CHAPTER 11

DISCRETE PRIOR PROBABILITIES – THE ENTROPY PRINCIPLE

At this point we return to the job of designing this robot. We have part of its brain designed, and we have seen how it would reason in a few simple problems of hypothesis testing and estimation. In every problem it has solved thus far, the results have either amounted to the same thing as, or were usually demonstrably superior to, those offered in the "orthodox" statistical literature. But it is still not a very versatile reasoning machine, because it has only one means by which it can translate raw information into numerical values of probabilities; the principle of indifference (2–74). Consistency requires it to recognize the relevance of prior information, and so in almost every problem it is faced at the onset with the problem of assigning initial probabilities, whether they are called technically prior probabilities or sampling probabilities. It can use indifference for this if it can break the situation up into mutually exclusive, exhaustive possibilities in such a way that no one of them is preferred to any other by the evidence. But often there will be prior information that does not change the set of possibilities but does give a reason for preferring one possibility to another. What do we do in this case?

Orthodoxy evades this problem by simply ignoring prior information for fixed parameters, and maintaining the fiction that sampling probabilities are known frequencies. Yet in some forty years of active work in this field, the writer has never seen a real problem in which one actually has prior information about sampling frequencies! In practice, sampling probabilities are always assigned from some standard theoretical model (binomial distribution, *etc.*) which starts from the principle of indifference. If the robot is to rise above such false pretenses, we must give it more principles for assigning initial probabilities by logical analysis of the prior information. In this Chapter and the following one we introduce two new principles of this kind, each of which has an unlimited range of useful applications. But the field is open–ended in all directions; we expect that more principles will be found in the future, leading to a still wider range of applications.

A New Kind of Prior Information.

Imagine a class of problems in which the robot's prior information consists of average values of certain things. Suppose, for example, that statistics were collected in a recent earthquake and that out of 100 windows broken, there were 976 pieces found. But we are not given the numbers 100, 976; we are told only that "The average window is broken into $\overline{m} = 9.76$ pieces." That is the way it would be reported. Given only that information, what is the probability that a window would be broken into exactly m pieces? There is nothing in the theory so far that will answer that question.

As another example, suppose we have a table which we cover with black cloth, and some dice, but for reasons that will be clear in a minute, they are black dice with white spots. A die is tossed onto the black table. Above there is a camera. Every time it is tossed, we take a snapshot. The camera will record only the white spots. Now we don't change the film in between, so we end up with a multiple exposure; uniform blackening of the film after we have done this a few thousand times. From the known density of the film and the number of tosses, we can infer the average number of spots which were on top, but not the frequencies with which various faces came up. Suppose that the average number of spots turned out to be 4.5 instead of the 3.5 that we might expect from an honest die. Given only this information (*i.e.*, not making use of anything else that you or I might know about dice except that they have six faces), what estimates should the robot make of the frequencies with which n spots came up? Supposing that successive tosses form an exchangeable sequence as defined in Chapter 3, what probability should it assign to the n'th face coming up on the next toss?

As a third example, suppose that we have a string of $N = 1{,}000$ cars, bumper to bumper, and they occupy the full length of $L = 3$ miles. As they drive onto a rather large ferry boat, the distance that it sinks into the water determines their total weight W. But the numbers N, L, W are withheld from us; we are told only their average length L/N and average weight W/N. We can look up statistics from the manufacturers, and find out how long the Volkswagen is, how heavy it is; how long a Cadillac is, and how heavy it is, and so on, for all the other brands. From knowledge only of the average length and the average weight of these cars, what can we then infer about the proportion of cars of each make that were in the cluster?

If we knew the numbers N, L, W, then this could be solved by direct application of Bayes' theorem; without that information we could still introduce the unknowns N, L, W as nuisance parameters and use Bayes' theorem, eliminating them at the end. We shall give an example of this procedure in the nonconglomerability problem in Chapter 15. The Bayesian solution is not really wrong, and for large N it would be for all practical purposes the same as the solution advocated below; but it would be tedious for three nuisance parameters and it would not really address our problem; it only transfers it to the problem of assigning priors to N, L, W, leaving us back in essentially the same situation. Is there a better procedure that will go directly to the real problem?

Now, it is not at all obvious how our robot should handle problems of this sort. Actually, we have defined two different problems; *estimating* a frequency distribution, and *assigning* a probability distribution. But in an exchangeable sequence these are almost identical mathematically. So let's think about how we would want the robot to behave in this situation. Of course, we want it to take into account fully all the information it has, of whatever kind. But we would not want it to jump to conclusions that are not warranted by the evidence it has. We have seen that a uniform probability assignment represents a state of mind completely noncommittal with regard to all possibilities; it favors no one over any other, and thus leaves the entire decision to the subsequent data which the robot may receive. The knowledge of average values does give the robot a reason for preferring some possibilities to others, but we would like it to assign a probability distribution which is, in some sense, as uniform as it can get while agreeing with the available information. The most conservative, noncommittal distribution is the one which is in some sense as "spread–out" as possible. In particular, the robot must not ignore any possibility – it must not assign zero probability to any situation unless its information really rules out that situation.

This sounds very much like defining a variational problem; the information available defines constraints fixing some properties of the initial probability distribution, but not all of them. The ambiguity remaining is to be resolved by the policy of honesty; frankly acknowledging the full extent of its ignorance by taking into account all possibilities allowed by its knowledge.† To cast it into mathematical form, the aim of avoiding unwarranted conclusions leads us to ask whether there is some reasonable numerical measure of how uniform a probability distribution is, which the robot could maximize subject to constraints which represent its available information. Let's approach this in the way most problems are solved; the time–honored method of trial and error. We just have to invent some measures of uncertainty, and put them to the test to see what they give us.

One measure of how broad an initial distribution is would be its variance. Would it make sense if the robot were to assign probabilities so as to maximize the variance subject to its information? But consider the distribution of maximum variance for a given $\overline{m}$, if the conceivable values of m are unlimited, as in the broken window problem. Then the maximum variance solution would be the one where the robot assigns a very large probability for no breakage at all, and an enormously small

† This is really an ancient principle of wisdom, recognized clearly already in such sources as Herodotus and the Old Testament.

probability of a window to be broken into billions and billions of pieces. You can get an arbitrarily high variance this way, while keeping the average at 9.76. In the dice problem, the solution with maximum variance would be to assign all the probability to the one and the six, in such a way that $p_1 + 6p_6 = 4.5$, or $p_1 = 0.3$, $p_6 = 0.7$. So that, evidently, is not the way we would want our robot to behave; it would be jumping to wildly unjustified conclusions, since nothing in its information says that it is impossible to have three spots up.

Minimum $\sum p_i^2$.

Another kind of measure of how spread out a probability distribution is, which has been used a great deal in statistics, is the sum of the squares of the probabilities assigned to each of the possibilities. The distribution which minimizes this expression, subject to constraints represented by average values, might be a reasonable way for our robot to behave. Let's see what sort of a solution this would lead to. We want to make

$$\sum_m p_m^2$$

a minimum, subject to the constraints that the sum of all p_m shall be unity, and the average over the distribution is $\overline{m}$. A formal solution is obtained at once from the variational problem

$$\delta\left[\sum_m p_m^2 - \lambda\sum_m m p_m - \mu\sum_m p_m\right] = \sum_m (2p_m - \lambda m - \mu)\delta p_m = 0 \qquad (11\text{–}1)$$

where λ and μ are Lagrange Multipliers. So p_m will be a linear function of m: $2p_m - \lambda m - \mu = 0$. Then μ and λ are found from

$$\sum_m p_m = 1, \qquad \sum_m m p_m = \overline{m}, \qquad (11\text{–}2)$$

where $\overline{m}$ is the average value of m, given to us in the statement of the problem..

Suppose that m can take on only the values $1, 2,$ and 3. Then the formal solution is

$$p_1 = \frac{4}{3} - \frac{\overline{m}}{2}, \qquad p_2 = \frac{1}{3}, \qquad p_3 = \frac{\overline{m}}{2} - \frac{2}{3}. \qquad (11\text{–}3)$$

This would be at least usable for some values of $\overline{m}$. But in principle, $\overline{m}$ could be anywhere in $1 \le \overline{m} \le 3$, and p_1 becomes negative when $\overline{m} > 8/3 = 2.667$, while p_3 becomes negative when $\overline{m} < 4/3 = 1.333$. The formal solution for minimum $\sum p_i^2$ lacks the property of nonnegativity. We might try to patch this up in an *ad hoc* way by replacing the negative values by zero and adjusting the other probabilities to keep the constraint satisfied. But then the robot is using different principles of reasoning in different ranges of $\overline{m}$; and it is still assigning zero probability to situations that are not ruled out by its information. This performance is not acceptable; it is an improvement over maximum variance, but the robot is still behaving inconsistently and jumping to unwarranted conclusions. We have taken the trouble to examine this criterion because some writers have rejected the entropy solution given next and suggested on intuitive grounds, without examining the actual results, that minimum $\sum p_i^2$ would be a more reasonable criterion.

