Chapter 5
A Few Elements of Quantitative Information Theory

1. The Amount of Information

The attempt to construct a quantitative theory in which the concepts of "production" and "transmission" of information are meaningful faces two main difficulties. The definition of a mathematical model in which such concepts are well defined and the assignment of a measure to the amount of information involved. Hereafter the basic elements of such a theory will be reviewed in a way which naturally lends itself to application in physics.

It is assumed that one receives information whenever one "learns" of an event whose occurrence was previously uncertain. Moreover, the more likely an event is, the less information is conveyed by the knowledge of its actual occurrence. Let x represent [the occurrence of] an event, and $\bar{x}$ its complement (i.e. its non-occurrence), and let p_x, $p_{\bar{x}}$, where

$$p_x + p_{\bar{x}} = 1 \tag{1.1}$$

denote the probabilities of two such events. Furthermore, let I_x denote the amount of information conveyed by the knowledge of the occurrence of x. Assuming that x is specified only by its probability p_x, let us define I_x to be a non-negative function I of p_x, defined over the range $0 < p_x \leq 1$,

$$I_x = I(p_x) \tag{1.2}$$

(notice that $p_x = 0$ is meaningless in the present context). Since the probability of receiving the amount of information I_x is p_x — and that of receiving $I_{\bar{x}}$ is $p_{\bar{x}}$ — the expected amount of information received is

$$H(x,\bar{x}) = p_x I(p_x) + p_{\bar{x}} I(p_{\bar{x}}) \quad . \tag{1.3}$$

Of course the symbolism can be easily generalized to more complex cases: if $\{x_1,\ldots,x_n\}$ are a set of mutually exclusive events of probabilities $\{p_{x_1},\ldots,p_{x_n}\}$ respectively, such that

$$p_{x_1} + \ldots + p_{x_n} = 1 \tag{1.4}$$

then the average amount of information conveyed by the knowledge of which x_i actually occurred is assumed to be

$$H(x_1,\ldots,x_n) = p_{x_1} I(p_{x_1}) + \ldots + p_{x_n} I(p_{x_n}) \tag{1.5}$$

(notice that if $p_{x_j} = 0$ for some j, the event x_j should simply be omitted from consideration).

It is interesting that this intuitive type of reasoning leads to strong constraints upon the form of $I(\ )$. For simplicity let us begin with $n = 3$ (and set $x_1 = x$, $x_2 = y$, $x_3 = z$). In order to determine which among the events x, y, z actually occurred, it is e.g. sufficient to determine whether or not x occurred, and in the case where it didn't, to determine which of y, z did occur. The amount of information conveyed by the first determination is evidently

$$H(x,\bar{x}) = p_x I(p_x) + (1 - p_x) I(1 - p_x) \quad . \tag{1.6}$$

If x didn't occur, then the conditional probabilities of y and z are respectively $p_y/p_{\bar{x}}$, $p_z/p_{\bar{x}}$. The amount of information conveyed by the second determination is therefore given by

$$H(y,z|\bar{x}) = \frac{p_y}{p_{\bar{x}}} I\left(\frac{p_y}{p_{\bar{x}}}\right) + \frac{p_z}{p_{\bar{x}}} I\left(\frac{p_z}{p_{\bar{x}}}\right) \quad . \tag{1.7}$$

However this latter amount of information is conveyed only when the event $\bar{x}$ occurs (i.e. x does not occur) so the total amount of information conveyed on the average, by the two determinations is

$$H(x,y,z) = H(x,\bar{x}) + p_{\bar{x}} H(y,z|\bar{x}) \quad . \tag{1.8}$$

The requirement to be imposed on the function $I(\ )$ is that the relation (1.8) be satisfied for all allowable values of p_x, p_y, p_z greater than zero. Explicitly (1.8) writes, expressing the H-functions in terms of probabilities,

$$H(p_x,p_y,p_z) = H(p_x,1-p_x) + (1-p_x)\, H\left(\frac{p_y}{1-p_x}, \frac{p_z}{1-p_x}\right) \quad . \tag{1.9}$$

Now the identical reasoning applies if x is considered to be a composite event, namely if x is assumed to consist of $(n-1)$ mutually exclusive events $u_1,\ldots,u_{n-1}$, whose probabilities are denoted, $p_1,\ldots,p_{n-1}$. Writing $q_1 \equiv p_y$, $q_2 \equiv p_z$, and $p_n \equiv p_{\bar{x}}$, $(p_n = q_1 + q_2)$, (1.9) reads in this case

$$H(p_1,\ldots,p_{n-1},q_1,q_2) = H(p_1,\ldots,p_n) + p_n H\left(\frac{q_1}{p_n}, \frac{q_2}{p_n}\right) \quad . \tag{1.10}$$

