}^{n} p_i = 1)$, or if X is a finite probability space with probability distribution $(p_1, \ldots, p_n)$, then we define the *uncertainty* or *entropy* of X, measured in *bits*, to be

$$H(X) := H_n(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

(Claude E. Shannon, 1948). (1916–2001; Bell Labs $\leqslant$ 1957, then MIT.)

The *information content* of an event x with probability $p > 0$ is defined to be $I(x) = I(p) := -\log_2 p$. Thus $H(X) = \sum_{i=1}^{n} p_i I(p_i)$ is the average or expected value of the information content of X.

Lemma 1.3.1. If $(p_1, \ldots, p_n)$ and $(q_1, \ldots, q_n)$ are probability distributions, then

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leqslant -\sum_{i=1}^{n} p_i \log_2 q_i,$$

with equality iff $p_i = q_i$ for each i.

Proof. Since $e^x \geqslant 1 + x$, with equality iff $x = 0$,
therefore $y \geqslant 1 + \ln y$ $(y > 0)$, with equality iff $y = 1$.
Hence $\ln \frac{q_i}{p_i} \leqslant \frac{q_i}{p_i} - 1$, with equality iff $p_i = q_i$.
Thus

$$\sum_{i=1}^n p_i \ln q_i - \sum_{i=1}^n p_i \ln p_i = \sum_{i=1}^n p_i \ln \frac{q_i}{p_i}$$
$$\leqslant \sum_{i=1}^n p_i\left(\frac{q_i}{p_i} - 1\right) = 1 - 1 = 0,$$

with equality iff $p_i = q_i$ for each i. Dividing by $\ln 2$ gives the result. //

Theorem 1.3. (a) $H_n(p_1, \ldots, p_n) \leqslant \log_2 n$, with equality iff $p_i = \frac{1}{n}\ \forall\, i$.
(b) $H_n(p_1, \ldots, p_n) \geqslant 0$, with equality iff $p_i = 1$ for some i.
(c) $H_{n+1}(p_1, \ldots, p_n, 0) = H_n(p_1, \ldots, p_n)$.

Proof. (a) By Lemma 1.3.1 with $q_i = \frac{1}{n}$ for each i,

$$H_n(p_1, \ldots, p_n) = -\sum_{i=1}^n p_i \log_2 p_i \leqslant -\sum_{i=1}^n p_i \log_2 \frac{1}{n} = \log_2 n,$$

with equality iff $p_i = \frac{1}{n}$ for each i.
(b) $-\sum_{i=1}^n p_i \log_2 p_i \geqslant 0$, with equality iff, for each i, either $p_i = 0$ or $\log_2 p_i = 0$, i.e., $p_i = 1$. But $\sum p_i = 1$, so that $p_i = 1$ for exactly one i.
(c) Obvious in view of our convention that $0 \log_2 0 = 0$. //

If X and Y are random variables, each taking finitely many values, write (X, Y) for the random variable that takes value (x, y) iff $X = x$ and $Y = y$. Suppose

$$X = \begin{pmatrix} x_1 \ldots x_n \\ p_1 \ldots p_n \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \ldots y_m \\ q_1 \ldots q_m \end{pmatrix}, \quad (X,Y) = \begin{pmatrix} (x_1, y_1) \ldots (x_n, y_m) \\ r_{11} \quad \ldots \quad r_{nm} \end{pmatrix}$$

(that is, $X = x_i$ with probability p_i, etc.). We say that X and Y are *independent* if $r_{ij} = p_i q_j$ for all i and j. We write $H(X, Y)$ for $H((X, Y))$.

Theorem 1.4. $H(X, Y) \leqslant H(X) + H(Y)$, with equality iff X, Y are independent.

Proof. Note that

$$p_i = \sum_{j=1}^m r_{ij}, \qquad q_j = \sum_{i=1}^n r_{ij} \tag{5}$$

whether or not X and Y are independent. Also,

$$\sum_{i=1}^n \sum_{j=1}^m p_i q_j = \Big(\sum_{i=1}^n p_i\Big)\Big(\sum_{j=1}^m q_j\Big) = 1.$$

Hence

$$\begin{aligned} H(X) + H(Y) &= -\sum_{i=1}^n p_i \log_2 p_i - \sum_{j=1}^m q_j \log_2 q_j \\ &= -\sum_{i=1}^n \sum_{j=1}^m (r_{ij} \log_2 p_i + r_{ij} \log_2 q_j) \\ &= -\sum_{i=1}^n \sum_{j=1}^m r_{ij} \log_2 p_i q_j \\ &\geqslant -\sum_{i=1}^n \sum_{j=1}^m r_{ij} \log_2 r_{ij} = H(X, Y), \end{aligned}$$

by Lemma 1.3.1, with equality iff $r_{ij} = p_i q_j$ for all i and j. //

With X, Y as before and fixed y_j, write $X|y_j$ for the random variable that takes value $x_i|y_j$ with probability $P(x_i|y_j) = \frac{P(x_i, y_j)}{P(y_j)} = \frac{r_{ij}}{q_j}$. (Note that $\sum_{i=1}^n \frac{r_{ij}}{q_j} = 1$ by (5).) By our previous definitions, therefore,

$$I(x_i|y_j) = -\log_2 \frac{r_{ij}}{q_j}$$

and

$$H(X|y_j) = \sum_{i=1}^{n} P(x_i|y_j) I(x_i|y_j) = -\sum_{i=1}^{n} \tfrac{r_{ij}}{q_j} \log_2 \tfrac{r_{ij}}{q_j}.$$