But the idea behind the variational approach still looks like a good one. There should be some consistent measure of the uniformity, or "amount of uncertainty" of a probability distribution which we can maximize, subject to constraints, and which will have the property that forces the robot to be completely honest about what it knows, and in particular it does not permit the robot to draw any conclusions unless those conclusions are really justified by the evidence it has. But we must pay more attention to the consistency requirement.

Entropy: Shannon's Theorem.

At this stage we turn to the most quoted theorem in Shannon's work on Information Theory (Shannon, 1948; Shannon & Weaver, 1949). If there exists a consistent measure of the "amount of uncertainty" represented by a probability distribution, there are certain conditions it will have to satisfy. We shall state them in a way which will remind you of the arguments we gave in Chapter 2; in fact, this is really a continuation of the basic development of probability theory:

(1) We assume that some numerical measure $H_n(p_1, p_2, \ldots, p_n)$ exists; *i.e.*, that it is possible to set up some kind of association between "amount of uncertainty" and real numbers.

(2) We assume a continuity property: H_n is a continuous function of the p_i. For otherwise an arbitrarily small change in the probability distribution would still lead to the same big change in the amount of uncertainty.

(3) We require that this measure should correspond qualitatively to common sense in that when there are many possibilities, we are more uncertain than when there are few. This condition takes the form that in case the p_i are all equal, the quantity

$$h(n) = H_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$

is a monotonic increasing function of n. This establishes the "sense of direction."

(4) We require that the measure H_n be consistent in the same sense as before; *i.e.* if there is more than one way of working out its value, we must get the same answer for every possible way.

Previously, our conditions of consistency took the form of the functional equations (2–4), (2–30). Now we have instead a hierarchy of functional equations relating the different H_n to each other. Suppose the robot perceives two alternatives, to which it assigns probabilities p_1 and $q \equiv 1 - p_1$. Then the "amount of uncertainty" represented by this distribution is $H_2(p_1, q)$. But now the robot learns that the second alternative really consists of two possibilities, and it assigns probabilities p_2, p_3 to them, satisfying $p_2 + p_3 = q$. What is now his full uncertainty $H_3(p_1, p_2, p_3)$ as to all three possibilities? Well, the process of choosing one of the three can be broken down into two steps. First, decide whether the first possibility is or is not true; the uncertainty removed by this decision is the original $H_2(p_1, q)$. Then, with probability q he encounters an additional uncertainty as to events 2, 3, leading to

$$H_3(p_1, p_2, p_3) = H_2(p_1, q) + qH_2\left(\frac{p_2}{q}, \frac{p_3}{q}\right) \qquad (11\text{–}4)$$

as the condition that we shall obtain the same net uncertainty for either method of calculation. In general, a function H_n can be broken down in many different ways, relating it to the lower order functions by a large number of equations like this.

Note that equation (11–4) says rather more than our previous functional equations did. It says not only that the H_n are consistent in the aforementioned sense, but also that they are to be additive. So this is really an additional assumption which we should have included in our list.

Exercise 11.1 It seems intuitively that the most general condition of consistency would be a functional equation which is satisfied by any monotonic increasing function of H_n. But this is ambiguous unless we say something about how the monotonic functions for different n are to be related; is it possible to invoke the same function for all n? Carry out some new research in this field by investigating this matter; try either to find a possible form of the new functional equations, or to explain why this cannot be done.

At any rate, the next step is perfectly straightforward mathematics; let's see the full proof of Shannon's theorem, now dropping the unnecessary subscript on H_n.

First, we find the most general form of the composition law (11–4) for the case that there are n mutually exclusive propositions $(A_1, \cdots, A_n)$ to consider, to which we assign probabilities $(p_1, \cdots, p_n)$ respectively. Instead of giving the probabilities of the $(A_1, \cdots, A_n)$ directly, we might first group the fist k of them together as the proposition denoted by $(A_1 + A_2 + \cdots + A_k)$ in Boolean algebra, and give its probability which by (2–64) is equal to $w_1 = (p_1 + \cdots + p_k)$; then the next m propositions are combined into $(A_{k+1} + \cdots + A_{k+m})$, for which we give the probability $w_2 = (p_{k+1} + \cdots + p_{k+m})$, *etc.* When this much has been specified, the amount of uncertainty as to the composite propositions is $H(w_1, \ldots, w_r)$.

Next we give the conditional probabilities $(p_1/w_1, \ldots, p_k/w_1)$ of the propositions $(A_1, \ldots, A_k)$, given that the composite proposition $(A_1 + \cdots + A_k)$ is true. The additional uncertainty, encountered with probability w_1, is then $H(p_1/w_1, \cdots, p_k/w_k)$. Carrying this out for the composite propositions $(A_{k+1} + \cdots + A_{k+m})$, *etc.*, we arrive ultimately at the same state of knowledge as if the $(p_1, \ldots, p_n)$ had been given directly; so consistency requires that these calculations yield the same ultimate uncertainty no matter how the choices were broken down in this way. Thus we have

$$\begin{aligned} H(p_1 \ldots p_n) = H(w_1 \ldots w_r) &+ w_1 H(p_1/w_1, \ldots, p_k/w_1) \\ &+ w_2 H(p_{k+1}/w_2, \ldots, p_{k+m}/w_2) + \cdots \end{aligned} \tag{11–5}$$

which is the general form of the functional equation (11–4). For example,

$$H(1/2, 1/3, 1/6) = H(1/2, 1/2) + (1/2)H(2/3, 1/3)\,.$$

Since $H(p_1, \ldots, p_n)$ is to be continuous, it will suffice to determine it for all rational values

$$p_i = \frac{n_i}{\sum n_i} \tag{11–6}$$

with n_i integers. But then (11–5) determines the function H already in terms of the quantities $h(n) \equiv H(1/n, \ldots, 1/n)$ which measure the "amount of uncertainty" for the case of n equally likely alternatives. For we can regard a choice of one of the alternatives $(A_1, \cdots, A_n)$ as the first step in the choice of one of

$$\sum_{i=1}^{n} n_i$$

equally likely alternatives in the manner just described, the second step of which is also a choice between n_i equally likely alternatives. As an example, with $n = 3$, we might choose $n_1 = 3$, $n_2 = 4$, $n_3 = 2$. For this case the composition law (11–5) becomes

$$h(9) = H\left(\frac{3}{9}, \frac{4}{9}, \frac{2}{9}\right) + \frac{3}{9}h(3) + \frac{4}{9}h(4) + \frac{2}{9}h(2)$$

For a general choice of the n_i, (11–5) reduces to

$$h(\sum n_i) = H(p_1, \ldots, p_n) + \sum_i p_i h(n_i) \tag{11–7}$$

Now we can choose all $n_i = m$; whereupon (11–7) collapses to

$$h(mn) = h(m) + h(n)\,. \tag{11–8}$$

Evidently, this is solved by setting

$$h(n) = K \log n \tag{11–9}$$

where K is a constant. But is this solution unique? If m, n were continuous variables, this would be easy to answer; differentiate with respect to m, set $m = 1$, and integrate the resulting differential equation with the initial condition $h(1) = 0$ evident from (11–8), and you have proved that (11–9) is the only solution. But in our case, (11–8) need hold only for integer values of m, n; and this elevates the problem from a trivial one of analysis to an interesting little exercise in number theory.

First, note that (11–9) is no longer unique; in fact, (11–8) has an infinite number of solutions for integer m, n. For, each positive integer N has a unique decomposition into prime factors; and so by repeated application of (11–8) we can express $h(N)$ in the form $\sum_i m_i h(q_i)$ where q_i are the prime numbers and m_i non-negative integers. Thus we can specify $h(q_i)$ *arbitrarily* for the prime numbers q_i, whereupon (11–8) is just sufficient to determine $h(N)$ for all positive integers.

To get any unique solution for $h(n)$, we have to add our qualitative requirement that $h(n)$ be monotonic increasing in n. To show this, note first that (11–8) ma;y be extended by induction:

$$h(nmr \cdots) = h(n) + h(m) + h(r) + \cdots$$

and setting the factors equal in the k'th order extension gives

$$h(n^k) = kh(n) \tag{11–10}$$

Now let t, s be any two integers not less than 2. Then for arbitrarily large n, we can find an integer m such that

$$\frac{m}{n} \le \frac{\log t}{\log s} < \frac{m+1}{n}, \qquad \text{or} \qquad s^m \le t^n < s^{m+1}\,. \tag{11–11}$$

Since h is monotonic increasing, $h(s^m) \le h(t^n) \le h(s^{m+1})$; or from (11–10),

$$mh(s) \le nh(t) \le (m+1)h(s)$$

which can be written as

$$\frac{m}{n} \le \frac{h(t)}{h(s)} \le \frac{m+1}{n}\,. \tag{11–12}$$

Comparing (11–11), (11–12), we see that

$$\left|\frac{h(t)}{h(s)} - \frac{\log t}{\log s}\right| \le \frac{1}{n}, \qquad \text{or} \qquad \left|\frac{h(t)}{\log t} - \frac{h(s)}{\log s}\right| \le \epsilon \tag{11–13}$$

where

$$\epsilon \equiv \frac{h(s)}{n \log t}$$

is arbitrarily small. Thus $h(t)/\log t$ must be a constant, and the uniqueness of (11–9) is proved.