Condition (1.10) is very strong; indeed it will practically suffice to determine the form of $H(p_1,\ldots,p_n)$ without regard to its definition in terms of $I(\ )$. The latter will only impose a further condition suggested by the fact that terms of the form $p_i I(p_i)$ should be dropped when $p_i = 0$, namely that $H(p_1,\ldots,p_n)$ be defined even when some of the p_i's vanish but that it be continuous in the domain

$$\mathcal{D}: \{p_i \geq 0 \ ; \ i = 1,\ldots,n \mid \sum_{i=1}^{n} p_i = 1\} \quad . \tag{1.11}$$

The function $H(p_1,\ldots,p_n)$ is completely determined up to a multiplicative constant — which fixes the size of the unit of information — by the only additional requirement that it must be a symmetric function of all its variables.

In order to explicitly construct H, let us begin with a few remarks. First, writing (1.9) with $p_x = p_y = 1/2$, $p_z = 0$, one has

$$H\left(\frac{1}{2},\frac{1}{2},0\right) = H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{1}{2}H(1,0) \quad . \tag{1.12}$$

Using now the hypothesis of complete symmetry, one has

$$H\left(\frac{1}{2},\frac{1}{2},0\right) = H\left(0,\frac{1}{2},\frac{1}{2}\right) \tag{1.13}$$

and finally using (1.9) again with $p_x = 0$, $p_y = p_z = 1/2$,

$$H\left(0,\frac{1}{2},\frac{1}{2}\right) = H(0,1) + H\left(\frac{1}{2},\frac{1}{2}\right) \quad . \tag{1.14}$$

Inserting (1.12) and (1.14) into (1.13) and recalling that one should have $H(0,1) = H(1,0)$, (1.13) implies

$$H(1,0) = 0 \quad . \tag{1.15}$$

Equation (1.15) in turn implies that

$$H(p_1,\ldots,p_n,0) = H(p_1,\ldots,p_n) \tag{1.16}$$

as one can straightforwardly derive from (1.10), setting in it $q_2 = 0$ (and hence $q_1 = p_n$). One can now by induction extend (1.10) to ($m \geq 2$)

$$H(p_1,\ldots,p_{n-1},q_1,\ldots,q_m) = H(p_1,\ldots,p_n) + p_n H\left(\frac{q_1}{p_n},\ldots,\frac{q_m}{p_n}\right) \tag{1.17}$$

where

$$p_n = q_1 + \ldots + q_m \quad . \tag{1.18}$$

Notice that (1.17) indeed coincides with (1.10) for $m = 2$. From (1.16), it is clear that one has to consider only the case when $q_i > 0$, $i = 1,\ldots,m$. Suppose there is an m such that (1.17), (1.18) are true for all n, then, setting

$$p' = q_2 + \ldots + q_{m+1} \tag{1.19}$$

and using (1.10) as well one can write

$$\begin{aligned} &H(p_1,\ldots,p_{n-1}, q_1,\ldots,q_{m+1}) = \\ &= H(p_1,\ldots,p_{n-1}, q_1,p') + p'H\left(\frac{q_2}{p'}, \ldots, \frac{q_{m+1}}{p'}\right) = \\ &= H(p_1,\ldots,p_n) + p_nH\left(\frac{q_1}{p_n}, \frac{p'}{p_n}\right) + p'H\left(\frac{q_2}{p'}, \ldots, \frac{q_{m+1}}{p'}\right) \quad . \end{aligned} \tag{1.20}$$

But, once more for the induction hypothesis (to be thought of for $n = 2$),

$$H\left(\frac{q_1}{p_n}, \ldots, \frac{q_{m+1}}{p_n}\right) = H\left(\frac{q_1}{p_n}, \frac{p'}{p_n}\right) + \frac{p'}{p_n} H\left(\frac{q_2}{p'}, \ldots, \frac{q_{m+1}}{p'}\right) \tag{1.21}$$

i.e.

$$p_nH\left(\frac{q_1}{p_n}, \frac{p'}{p_n}\right) + p'H\left(\frac{q_2}{p'}, \ldots, \frac{q_{m+1}}{p'}\right) = p_nH\left(\frac{q_1}{p_n}, \ldots, \frac{q_{m+1}}{p_n}\right) \tag{1.22}$$

which, inserted at the right-hand side of (1.20) gives the assertion of (1.17) for $m+1$.