The *conditional uncertainty* or *conditional entropy* of X given Y is defined to be

$$\begin{aligned} H(X|Y) &:= \sum_{j=1}^{m} q_j H(X|y_j) && (6) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \log_2 \tfrac{r_{ij}}{q_j} && (7) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} I(x_i|y_j), \end{aligned}$$

which is the average or expected value of $H(X|y_j)$ over all y_j, or (in a sense) of $I(x_i|y_j)$ over all x_i and y_j.

Theorem 1.5. (a) $H(X|Y) = H(X,Y) - H(Y)$.
(b) $H(X|X) = 0$.
(c) $H(X|Y) \geqslant 0$, with equality iff X is uniquely determined by Y.
(d) $H(X|Y) \leqslant H(X)$, with equality iff X and Y are independent.

Proof. (a)

$$\begin{aligned} H(X|Y) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \log_2 \tfrac{r_{ij}}{q_j} \qquad \text{by (7)} \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \log_2 r_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \log_2 q_j \\ &= H(X,Y) - H(Y) \end{aligned}$$

since $\sum_{i=1}^{n} r_{ij} = q_j$ by (5).

(b) If $Y = X$ then $y_i = x_i$ and $r_{ij} = P(y_i, y_j) = \begin{cases} q_j & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$
Thus $r_{ij} \log_2 \frac{r_{ij}}{q_j} = 0$ for every term contributing to $H(X|X)$.

(c) $H(X|Y) \geqslant 0$ by (7), and $H(X|Y) = 0$ iff $r_{ij} = 0$ or q_j for each i and j. By (5), this means that, for each j, $r_{ij} = q_j$ for exactly one

i, say $i = i(j)$, which means that if $Y = y_j$ then $X = x_{i(j)}$. Thus X is uniquely determined by Y. The converse follows similarly.

(d) By (a) and Theorem 1.4,

$$H(X|Y) = H(X,Y) - H(Y) \leqslant H(X) + H(Y) - H(Y) = H(X),$$

with equality iff X and Y are independent. //

The *information about* X *given by* y_j is

$$I(X|y_j) := H(X) - H(X|y_j) \qquad \text{(can be negative!).} \qquad (8)$$

The *information about* X *given by* Y is

$$I(X|Y) := \sum_{j=1}^{m} q_j I(X|y_j) = H(X) - H(X|Y) \qquad (9)$$

by (6) and (8), which is the expected amount of uncertainty in X that is removed by Y.

Example. Three horses are entered for a race. Their probabilities of winning are $\frac{7}{8}$, $\frac{1}{16}$, $\frac{1}{16}$. The uncertainty as to the result is

$$\begin{aligned} H(X) = H_3(\tfrac{7}{8}, \tfrac{1}{16}, \tfrac{1}{16}) &= -\tfrac{7}{8} \log_2 \tfrac{7}{8} - \tfrac{2}{16} \log_2 \tfrac{1}{16} \\ &\approx 0.169 + 0.5 \\ &= 0.669. \end{aligned}$$

I tell you that the favourite has broken its leg and will not run. If the probability of this is $2^{-14} \approx \frac{1}{16000}$, then I have given you $-\log_2 2^{-14} = 14$ bits of information. But the uncertainty in the result of the race is now

$$H(X|y) = H_2(\tfrac{1}{2}, \tfrac{1}{2}) = U(2) = 1,$$

and so I have given you $0.669 - 1 = -0.331$ bits of information *about the result of the race.*

Theorem 1.6. (a) $I(X|Y) = H(X) + H(Y) - H(X,Y) = I(Y|X)$.
(b) $I(X|X) = H(X)$.
(c) $I(X|Y) \leqslant H(X)$, with equality iff X is uniquely determined by Y.
(d) $I(X|Y) \geqslant 0$, with equality iff X and Y are independent.
(e) $I(X|Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \log_2 \frac{r_{ij}}{p_i q_j}$.

Proof.(a) By (9) and Theorem 1.5(a),

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

as required. $I(Y|X) = I(X|Y)$ because clearly $H(Y,X) = H(X,Y)$.

(b), (c) and (d) follow immediately from the corresponding parts of Theorem 1.5, since $I(X|Y) + H(X|Y) = H(X)$.

(e)
$$\begin{aligned} I(X|Y) &= H(X) - H(X|Y) && \text{by (9)} \\ &= -\sum_{i=1}^{n} p_i \log_2 p_i + \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \log_2 \frac{r_{ij}}{q_j} && \text{by (7)} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \left[-\log_2 p_i + \log_2 \frac{r_{ij}}{q_j} \right] && \text{by (5)} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} r_{ij} \log_2 \frac{r_{ij}}{p_i q_j}. \quad // \end{aligned}$$