Now different choices of K in (11–9) amount to the same thing as taking logarithms to different bases; so if we leave the base arbitrary for the moment, we can just as well write $h(n) = \log n$. Substituting this into (11–7), we have Shannon's theorem: The only function $H(p_1, \ldots, p_n)$ satisfying the conditions we have imposed on a reasonable measure of "amount of uncertainty" is

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i \tag{11–14}$$

Accepting this interpretation, it follows that the distribution $(p_1 \cdots p_n)$ which maximizes (11–14) subject to constraints imposed by the available information, will represent the "most honest" description of what the robot knows about the propositions $(A_1, \ldots, A_n)$. The only arbitrariness is that we have the option of taking the logarithm to any base we please, corresponding to a multiplicative constant in H. This, of course, has no effect on the values of $(p_1, \ldots, p_n)$ which maximize H.

As in Chapter 2, we note the logic of what has and has not been proved. We have shown that use of the measure (11–14) is a *necessary* condition for consistency; but in accordance with Gödel's theorem one cannot prove that it actually is consistent unless we move out into some as yet unknown region beyond that used in our proof. From the above argument, given originally in Jaynes (1957a) and leaning heavily on Shannon, we conjectured that any other choice of "information measure" will lead to inconsistencies if carried far enough; and a direct proof of this was found subsequently by Shore & Johnson (1980) using an argument entirely independent of ours. Many years of use of the Maximum Entropy Principle (variously abbreviated to PME, MEM, MENT, MAXENT by various writers) has not revealed any inconsistency; and of course we do not believe that one will ever be found.

The function H is called the *entropy*; or better the *information entropy* of the distribution $\{p_i\}$. This is an unfortunate terminology which now seems impossible to correct. We must warn at the outset that the major occupational disease of this field is a persistent failure to distinguish between the *information entropy*, which is a property of any probability distribution, and the *experimental entropy* of thermodynamics, which is instead a property of a thermodynamic state as defined, for example by such observed quantities as pressure, volume, temperature, magnetization, of some physical system. They should never have been called by the same name; the experimental entropy makes no reference to any probability distribution, and the information entropy makes no reference to thermodynamics.† Many textbooks and research papers are flawed fatally by the author's failure to distinguish between these entirely different things; and in consequence proving nonsense theorems.

We have seen the mathematical expression $\Sigma p \log p$ appearing incidentally in several previous Chapters, generally in connection with the multinomial distribution; now it has acquired a new meaning as a fundamental measure of how uniform a probability distribution is.

† But in case the problem happens to be one of thermodynamics, there is a relation between them, which we shall find presently.

Exercise 11.2 Prove that any change in the direction of equalizing two probabilities will increase the information entropy. That is, if $p_i < p_j$, then the change $p_i \to p_i + \epsilon$, $p_j \to p_j - \epsilon$ where ϵ is infinitesimal and positive, will increase $H(p_1 \cdots p_n)$ by an amount proportional to ϵ. Applying this repeatedly, it follows that the maximum attainable entropy is one for which all the differences $|p_i - p_j|$ are as small as possible. This shows also that information entropy is a *global* property, not a local one; a difference $|p_i - p_j|$ has just as great an effect on entropy whether $|i - j|$ is 1 or 1000.

Although the above demonstration appears satisfactory mathematically, it is not yet in completely satisfactory form conceptually. The functional equation (11–4) does not seem quite so intuitively compelling as our previous ones were. In this case, the trouble is probably that we have not yet learned how to verbalize the argument leading to (11–4) in a fully convincing manner. Perhaps this will inspire you to try your hand at improving the verbiage that we used just before writing (11–4). Then it is comforting to know that there are several other possible arguments, like the aforementioned one of Shore & Johnson, which also lead uniquely to the same conclusion (11–14). We note another of them.

The Wallis Derivation.

This resulted from a suggestion made to the writer in 1962 by Dr. Graham Wallis (although the argument to follow differs slightly from his). We are given information I, which is to be used in assigning probabilities $\{p_1 \cdots p_m\}$ to m different possibilities. We have a total amount of probability

$$\sum_{i=1}^{m} p_i = 1$$

to allocate among them. Now in judging the reasonableness of any particular allocation we are limited to a consideration of I and the rules of probability theory; for to call upon any other evidence would be to admit that we had not used all the available information in the first place.

The problem can also be stated as follows. Choose some integer $n \gg m$, and imagine that we have n little "quanta" of probability, each of magnitude $\delta = n^{-1}$, to distribute in any way we see fit. In order to ensure that we have "fair" allocation, in the sense that none of the m possibilities shall knowingly be given either more or fewer of these quanta than it "deserves," in the light of the information I, we might proceed as follows.

Suppose we were to scatter these quanta at random among the m choices – you can make this a blindfolded penny–pitching game into m equal boxes if you like. If we simply toss these "quanta" of probability at random, so that each box has an equal probability of getting them, nobody can claim that any box is being unfairly favored over any other. If we do this, and the first box receives exactly n_1 quanta, and the second n_2, *etc.*, we will say that the random experiment has generated the probability assignment

$$p_i = n_i \delta = n_i / n, \qquad i = 1, 2, \ldots, m\,.$$

The probability that this will happen is the multinomial distribution

$$m^{-n} \frac{n!}{n_1! \cdots n_m!}\,. \qquad (11\text{–}15)$$

Now imagine that a blindfolded friend repeatedly scatters the n quanta at random among the m boxes. Each time he does this we examine the resulting probability assignment. If it happens to conform to the information I, we accept it; otherwise we reject it and tell him to try again. We continue until some probability assignment $\{p_1, \ldots, p_m\}$ is accepted.

What is the most likely probability distribution to result from this game? From (11–15) it is the one which maximizes

$$W = \frac{n!}{n_1! \cdots n_m!} \tag{11–16}$$

subject to whatever constraints are imposed by the information I. We can refine this procedure by choosing smaller quanta; *i.e.* large n. In this limit we have, by the Stirling approximation

$$\log n! = n \log n - n + \sqrt{2\pi n} + \frac{1}{12n} + O(\frac{1}{n^2}) \tag{11–17}$$

where $O(1/n^2)$ denotes terms that tend to zero as $n \to \infty$, as $(1/n^2)$ or faster. Using this result, and writing $n_i = np_i$, we find easily that as $n \to \infty$, $n_i \to \infty$, in such a way that $n_i/n \to p_i =$ const.,

$$\frac{1}{n} \log W \to -\sum_{i=1}^{m} p_i \log p_i = H(p_1, \ldots, p_m) \tag{11–18}$$

and so, the *most likely* probability assignment to result from this game, is just the one that has maximum entropy subject to the given information I.

You might object that this game is still not entirely "fair," because we have stopped at the first acceptable result without seeing what other acceptable ones might also have turned up. In order to remove this objection, we can consider all possible acceptable distributions and choose the average $\overline{p_i}$ of them. But here the "laws of large numbers" come to our rescue. We leave it as an exercise for the reader to prove that in the limit of large n, *the overwhelming majority of all acceptable probability allocations that can be produced in this game are arbitrarily close to the maximum–entropy distribution.*†

From a conceptual standpoint, the Wallis derivation is quite attractive. It is entirely independent of Shannon's functional equations (11–5), it does not require any postulates about connections between probability and frequency; nor does it suppose that the different possibilities $\{1, \ldots, m\}$ are themselves the result of any repeatable random experiment. Furthermore, it leads automatically to the prescription that H is to be *maximized* – and not treated in some other way – without the need for any quasi–philosophical interpretation of H in terms of such a vague notion as "amount of uncertainty." Anyone who accepts the proposed game as a fair way to allocate probabilities that are not determined by the prior information, is thereby led inexorably to the Maximum Entropy Principle.

Let us stress this point. It is a big mistake to try to read too much philosophical significance into theorems which lead to equation (11–14). In particular, the association of the word "information" with entropy expressions seems in retrospect quite unfortunate, because it persists in carrying the wrong connotations to so many people. Shannon himself, with prophetic insight into the reception his work would get, tried to play it down by pointing out immediately after stating the theorem, that it was in no way necessary for the theory to follow. By this he meant that the inequalities which H satisfies are already quite sufficient to justify its use; it does not really need the further support of the theorem which deduces it from functional equations expressing intuitively the properties of "amount of uncertainty."

† This result is formalized more completely in the Entropy Concentration Theorem given later.

However, while granting that this is perfectly true, we would like now to show that *if we do* accept the expression for entropy, very literally, as *the* correct expression for the "amount of uncertainty" represented by a probability distribution, this will lead us to a much more unified picture of probability theory in general. It will enable us to see that the principle of indifference, and many frequency connections of probability are special cases of a single principle, and that statistical mechanics, communication theory, and a mass of other applications are all instances of a single method of reasoning.

An Example.