It is now easy matter to prove that if

$$p_i = \sum_{j=1}^{m_i} q_{i,j} \quad , \qquad i = 1,\ldots,n \tag{1.23}$$

then

$$H(q_{1,1},\ldots,q_{1,m_1},\ldots,q_{n,1},\ldots,q_{n,m_n}) =$$

$$= H(p_1,\ldots,p_n) + \sum_{i=1}^{n} p_i H\left(\frac{q_{i,1}}{p_i},\ldots,\frac{q_{i,m_i}}{p_i}\right) . \tag{1.24}$$

In fact, using (1.17), one has

$$H(q_{1,1},\ldots,q_{n,m_n}) = p_n H\left(\frac{q_{n,1}}{p_n},\ldots,\frac{q_{n,m_n}}{p_n}\right) =$$

$$+ H(q_{1,1},\ldots,q_{n-1,m-1},p_n) \quad . \tag{1.25}$$

After shifting p_n in the last factor to the extreme left by the assumed symmetry of H, the reduction process can be repeated on it. After n steps the result (1.24) is obtained. Let us now set

$$F(n) \doteq H\left(\frac{1}{n},\ldots,\frac{1}{n}\right) \quad , \qquad n \geq 2 \tag{1.26}$$

and $F(1) = 0$. Applying (1.24) to the case $m_1 = \ldots = m_n = m$, $q_{ij} = 1/mn$, $\forall\, i,j = 1,\ldots,n$, one has

$$F(mn) = F(m) + F(n) \quad . \tag{1.27}$$

Applying further (1.17) one obtains

$$H\left(\frac{1}{n},\ldots,\frac{1}{n}\right) = H\left(\frac{1}{n},\frac{n-1}{n}\right) + \frac{n-1}{n} H\left(\frac{1}{n-1},\ldots,\frac{1}{n-1}\right) \tag{1.28}$$

from which

$$\eta_n \doteq H\left(\frac{1}{n}, \frac{n-1}{n}\right) = F(n) - \frac{n-1}{n} F(n-1) \quad . \tag{1.29}$$

Now from the continuity of $H(p,1-p)$ there follows that as $n \to \infty$, $\eta_n \to H(0,1) = 0$ (see (1.15)). Moreover the recursion relation

$$n\,\eta_n = nF(n) - (n-1)F(n-1) \tag{1.30}$$

implies

$$nF(n) = \sum_{k=1}^{n} k \cdot \eta_k \tag{1.31}$$

or

$$\frac{F(n)}{n} = \frac{1}{n^2} \sum_{k=1}^{n} k \cdot \eta_k = \frac{n+1}{2n} \frac{2}{n(n+1)} \sum_{k=1}^{n} k \cdot \eta_k \quad . \tag{1.32}$$

But $\frac{2}{n(n+1)} \sum_{k=1}^{n} k \cdot \eta_k$ is simply the arithmetic mean of the first $\frac{n(n+1)}{2}$ terms of the sequence $\eta_1, \eta_2, \eta_2, \eta_3, \eta_3, \eta_3, \eta_4, \eta_4, \eta_4, \eta_4, \ldots$ whose limit, as it was shown above is zero. Thus as $n \to \infty$, $\frac{2}{n(n+1)} \sum_{k=1}^{n} k \cdot \eta_k \to 0$, from which follows that also $\frac{F(n)}{n} \to 0$.

Finally let

$$\lambda_n \equiv F(n) - F(n-1) = \eta_n - \frac{1}{n} F(n-1) \tag{1.33}$$

$\lambda_n \to 0$ as $n \to \infty$, as well. We have thus all the ingredients needed to determine the form of $F(n)$. It is clear from (1.27) that we only need to know the value of $F(n)$ for n prime. Indeed, for arbitrary n, let

$$n = p_1^{\alpha_1} \ldots p_s^{\alpha_s} \quad , \quad s \geq 1 \quad , \quad \alpha_s \in \mathbb{Z}_+ \tag{1.34}$$

be the prime factorization of n. Repeated application of (1.27) gives then

$$F(n) = \alpha_1 F(p_1) + \ldots + \alpha_s F(p_s) \quad . \tag{1.35}$$

We now put, for all prime p

$$F(p) = c_p \ln p \tag{1.36}$$

so that (1.35) reads

$$F(n) = \alpha_1 c_{p_1} \ln p_1 + \ldots + \alpha_s c_{p_s} \ln p_s \quad . \tag{1.37}$$

It is straightforward to show that the sequence $\{c_p, p = \text{prime}\}$ contains a largest member. In fact, if the contrary were true, it would be possible to construct an infinite sequence of primes $p_1 < p_2 < p_3 < \ldots$ with $p_1 = 2$ such that p_{i+1} were the first prime greater than p_i for which $c_{p_{i+1}} > c_{p_i}$. There would follow from this construction that if q is a prime less than p_i, then $c_q < c_{p_i}$. Now, for $i > 1$, let $p_i - 1 = q_1^{\beta_1} \ldots q_r^{\beta_r}$, $r \geq 1$, $\beta_r \in \mathbb{Z}_+$ be the prime factorization of $(p_i - 1)$, and consider

$$\begin{aligned} \lambda_{p_i} &= F(p_i) - F(p_i - 1) = \\ &= F(p_i) - \frac{F(p_i)}{\ln p_i} \ln(p_i - 1) + c_{p_i} \ln(p_i - 1) - F(p_i - 1) = \\ &= \frac{F(p_i)}{p_i} \frac{p_i}{\ln p_i} \ln \frac{p_i}{p_i - 1} + \sum_{j=1}^{r} \beta_j (c_{p_i} - c_{q_j}) \ln q_j \quad . \end{aligned} \tag{1.38}$$

Since $(p_i - 1)$ is necessarily even, one of the q_j must take on the value 2. Moreover since $p_i > q_j$, $j = 1, \ldots, r$, by the hypothesis we should have

be the prime factorization of n. Repeated application of (1.27) gives then

$$F(n) = \alpha_1 F(p_1) + \ldots + \alpha_s F(p_s) \quad . \tag{1.35}$$

We now put, for all prime p

$$F(p) = c_p \ell n\, p \tag{1.36}$$

so that (1.35) reads

$$F(n) = \alpha_1 c_{p_1} \ell n\, p_1 + \ldots + \alpha_s c_{p_s} \ell n\, p_s \quad . \tag{1.37}$$

It is straightforward to show that the sequence $\{c_p,\ p = \text{prime}\}$ contains a largest member. In fact, if the contrary were true, it would be possible to construct an infinite sequence of primes $p_1 < p_2 < p_3 < \ldots$ with $p_1 = 2$ such that p_{i+1} were the first prime greater than p_i for which $c_{p_{i+1}} > c_{p_i}$. There would follow from this construction that if q is a prime less than p_i, then $c_q < c_{p_i}$. Now, for $i > 1$, let $p_i - 1 = q_1^{\beta_1} \ldots q_r^{\beta_r}$, $r \geq 1$, $\beta_r \in \mathbb{Z}_+$ be the prime factorization of $(p_i - 1)$, and consider

$$\lambda_{p_i} = F(p_i) - F(p_i - 1) =$$

$$= F(p_i) - \frac{F(p_i)}{\ell n p_i} \ell n(p_i - 1) + c_{p_i} \ell n(p_i - 1) - F(p_i - 1) =$$

$$= \frac{F(p_i)}{p_i} \frac{p_i}{\ell n p_i} \ell n \frac{p_i}{p_i - 1} + \sum_{j=1}^{r} \beta_j (c_{p_i} - c_{q_j}) \ell n q_j \quad . \tag{1.38}$$

Since $(p_i - 1)$ is necessarily even, one of the q_j must take on the value 2. Moreover since $p_i > q_j$, $j = 1, \ldots, r$, by the hypothesis we should have

$$\sum_{j=1}^{r} \beta_j (c_{p_i} - c_{q_j}) \ell n q_j \geq (c_{p_i} - c_2) \ell n 2 \geq (c_{p_2} - c_2) \ell n 2 \quad . \tag{1.39}$$