First, let's test this principle by seeing how it would work out in the example discussed above, in which m can take on only the values 1, 2, 3, and $\overline{m}$ is given. We can use our Lagrange multiplier argument again to solve this problem; as in (11–1),

$$\delta \left[H(p_1, \ldots, p_3) - \lambda \sum_{m=1}^{3} m p_m - \mu \sum_{m=1}^{3} p_m \right] = \sum_{m=1}^{3} \left[\frac{\partial H}{\partial p_m} - \lambda m - \mu \right] \delta p_m = 0. \qquad (11\text{–}19)$$

Now,

$$\frac{\partial H}{\partial p_m} = -\log p_m - 1 \qquad (11\text{–}20)$$

so our solution is

$$p_m = \mathrm{e}^{-\lambda_0 - \lambda m} \qquad (11\text{–}21)$$

where $\lambda_0 \equiv \mu + 1$.

So the distribution which has maximum entropy, subject to a given average value, will be in exponential form, and we have to fit the constants λ_0 and λ by forcing this to agree with the constraints that the sum of the p's must be one and the average value must be equal to the average $\overline{m}$ that we assigned. This is accomplished quite neatly if you define a function

$$Z(\lambda) \equiv \sum_{m=1}^{3} e^{-\lambda m} \qquad (11\text{–}22)$$

which we called the *partition function* in Chapter 9. The equations (11–2) which fix our Lagrange multipliers then take the form

$$\lambda_0 = \log Z(\lambda)\,, \qquad \overline{m} = -\frac{\partial}{\partial \lambda} \log Z(\lambda)\,. \qquad (11\text{–}23)$$

We find easily that $p_1(\overline{m})$, $p_2(\overline{m})$, $p_3(\overline{m})$ are given in parametric form by

$$p_k = \frac{e^{-k\lambda}}{e^{-\lambda} + e^{-2\lambda} + e^{-3\lambda}} = \frac{e^{(3-k)\lambda}}{e^{2\lambda} + e^{\lambda} + 1}\,, \qquad k = 1, 2, 3. \qquad (11\text{–}24)$$

$$\overline{m} = \frac{e^{2\lambda} + 2e^{\lambda} + 3}{e^{2\lambda} + e^{\lambda} + 1}. \qquad (11\text{–}25)$$

In a more complicated problem we would just have to leave it in parametric form, but in this particular case we can eliminate the parameter λ algebraically, leading to the explicit solution

$$\begin{aligned} p_1 &= \frac{3 - \overline{m} - p_2}{2} \\ p_2 &= \frac{1}{3}\left[\sqrt{4 - 3(\overline{m} - 2)^2} - 1\right] \\ p_3 &= \frac{\overline{m} - 1 - p_2}{2} \end{aligned} \tag{11–26}$$

As a function of $\overline{m}$, p_2 is the arc of an ellipse which comes in with unit slope at the end points. p_1 and p_3 are also arcs of ellipses, but slanted one way and the other.

We have finally arrived here at a solution which meets the objections we had to the first two criteria. The maximum entropy distribution (11–24) has automatically the property $p_k \geq 0$ because the logarithm has a singularity at zero which we could never get past. It has, furthermore, the property that it never allows the robot to assign zero probability to any possibility unless the evidence forces that probability to be zero.‡ The only place where a probability goes to zero is in the limit where $\overline{m}$ is exactly one or exactly three. But of course, in those limits, some probabilities did have to be zero by deductive reasoning, whatever principle we invoked.

Generalization: A More Rigorous Proof.

The maximum–entropy solution can be generalized in many ways. Suppose a variable x can take on n different discrete values $(x_1, \ldots, x_n)$, which correspond to the n different propositions $(A_1, \ldots, A_n)$ above; and that there are m different functions of x

$$f_k(x), \qquad 1 \leq k \leq m < n, \tag{11–27}$$

and the constraints are that we want them to have expectations, $\langle f_k(x) \rangle = F_k, \quad 1 \leq k \leq m$, where the $\{F_k\}$ are numbers given to us in the statement of the problem. What probabilities $(p_1, \ldots, p_n)$ will the robot assign to the possibilities $(x_1, \ldots, x_n)$? We shall have

$$F_k = \langle f_k(x) \rangle = \sum_{i=1}^{n} p_i f_k(x_i), \tag{11–28}$$

and to find the set of p_i's which has maximum entropy subject to all these constraints simultaneously, we just have to introduce as many Lagrange multipliers as there are constraints imposed on the problem

$$\begin{aligned} &\delta\Big[H(p_1 \cdots p_n) - (\lambda_0 - 1)\sum_i p_i - \lambda_1 \sum_i p_i f_1(x_i) - \cdots - \lambda_m \sum_i p_i f_m(x_i)\Big] \\ &\qquad = \sum_i \left[\frac{\partial H}{\partial p_i} - (\lambda_0 - 1) - \lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)\right] \delta p_i = 0 \end{aligned}$$

and so from (11–19) our solution is the following:

$$p_i = e^{-\lambda_0 - \lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)}, \tag{11–29}$$

as always, exponential in the constraints. To evaluate the λ's, the sum of all probabilities will have to be unity:

‡ This property was stressed by Dr. David Blackwell, who considered it the most fundamental requirement of a rational procedure for assigning probabilities.

$$1 = \sum_i p_i = e^{-\lambda_0} \sum_i \exp\left[-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)\right] \tag{11–30}$$

If we now define a partition function as

$$Z(\lambda_1, \cdots, \lambda_m) \equiv \sum_{i=1}^{n} \exp\left[-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)\right] \tag{11–31}$$

then (11–30) reduces to

$$\lambda_0 = \log Z(\lambda_1, \ldots, \lambda_m)\,. \tag{11–32}$$

The average value (11–28) of $f_k(x)$ is then

$$F_k = e^{-\lambda_0} \sum_i f_k(x_i) \exp\left[-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)\right]$$

or,

$$F_k = -\frac{\partial}{\partial \lambda_k} \log Z \tag{11–33}$$

What is the maximum value of the entropy that we get from this probability distribution?

$$H_{max} = \left[-\sum_{i=1}^{n} p_i \log p_i\right]_{max} \tag{11–34}$$

From (11–29) we find that

$$H_{max} = \lambda_0 + \lambda_1 F_1 + \cdots + \lambda_m F_m\,. \tag{11–35}$$

Now these results open up so many new applications that it is important to have as rigorous a proof as possible. But to solve a maximization problem by variational means, as we just did, isn't 100 percent rigorous. Our Lagrange multiplier argument has the nice feature that it gives you the answer instantaneously. It has the bad feature that after you've done it, you're not quite sure it *is* the answer. Suppose we wanted to locate the maximum of a function whose absolute maximum happened to occur at a cusp (discontinuity of slope) instead at a rounded top. If we state it as a variational problem, it will locate any subsidiary rounded maxima, but it will not find the cusp. Even after we've proved that we have the highest value that can be reached by variational methods, it is possible that the function reaches a still higher value at some cusp that we can't locate by variational methods. There would always be a little grain of doubt remaining if we do only the variational problem.

So, now we give an entirely different derivation which is strong just where the variational argument is weak. For this we need a lemma. Let p_i be any set of numbers which could be a possible probability distribution; in other words,

$$\sum_{i=1}^{n} p_i = 1, \qquad p_i \geq 0 \tag{11–36}$$

and let u_i be another possible probability distribution,

$$\sum_{i=1}^{n} u_i = 1, \qquad u_i \geq 0\,. \tag{11–37}$$

Now

$$\log x \le (x-1), \qquad 0 \le x < \infty \tag{11–38}$$

with equality if and only if $x = 1$. Therefore,

$$\sum_{i=1}^{n} p_i \log \frac{u_i}{p_i} \le \sum_{i=1}^{n} p_i \left(\frac{u_i}{p_i} - 1 \right) = 0$$

or,

$$H(p_1, \ldots, p_n) \le \sum_{i=1}^{n} p_i \log \left(\frac{1}{u_i} \right) \tag{11–39}$$

with equality if and only if $p_i = u_i$, $i = 1, 2, \ldots, n$. This is the lemma we need.

Now we simply pull a distribution u_i out of the hat;

$$u_i \equiv \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \exp\{-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)\}. \tag{11–40}$$

where $Z(\lambda_1, \ldots, \lambda_m)$ is defined by (11–31). Never mind why we chose u_i this particular way; we'll see why in a minute. We can now write the inequality (11–39) as

$$H \le \sum_{i=1}^{n} p_i [\log Z + \lambda_1 f_1(x_i) + \cdots + \lambda_m f_m(x_i)]$$

or

$$H \le \log Z + \lambda_1 \langle f_1(x) \rangle + \cdots + \lambda_m \langle f_m(x) \rangle. \tag{11–41}$$

Now, let the p_i vary over the class of all possible probability distributions that satisfy the constraints (11–28) of the problem. The right–hand side of (11–41) stays constant. Our lemma now says that H attains its absolute maximum H_{max}, making (11–41) an equality, if and only if the p_i are chosen as the canonical distribution (11–40).

This is the rigorous proof, which is independent of the things that might happen if you try to do it as a variational problem. This argument is, as we see, strong just where the variational argument is weak. On the other hand, this argument is weak where the variational argument is strong, because we just had to pull the answer out of a hat in writing (11–40). We had to know the answer before we could prove it. If you have both arguments side by side, then you have the whole story.

Formal Properties of Maximum-Entropy Distributions.