However, as $i \to \infty$, by the properties proved above, both λ_{p_i} and $\frac{F(p_i)}{p_i}$ tend to zero $\left(\text{and so does } \frac{p_i}{\ell n p_i} \ell n \frac{p_i}{p_i - 1}\right)$. Therefore, by (1.38), (1.39) it should be $(c_{p_2} - c_2) \ell n 2 \leq 0$ or $c_{p_2} \leq c_2$, which contradicts the definition of p_2. In precisely the same manner one shows that the sequence $\{c_p, p = \text{prime}\}$ contains a smallest member. Suppose now that there is at least a prime $\bar{p}$ such that $c_{\bar{p}} > c_2$, and let p_0 be that prime for which c_{p_0} is a maximum. Of course then $c_{p_0} > c_2$. Let m be a positive integer ≥ 1 and let $\bar{q}_1^{\gamma_1} \ldots \bar{q}_t^{\gamma_t}$ be the prime factorization of $p_0^m - 1$. From (1.27) it follows that

$$\frac{F(p_0^m)}{\ell n\, p_0^m} = c_{p_0} \quad . \tag{1.40}$$

Then we can repeat all the steps leading to (1.38) and (1.39) (β_j being replaced by γ_j, r by t, q_j by $\bar{q}_j$ and p_i by p_0^m) obtaining

$$\lambda_{p_0^m} \geq \frac{F(p_0^m)}{p_0^m} \frac{p_0^m}{\ell n\, p_0^m} \ell n \frac{p_0^m}{p_0^m - 1} + (c_{p_0} - c_2) \ell n 2 \quad . \tag{1.41}$$

Letting $m \to \infty$, one gets $(c_{p_0} - c_2) \ell n 2 \leq 0$, which contradicts $c_{p_0} > c_2$. In precisely the same manner, one shows the non-existence of any prime q_0 for which $c_{q_0} < c_2$. Thus, all the c_p's are equal and (from (1.34), (1.37))

$$F(n) = c\, \ell n\, n \tag{1.42}$$

where c is a constant (equal) to the common value of the c_p's).

Let finally

$$p = \frac{r}{s} \quad , \quad r,s \in \mathbb{Z}_+ \quad , \quad s \geq r \quad . \tag{1.43}$$

By (1.24) one can write

$$H\left(\frac{1}{s}, \ldots, \frac{1}{s}\right) = H\left(\frac{r}{s}, \frac{s-r}{s}\right) + \frac{r}{s} H\left(\frac{1}{r}, \ldots, \frac{1}{r}\right) +$$

$$+ \frac{s-r}{s} H\left(\frac{1}{s-r}, \ldots, \frac{1}{s-r}\right) \tag{1.44}$$

from which

$$H(p, 1-p) = F(s) - pF(r) - (1-p)F(s-r) =$$

$$= c \ln s - pc \ln r - (1-p)c \ln (s-r) =$$

$$= c\left(p \ln \frac{s}{r} + (1-p) \ln \frac{s}{s-r}\right) =$$

$$= c\left(p \ln \frac{1}{p} + (1-p) \ln \frac{1}{(1-p)}\right) \quad . \tag{1.45}$$

By continuity this extends to all irrational p's, and using (1.10) inductively on n, we have

$$H(p_1,\ldots,p_n) = - c \sum_{i=1}^{n} p_i \ln p_i \quad . \tag{1.46}$$

Notice that $H(p_1,\ldots,p_n)$ in (1.46) is just of the form (1.5),

$$H(p_1,\ldots,p_n) = \sum_{i=1}^{n} p_i \, I(p_i) \quad . \tag{1.47}$$

This suggests that c must be taken > 0, so that

$$I(p) = - c \ln p \tag{1.48}$$

is a monotone increasing function of $\bar{p} = 1 - p$.

2. Basic Properties of the H Function

Let X be an abstract set, consisting of a finite number of elements x. Let $p(\ )$ be a probability function defined over X, i.e. $p(Q)$ is a non-negative number defined for each subset Q of X, with the properties that:

$$p(X) = 1 \tag{2.1}$$

and

$$p(Q_1 \cup Q_2) = p(Q_1) + p(Q_2) \tag{2.2}$$

if Q_1 and Q_2 are disjoint subsets of X. The totality of objects $\{(X,x),p(\ )\}$ is what was defined as a finite probability space. Recalling the discussion of previous section, any finite probability space can be considered an information source. By obvious extension of (1.46), we will define as information content of such a source the non-negative quantity

$$H(X) = - \sum_{x \in X} p(x) \ln p(x) \tag{2.3}$$

(where units chosen in such a way that $c = 1$).

Let now (X,x) and (Y,y) be two finite abstract spaces, and denote by $X \otimes Y$ the finite abstract space consisting of all pairs (x,y), and by $p(\ ,\ )$ a probability distribution over $X \otimes Y$. The information content of this source is obviously

$$H(X \otimes Y) = - \sum_{x \in X} \sum_{y \in Y} p(x,y) \ln p(x,y) \quad . \tag{2.4}$$

However in this case the distribution $p(\ ,\ )$ gives rise also to a distribution