Now we want to list the general formal properties of this canonical distribution (11–40). This is a bad way of doing it in one sense; at this point it all sounds very abstract and you don't see the connection to any real problem yet. On the other hand, we get all the things we want a lot faster if we first become aware of all the formal properties that are going to be in this theory in any application; and then later we'll go into specific physical problems and we'll see that every one of these formal relations turns out to have many different useful meanings, depending on the particular problem.

Now the maximum attainable H that we can get by holding these averages fixed depends, of course, on the average values we specified,

$$H_{max} = S(F_1, \cdots, F_m) = \log Z + \sum_{k=1}^{m} \lambda_k F_k. \tag{11–42}$$

H itself we can regard as a measure of the "amount of the uncertainty" in any probability distribution. After we have maximized it, it becomes a function of the definite data of the problem $\{F_i\}$, and we'll call it S with a view to the original application in physics. It's still a measure of "uncertainty," but it's uncertainty *when all the information we have consists of just these numbers.* It is "subjective" in the sense that it still measures uncertainty; but it is completely "objective" in the sense that it depends only on the *given data of the problem,* and not on anybody's personality or wishes. As we must stress repeatedly, it is "objectivity" in this sense that we need in scientific inference.

If S is to be a function only of $(F_1, \ldots, F_m)$, then in (11–42) the $(\lambda_1, \ldots, \lambda_m)$ must also be thought of as functions of $(F_1, \ldots, F_m)$. At first, the λ's were just unspecified constants, but eventually we want to know what they are. If we choose different λ_i, we are writing down different probability distributions (11–40); and we saw in (11–33) that the averages over this distribution agree with the given averages F_k if

$$F_k = \langle f_k \rangle = -\frac{\partial}{\partial \lambda_k}(\log Z), \qquad k = 1, 2, \ldots, m \tag{11–43}$$

So we are now to regard (11–43) as a set of m simultaneous equations which are to be solved for the λ's in terms of the given data F_k; at least one would like to dream about this. Generally, in a nontrival problem, it is impractical to solve for the λ's explicitly (although there is a simple formal solution in (11–44) below) and we leave the λ_k where they are, expressing things in parametric form. Actually, this isn't such a tragedy, because the λ's usually turn out to have such important physical meanings that we are quite happy to use them as the independent variables. However, if we can evaluate the function $S(F_1, \ldots, F_m)$ explicitly, then we *can* give the λ's as explicit functions of the given data, as follows.

Suppose we make a small change in one of the constraint values F_k; how does this change the maximum attainable H? We have from (11–42),

$$\frac{\partial S}{\partial F_k} = \sum_{j=1}^{m} \frac{\partial \log Z}{\partial \lambda_j} \frac{\partial \lambda_j}{\partial F_k} + \sum_{j=1}^{m} \frac{\partial \lambda_j}{\partial F_k} F_k + \lambda_k$$

which, thanks to (11–43), collapses to

$$\lambda_k = \frac{\partial S}{\partial F_k} \tag{11–44}$$

in which λ_k is given explicitly.

Compare this equation with (11–43); one gives F_k explicitly in terms of the λ_k, the other gives the λ_k explicitly in terms of the F_k. If we specify $\log Z$ as a function of the λ_k; or if we specify S as a function of the given data F_k, these are equivalent in the sense that each gives full information about the probability distribution. The complete story is contained in either function, and in fact you see that (11–42) is just the Legendre transformation that takes us from one representative function to another.

We can derive some more interesting laws simply by differentiating the two we already have. Differentiate (11–43) with respect to λ_j:

$$\frac{\partial F_k}{\partial \lambda_j} = \frac{\partial^2}{\partial \lambda_j \partial \lambda_k}(\log Z) = \frac{\partial F_j}{\partial \lambda_k} \tag{11–45}$$

since the second cross derivatives of $\log Z$ are symmetric in j and k. So here's a general reciprocity law which will hold in any problem that we do by maximizing the entropy. Likewise, if we differentiate (11–44) a second time, we have

$$\frac{\partial \lambda_k}{\partial F_j} = \frac{\partial^2 S}{\partial F_j \partial F_k} = \frac{\partial \lambda_j}{\partial F_k} \tag{11–46}$$

another reciprocity law, which is however not independent of (11–45), because if we define the matrices $A_{jk} \equiv \partial\lambda_j/\partial F_k$, $B_{jk} \equiv \partial F_j/\partial \lambda_k$, you see easily that they are inverse matrices: $A = B^{-1}$, $B = A^{-1}$. These reciprocity laws might appear trivial from the ease with which we derived them here; but when we get around to applications we'll see that they have highly nontrivial and nonobvious physical meanings. In the past, some of them were found by tedious means that made them seem mysterious and arcane.

Now let's consider the possibility that one of these functions $f_k(x)$ contains an extra parameter α which can be varied. If you want to think of applications, you can say $f_k(x_i;\alpha)$ stands for the i'th energy level of some system and α represents the volume of the system. The energy levels depend on the volume. Or, if it's a magnetic resonance system, you can say that $f_k(x_i)$ represents the energy of the i'th stationary state of the spin system and α represents the magnetic field $\mathbf{H}$ applied to it. Often we want to make a prediction of how certain quantities change as we change α. We may want to calculate the pressure or the susceptibility. By the criterion of minimum mean square error, the best estimate of the derivative would be the mean value over the probability distribution

$$\left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = \frac{1}{Z}\sum_i \exp\left[-\lambda_1 f_1(x_i) - \cdots - \lambda_k f_k(x_i;\alpha) - \cdots - \lambda_m f_m(x_i)\right] \frac{\partial f_k(x_i,\alpha)}{\partial \alpha}$$

which reduces to

$$\left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{1}{\lambda_k}\frac{\partial}{\partial \alpha}\log Z. \tag{11–47}$$

In this derivation, we supposed that this parameter α appearers in only one function f_k. If the same parameter is in several different f_k, then you verify easily that this generalizes to

$$\sum_{k=1}^{m} \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{\partial}{\partial \alpha}\log Z. \tag{11–48}$$

This general rule contains, among other things, the equation of state of any thermodynamic system.

When we add α to the problem, the maximum entropy S is a function not only of the specified average values $\langle f_k \rangle = F_k$, but it depends on α too. Likewise, Z depends on α. But if we differentiate $\log Z$ or S, we get the same thing:

$$-\frac{\partial S}{\partial \alpha} = \sum_{k=1}^{m} \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{\partial}{\partial \alpha}\log Z \tag{11–49}$$

with one tricky point that isn't brought out too clearly in this notation. In S the independent variables are $\{F_k, \alpha\}$. In other words, $S = S(F_1, \ldots, F_m; \alpha)$. But in $\log Z$ they are $\{\lambda_k; \alpha\}$: $\log Z = \log Z(\lambda_1, \ldots, \lambda_m; \alpha)$. So in (11–49) we have to understand that in $(\partial S/\partial \alpha)$ we are holding the F_k fixed, while in $(\partial \log Z/\partial\alpha)$ we are holding the λ_k fixed. The equality of these derivatives then follows from the Legendre transformation (11–42). Evidently, if there are several different parameters $\{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ in the problem, a relation of the form (11–49) will hold for each of them.

Now let's note some general "fluctuation laws," or moment theorems. First, a comment about notation: we're using the F_k, $\langle f_k\rangle$ to stand for the same *number.* They are equal because we specified that the expectation values $\{\langle f_1\rangle, \ldots, \langle f_m\rangle\}$ are to be set equal to the given data $\{F_1, \ldots, F_m\}$ of the problem. When we want to emphasize that these quantities are expectation values over the canonical distribution (11–40), we'll use the notation $\langle f_k\rangle$. When we want to emphasize that they are the given data, we'll call them F_k. At the moment, we want to do the former, and so the reciprocity law (11–45) can be written equally well as

$$\frac{\partial\langle f_k\rangle}{\partial\lambda_j} = \frac{\partial\langle f_j\rangle}{\partial\lambda_k} = \frac{\partial^2}{\partial\lambda_j\partial\lambda_k}\log Z \tag{11–50}$$

In varying the λ's here, we're changing from one canonical distribution (11–40) to a slightly different one in which the $\langle f_k\rangle$ are slightly different. Since the new distribution corresponding to $(\lambda_k + d\lambda_k)$ is still of canonical form, it is still a maximum–entropy distribution corresponding to slightly different data $(F_k + dF_k)$. Thus we are comparing two slightly different maximum entropy problems. For later physical applications it will be important to recognize this in interpreting the reciprocity law (11–50).

But now we want to show that the quantities in (11–50) also have an important meaning with reference to a *single* maximum entropy problem. In the canonical distribution (11–31), how are the different quantities $f_k(x)$ correlated with each other? More specifically, how are departures from their mean values $\langle f_k\rangle$ correlated? The measure of this is the *covariance* or second central moments of the distribution:

$$\langle (f_j - \langle f_j\rangle)(f_k - \langle f_k\rangle)\rangle = \langle f_j f_k - f_j\langle f_j\rangle - \langle f_j\rangle f_k + \langle f_j\rangle\langle f_k\rangle\rangle = \langle f_j f_k\rangle - \langle f_j\rangle\langle f_k\rangle \tag{11–51}$$

If a value of f_k greater than the average $\langle f_k\rangle$ is likely to be accompanied by a value of f_j greater than its average $\langle f_j\rangle$, the covariance is positive; if they tend to fluctuate in opposite directions, it is negative; and if their variations are uncorrelated, the covariance is zero. If $j = k$, this reduces to the *variance*:

$$\langle (f_k - \langle f_k\rangle)^2\rangle = \langle f_k^2\rangle - \langle f_k\rangle^2 \geq 0. \tag{11–52}$$

To calculate these quantities directly from the canonical distribution (11–40), we can first find

$$\begin{aligned}\langle f_j f_k\rangle &= \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \sum_{i=1}^{n} f_j(x_i) f_k(x_i) \exp[-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)] \\ &= \frac{1}{Z}\sum_{i-1}^{n} \frac{\partial^2}{\partial\lambda_j\partial\lambda_k} \exp[-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)] \\ &= \frac{1}{Z}\frac{\partial^2 Z}{\partial\lambda_j\partial\lambda_k}\end{aligned} \tag{11–53}$$

Then, using (11–43), the covariance becomes

$$\langle f_j f_k \rangle - \langle f_j \rangle \langle f_k \rangle = \frac{1}{Z} \frac{\partial^2 Z}{\partial \lambda_j \partial \lambda_k} - \frac{1}{Z^2} \frac{\partial Z}{\partial \lambda_j} \frac{\partial Z}{\partial \lambda_k} = \frac{\partial^2}{\partial \lambda_j \partial \lambda_k} \log Z \tag{11–54}$$

But this is just the quantity (11–50); therefore the reciprocity law takes on a bigger meaning,

$$\langle f_j f_k \rangle - \langle f_j \rangle \langle f_k \rangle = -\frac{\partial \langle f_j \rangle}{\partial \lambda_k} = -\frac{\partial \langle f_k \rangle}{\partial \lambda_j} . \tag{11–55}$$

The second derivatives of $\log Z$ which gave us the reciprocity law also give us the covariance of f_j and f_k in our distribution.

Note that (11–55) is in turn only a special case of a more general rule: Let $q(x)$ be any function; then the covariance with $f_k(x)$ is, as you verify easily,

$$\langle q f_k \rangle - \langle q \rangle \langle f_k \rangle = -\frac{\partial \langle q \rangle}{\partial \lambda_k} . \tag{11–56}$$

Exercise 11.3. From comparing (11–43), (11–50), (11–55) we might expect that still higher derivatives of $\log Z$ would correspond to higher central moments of the distribution (11–40). Check this conjecture by calculating the third and fourth central moments in terms of $\log Z$. *Hint*: See Appendix C on the theory of cumulants.

For noncentral moments, it is customary to define a *moment generating function*

$$\Phi(\beta_1, \ldots, \beta_m) \equiv \langle \exp[\beta_1 f_1 + \cdots + \beta_m f_m] \rangle \tag{11–57}$$

which evidently has the property

$$\langle f_i^{m_i} f_j^{m_j} \cdots \rangle = \left(\frac{\partial^{m_i}}{\partial \beta_i^{m_i}} \frac{\partial^{m_j}}{\partial \beta_j^{m_j}} \cdots \right) \Phi(\beta_1, \ldots, \beta_m) \Big|_{\beta_k = 0} \tag{11–58}$$

However, we find from (11–57)

$$\Phi(\beta_1, \ldots, \beta_m) = \frac{Z[(\lambda_1 - \beta_1), \ldots, (\lambda_m - \beta_m)]}{Z(\lambda_1, \ldots, \lambda_m)} \tag{11–59}$$

so that the partition function Z serves this purpose; instead of (11–58) we may write equally well,

$$\langle f_i^{m_i} f_j^{m_j} \cdots \rangle = \frac{1}{Z} \left(\frac{\partial^{m_i}}{\partial \lambda_i^{m_i}} \frac{\partial^{m_j}}{\partial \lambda_j^{m_j}} \cdots \right) Z \tag{11–60}$$

which is the generalization of (11–53).

Now, we might ask, what are the covariances of the derivatives of f_k with respect to a parameter α? Define

$$g_k \equiv \frac{\partial f_k}{\partial \alpha} . \tag{11–61}$$

Then, for example, if f_k is the energy and α is the volume then $-g_k$ is the pressure. We easily verify another reciprocity relation:

$$\frac{\partial\langle g_j\rangle}{\partial\lambda_k} = -[\langle g_j f_k\rangle - \langle g_j\rangle\langle g_k\rangle] = \frac{\partial\langle g_k\rangle}{\partial\lambda_j\rangle} \tag{11–62}$$

analogous to (11–55). By a similar derivation, we find the identity

$$\sum_{j=1}^{m}\lambda_j[\langle g_j g_k\rangle - \langle g_j\rangle\langle g_k\rangle] = \left\langle\frac{\partial g_k}{\partial\alpha}\right\rangle - \frac{\partial\langle g_k\rangle}{\partial\alpha}\,. \tag{11–63}$$

We had found and used special cases of this for some time before realizing its generality.

Other derivatives of $\log Z$ are related to various moments of the f_k and their derivatives with respect to α. For example, closely related to (11–63) is

$$\frac{\partial^2\log Z}{\partial\alpha^2} = \sum_{jk}\lambda_j\lambda_k[\langle g_j g_k\rangle - \langle g_j\rangle\langle g_k\rangle] - \sum_k\lambda_k\left\langle\frac{\partial g_k}{\partial\alpha}\right\rangle \tag{11–64}$$

The cross–derivatives give us a simple and useful relation

$$\frac{\partial^2\log Z}{\partial\alpha\partial\lambda_k} = -\frac{\partial\langle f_k\rangle}{\partial\alpha} = \sum_j\lambda_j[\langle f_k g_j\rangle - \langle f_k\rangle\langle g_j\rangle] - \langle g_k\rangle \tag{11–65}$$

which also follows from (11–50) and (11–56); and by taking further derivatives an infinite hierarchy of similar moment relations is obtained. As we will see later, the above theorems have as special cases many relations such as the Einstein fluctuation laws for black–body radiation and for density of a gas or liquid, the Nyquist voltage fluctuations, or "noise" generated by a reversible electric cell, *etc.*

It is evident that if several different parameters $\{\alpha_1,\cdots,\alpha_r\}$ are present, relations of the above form will hold for each of them; and new ones like

$$\frac{\partial^2\log Z}{\partial\alpha_1\partial\alpha_2} = \sum_k\lambda_k\left\langle\frac{\partial^2 f_k}{\partial\alpha_1\partial\alpha_2}\right\rangle - \sum_{kj}\lambda_j\lambda_k\left[\left\langle\frac{\partial f_k}{\partial\alpha_1}\frac{\partial f_j}{\partial\alpha_2}\right\rangle - \left\langle\frac{\partial f_k}{\partial\alpha_1}\right\rangle\left\langle\frac{\partial f_j}{\partial\alpha_2 1}\right\rangle\right] \tag{11–66}$$

will appear.

The relation (11–42) between $\log Z(\lambda_1,\ldots,\lambda_m;\alpha_1,\ldots,\alpha_m)$ and $S(\langle f_1\rangle,\ldots,\langle f_m\rangle;\alpha_1,\ldots,\alpha_r)$, shows that they can all be stated also in terms of derivatives (*i.e.*, variational properties) of S. In the case of S, however, there is a still more general and important variational property.

In (11–44) we supposed that the definitions of the functions $f_k(x)$ were fixed once and for all, the variation of $\langle f_k\rangle$ being due only to variations in the p_i. We now derive a more general variational statement in which both of these quantities are varied. Let $\delta f_k(x_i)$ be specified arbitrarily and independently for each value of k and i, let $\delta\langle f_k\rangle$ be specified independently of the $\delta f_k(x_i)$, and consider the resulting change from one maximum–entropy distribution p_i to a slightly different one $p'_i = p_i + \delta p_i$, the variations δp_i and $\delta\lambda_k$ being determined in terms of $\delta f_k(x_i)$ and $\delta\langle f_k\rangle$ through the above equations. In other words, we are now considering two slightly different maximum–entropy problems in which all conditions of the problem – including the definitions of the functions $f_k(x)$ on which it is based – are varied arbitrarily. The variation in $\log Z$ is

$$\begin{aligned}\delta \log Z &= \frac{1}{Z}\sum_{i=1}^{n}\left\{\sum_{k=1}^{m}[-\lambda_k \delta f_k(x_i) - \delta\lambda_k f_k(x_i)]\exp[-\sum_{j=1}^{m}\lambda_j f_j(x_i)]\right\} \\ &= -\sum_{k=1}^{m}[\lambda_k\langle\delta f_k\rangle + \delta\lambda_k\langle f_k\rangle]\end{aligned} \tag{11–67}$$

and thus from the Legendre transformation (11–42)

$$\delta S = -\sum_k \lambda_k[\delta\langle f_k\rangle - \langle\delta f_k\rangle]\,, \qquad \text{or} \qquad \delta S = \sum_k \lambda_k \delta Q_k \tag{11–68}$$

where

$$\delta Q_k \equiv \delta\langle f_k\rangle - \langle \delta f_k\rangle = \sum_{i=1}^{N} f_k(x_i)\delta p_i\,. \tag{11–69}$$

This result, which generalizes (11–44), shows that the entropy S is stationary not only in the sense of the maximization property which led to the canonical distribution (11–40); it is also stationary with respect to small variations in the functions $f_k(x_i)$ if the p_i are held fixed.

As a special case of (11–68), suppose that the functions f_k contain parameters $\{\alpha_1, \ldots, \alpha_r\}$ as in (11–66), which generate the $\delta f_k(x_i)$ by

$$\delta f_k(x_i, \alpha_j) = \sum_{j=1}^{r}\frac{\partial f_k(x_i,\alpha)}{\partial \alpha_j}\delta\alpha_j. \tag{11–70}$$

While δQ_k is not in general the exact differential of any function $Q_k(\langle f_i\rangle; \alpha_j)$, (11–68) shows that λ_k is an integrating factor such that $\sum \lambda_k \delta Q_k$ is the exact differential of a "state function" $S(\langle f_i\rangle;\alpha_j)$. At this point, perhaps all this is beginning to sound familiar to those who have studied thermodynamics.

Finally, we leave it for you to prove from (11–68) that

$$\sum_{k=1}^{m}\langle f_k\rangle\frac{\partial\lambda_k}{\partial\alpha} = 0 \tag{11–71}$$

where $\langle f_1\rangle, \ldots, \langle f_r\rangle$ are held constant in the differentiation.

Evidently, there's now a large new class of problems which we can ask the robot to do, which it can solve in rather a wholesale way. It first evaluates this partition function Z, or better still, $\log Z$. Then just by differentiating that with respect to all its arguments in every possible way, it obtains all sorts of predictions in the form of mean values over the maximum–entropy distribution. This is quite a neat mathematical procedure, and, of course, you recognize what we have been doing here. These relations are all just the standard equations of statistical mechanics given to us by J. Willard Gibbs, but now in a disembodied form with all the physics removed.

Indeed, virtually all known thermodynamic relations, found over more than a Century by the most diverse and difficult kinds of physical reasoning and experimentation, are now seen as straightforward mathematical identities of the Maximum Entropy formalism. This makes it clear that those relations are actually independent of any particular physical assumptions and are properties of extended logic in general, giving us a new insight into why the relations of thermodynamics are so general, independent of the properties of any particular substance. This historically oldest application, Gibbsian statistical mechanics, is still the most used (although many of its users are still unaware of its generality). We stress that generality by considering statistical mechanics only in Chapter 29, after a few other applications have been expounded.

The maximum entropy mathematical formalism has a mass of other applications outside of physics. In Chapter 14 we work out the full numerical solution to a nontrivial problem of inventory control, and in Chapter 27 we give a highly nontrivial analytical solution of a problem of optimal encoding in communication theory. In a sense, once we have understood the maximum entropy principle as explained in this Chapter, most applications of probability theory are seen as invoking it to assign the initial probabilities – whether called technically prior probabilities or sampling probabilities. Whenever we assign uniform prior probabilities, we can say truthfully that we are applying maximum entropy (although in that case the result is so simple and intuitive that we do not need any of the above formalism). As we saw already in Chapter 7, whenever we assign a Gaussian sampling distribution, this is the same as applying maximum entropy for given first and second moments. And we saw in Chapter 9 that whenever we assign a binomial sampling distribution, this is mathematically equivalent to assigning the uniform maximum entropy distribution on a deeper hypothesis space.

Conceptual Problems – Frequency Correspondence.

The principle of maximum entropy is basically a simple and straightforward idea, and in the case that the given information consists of average values it leads, as we have just seen, to a surprisingly concise mathematical formalism, since essentially everything is known if we can evaluate a single function $\log Z(\lambda_1, \ldots, \lambda_m; \alpha_1, \ldots, \alpha_r)$. Nevertheless, it seems to generate some serious conceptual difficulties, particularly to people who have been trained to think of probability only in the frequency sense. Therefore, before turning to applications, we want to examine, and hopefully resolve, some of these difficulties. Here are some of the objections that have been raised against the principle of maximum entropy:

(A) If the only justification for the canonical distribution (11–40) is "maximum uncertainty," that is a negative thing which can't possibly lead to any useful predictions; you can't get reliable results out of mere ignorance.

(B) The probabilities obtained by maximum entropy cannot be relevant to physical predictions because they have nothing to do with frequencies – there is absolutely no reason to suppose that distributions observed experimentally would agree with ones found by maximizing entropy.

(C) The principle is restricted to the case where the constraints are average values – but almost always the given data $\{F_1, \ldots, F_n\}$ are *not* averages over anything. They are definite measured numbers. When you set them equal to averages, $F_k = \langle f_k \rangle$, you are committing a logical contradiction, for the given data said that f_k had the value F_k; yet you immediately write down a probability distribution that assigns nonzero probabilities to values of $f_k \neq F_k$.

(D) The principle cannot lead to any definite physical results because different people have different information, which would lead to different distributions – the results are basically arbitrary.

Objection (A) is, of course, nothing but a play on words. The "uncertainty" was always there. Our maximizing the entropy did not *create* any "ignorance" or "uncertainty;" it is rather the means of determining quantitatively the full extent of the uncertainty already present. It is *failure* to do this – and as a result using a distribution that implies more knowledge than we really have – that would lead to dangerously unreliable conclusions.

Of course, the information put into the theory as constraints on our maximum–entropy distribution, may be so meager – the distribution is so weakly constrained from the uninformative uniform one – that no reliable predictions can be made from it. But in that case, as we will see later, the theory automatically tells us this. If we emerge with a very broad probability distribution for some quantity θ of interest (such as pressure, magnetization, electric current density, rate of diffusion, *etc.*, that is the robot's way of telling us: "You haven't given me enough information to determine any definite prediction." But if we get a very sharp distribution for θ [for example – and typical of what does happen in many real problems – if the theory says the odds on θ being in the interval $\theta_0(1 \pm 10^{-6})$ are greater than $10^{10} : 1$], then the given information *was* sufficient to make a very definite prediction.

But in both cases, and in the intermediate ones between these extremes, the distribution for θ always tells us just what conclusions we *are* entitled to draw about θ, on the basis of the information *which was put into the equations.* If someone has additional cogent information, but fails to incorporate it into his calculation, the result is not a failure, only a misuse, of the maximum entropy method.

To answer objection (B), we show that the situation is vastly more subtle than that. The principle of maximum entropy has, fundamentally, nothing to do with any repeatable "random experiment." Some of the most important applications are to cases where the probabilities p_i in (11–40) have no frequency connection – the x_i are simply an enumeration of the *possibilities,* in the single situation being considered, as in the cars on the ferry problem.

However, nothing prevents us from applying the principle of maximum entropy also to cases where the x_i are generated by successive repetitions of some experiment as in the dice problem; and in this case, the question of the relation between the maximum–entropy probability $p(x_i)$ and the frequency with which x_i is observed, is capable of mathematical analysis. We demonstrate that (1) in this case the maximum–entropy probabilities *do* have a precise connection with frequencies; (2) in most real problems, however, this relation is unnecessary for the usefulness of the method; (3) in fact, the principle of maximum entropy is most useful to us in just those cases where the observed frequencies do *not* agree with the maximum–entropy probabilities.

Suppose now that the value of x is determined by some random experiment; at each repetition of the experiment the final result is one of the values x_i, $i = 1, 2, \ldots, n$; in the dice problem, $n = 6$. But now, instead of asking for the probability p_i, let's ask an entirely different question: on the basis of the available information, what can we say about the relative *frequencies* f_i with which the various x_i occur?

Let the experiment consist of N trials (we are particularly interested in the limit $N \to \infty$, because that is the situation contemplated in the usual frequency theory of probability), and let every conceivable sequence of results be analyzed. Each trial could give, independently, any one of the results $\{x_1, \ldots, x_n\}$; and so there *a priori* n^N conceivable outcomes of the whole experiment. But many of these will be incompatible with the given information (let's suppose again that his consists of average values of several functions $f_k(x)$, $k = 1, 2, \ldots, m$; in the end it will be clear that the final conclusions are independent of whether it takes this form or some other). We will, of course, assume that the result of the experiment agrees with this information – if it didn't, then the given information was false and we are doing the wrong problem. In the whole experiment, the results x_i will be obtained n_1 times, x_2 will be obtained n_2 times, *etc.*. Of course,

$$\sum_{i=1}^{n} n_i = N \tag{11–72}$$

and if the specified mean values F_k given to us are in fact observed in the actual experiment, we have the additional relation

$$\sum_{i=1}^{n} n_i f_k(x_i) = N F_k, \qquad d = 1, 2, \ldots, m \tag{11–73}$$

If $m < n - 1$, the relations (11–72), (11–73) are insufficient to determine the relative frequencies $f_i = n_i/N$. Nevertheless, we do have grounds for preferring some choices of the f_i to others. For, out of the original n^N conceivable outcomes, how many would lead to a given set of sample numbers $\{n_1, n_2, \ldots, n_n\}$? The answer is, of course, the multinomial coefficient

$$W = \frac{N!}{n_1! n_2! \ldots n_n!} = \frac{N!}{(Nf_1)!(Nf_2)! \ldots (Nf_n)!} \tag{11–74}$$

The set of frequencies $\{f_1, \cdots, f_n\}$ which can be realized in the greatest number of ways is therefore the one which maximizes W subject to the constraints (11–72), (11–73). Now we can equally well maximize any monotonic increasing function of W, in particular $N^{-1} \log W$; but as $N \to \infty$ we have, as we saw already in (11–18),

$$\frac{1}{N} \log W \to -\sum_{i=1}^{n} f_i \log f_i = H_f \tag{11–75}$$

So you see that, in (11–72), (11–73), (11–75) we have formulated exactly the same mathematical problem as in the maximum–entropy derivation, so the two problems will have the same solution. This argument is mathematically reminiscent of the Wallis derivation noted above; and the same result could have been found as well by direct application of Bayes' theorem, assigning uniform prior probabilities over all the n^N conceivable outcomes and passing to the limit $N \to \infty$.

You see also, in partial answer to objection (C), that this identity of the mathematical problems will persist whether or not the constraints take the form of mean values. If the given information does consist of mean values, then the mathematics is particularly neat, leading to the partition function, *etc.* But, for given information which places *any* definite kind of constraint on the problem, we have the same conclusion: the *probability* distribution which maximizes the entropy is numerically identical with the *frequency* distribution which can be realized in the greatest number of ways.

The maximum in W is, furthermore, enormously sharp. To show this, let $\{f_1, \ldots, f_n\}$ be the set of frequencies which maximizes W and has entropy H_f; and let $\{f'_1, \ldots, f'_n\}$ be any other set of possible frequencies [that is, a set which satisfies the constraints (11–72), (11–73) and has entropy $H_{f'} < H_f$. The ratio (number of ways in which f_i could be realized)/(number of ways in which f'_i could be realized) grows asymptotically, according to (11–75), as

$$\frac{W}{W'} \to \exp\{N(H_f - H_{f'})\} \tag{11–76}$$

and passes all bounds as $N \to \infty$. Therefore, the frequency distribution predicted by maximum entropy can be realized experimentally in *overwhelmingly* more ways than can any other that satisfies the same constraints. This is related later to the more complete entropy concentration theorem.

We have here another precise and quite general connection between probability and frequency; it had nothing to do with the definition of probability, but emerged as a mathematical *consequence* of probability theory, interpreted as extended logic. Another kind of connection between probability and frequency, whose precise mathematical statement is different in form, but which has the same practical consequences, will appear in Chapter 12.

Turning to objection (C), our purpose in imposing constraints is to incorporate certain information into our probability distribution. Now, what does it mean to say that a probability distribution "contains" some information? We take this as meaning that the information can be extracted from it by using the usual rule for estimation, calculating the mean value. Usually, the datum F_k is of unknown accuracy, and so using it to constrain only the $\langle F_k \rangle$ is just the process of being honest, leaving the width of the distribution for $f_k(x)$ to be determined by the range and density of the set of possibilities x_i. But if we do have independent information about the accuracy of F_1, that can be incorporated by adding a new constraint on $\langle f_1(x_i)^2 \rangle$; the formalism already allows for this. But this seldom makes any substantive difference in the final conclusions, because the variance of the maximum–entropy distribution for $f_1(x)$ is usually small compared to any reasonable mean–square experimental error.

Now let's turn to objection (D) and analyze the situation with come care, because it is perhaps the most common of all of them. Does the above connection between probability and frequency justify our predicting that the maximum–entropy distribution will in fact be observed as a frequency distribution in a real experiment? Clearly not, in the sense of deductive proof; for just as objection (D) points out, we have to concede that different people may have different amounts of information, which will lead them to writing down different distributions which make different predictions of observable facts, and they can't all be right. But this misses the point about what we are trying to do; let's look at it more closely.

Consider a specific case: Mr. A imposes constraints on the mean values $\langle f_1(x) \rangle$, $\langle f_2(x) \rangle$ to agree with his data F_1, F_2. Mr. B, better informed, imposes in addition a constraint on $\langle f_3(x) \rangle$ to agree with his extra datum F_3. Each sets up a maximum–entropy distribution on the basis of his information. Since Mr. B's entropy is maximized subject to one further constraint, we will have

$$S_B \le S_A\,. \tag{11–77}$$

Suppose that Mr. B's extra information was redundant, in the sense that it was only what Mr. A would have predicted from his distribution. Now Mr. A has maximized his entropy with respect to all variations of the probability distribution which hold $\langle f_1 \rangle$, $\langle f_2 \rangle$ fixed at the specified values F_1, F_2. Therefore, he has *a fortiori* maximized it with respect to the smaller class of variations which also hold $\langle f_3 \rangle$ fixed at the value finally attained. Therefore Mr. A's distribution also solves Mr. B problem in this case; $\lambda_3 = 0$, and Mr. A and Mr. B have identical probability distributions. In this case, and only in this case, we have equality in (11–77).

From this we learn two things: (1) Two people with different given information do not necessarily arrive at different maximum–entropy distributions; this is the case only when Mr. B's extra information was "surprising" to Mr. A. (2) In setting up a maximum–entropy problem, it is not necessary to determine whether the different pieces of information used are independent: any redundant information will not be "counted twice," but will drop out of the equations automatically. Indeed, this not only agrees with our basic desideratum that $AA = A$ in Boolean algebra; it would be true of any variational principle (imposing a new constraint cannot change the solution if the old solution already satisfied that constraint).

Now suppose the opposite extreme: Mr. B's extra information was logically contradictory to what Mr. A knows. For example, it might turn out that $f_3(x) = f_1(x) + 2f_2(x)$, but Mr. B's data failed to satisfy $F_3 = F_1 + 2F_2$. Evidently, there is *no* probability distribution that fits Mr. B's supposed data. How does our robot tell us this? Mathematically, you will then find that the equations

$$F_k = -\frac{\partial}{\partial \lambda_k} \log Z(\lambda_1, \lambda_2, \lambda_3) \tag{11–78}$$

have no simultaneous solution with real λ_k. In the example just mentioned,

$$\begin{aligned} Z(\lambda_1, \lambda_2, \lambda_3) &= \sum_{i=1}^{n} \exp[-\lambda_1 f_1(x_i) - \lambda_2 f_2(x_i) - \lambda_3 f_3(x_i)] \\ &= \sum_{i=1}^{n} \exp[-(\lambda_1 + \lambda_3) f_1(x_i) - (\lambda_2 + 2\lambda_3) f_2(x_i)] \end{aligned} \tag{11–79}$$

and so

$$\frac{\partial Z}{\partial \lambda_3} = \frac{\partial Z}{\partial \lambda_1} + 2\frac{\partial Z}{\partial \lambda_2} \tag{11–80}$$

and so (11–78) cannot have solutions for λ_1, λ_2, λ_3 unless $F_3 = F_1 + 2F_2$. So, when a new piece of information logically contradicts previous information, the principle of maximum entropy breaks down, as it should, refusing to give us any distribution at all.

The most interesting case is the intermediate one where Mr. B's extra information was neither redundant nor contradictory. He then finds a maximum–entropy distribution different from that of Mr. A, and the inequality holds in (11–77), indicating that Mr. B's extra information was "useful" in further narrowing down the range of possibilities allowed by Mr. A's information. The measure of this range is just W; and from (11–76) we have asymptotically,

$$\frac{W_A}{W_B} \sim \exp\{N(S_A - S_B)\} \tag{11–81}$$

For large N, even a slight decrease in the entropy leads to an enormous decrease in the number of possibilities.

Suppose now that we start performing the experiment with Mr. A and Mr. B watching. Since Mr. A predicts a mean value $\langle f_3 \rangle$ different from the correct one known to Mr. B, it is clear that the experimental distribution cannot agree in all respects with Mr. A's prediction. We cannot be sure in advance that it will agree with Mr. B's prediction either, for there may be still further constraints on $f_4(x)$, $f_5(x)$, ..., *etc.*, operating in the experiment unknown to Mr. B.

However, the property demonstrated above does justify the following weaker statement of frequency correspondence: if the information incorporated into the maximum–entropy analysis includes all the constraints actually operating in the random experiment, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally. Indeed, most frequency distributions observed in Nature are maximum–entropy distributions, simply because they can be realized in so many more ways than can any other.

Conversely, suppose the experiment fails to confirm the maximum–entropy prediction, and this disagreement persists indefinitely on repetition of the experiment. Then, since by hypothesis the data F_i were true if incomplete, we will conclude that the physical mechanism of the experiment must contain some additional constraint which was not taken into account in the maximum–entropy calculation. The observed deviations then provide a clue as to the nature of this new constraint. In this way, Mr. A can discover empirically that his information was incomplete.

In summary, the principle of maximum entropy is not an Oracle telling which predictions *must* be right; it is a rule for inductive reasoning that tells us which predictions *are most strongly indicated by our present information.*

COMMENTS

The little scenario just described is an accurate model of just what did happen in one of the most important applications of statistical analysis, carried out by J. Willard Gibbs. By the year 1901 it was known that in classical statistical mechanics, use of the canonical ensemble (which Gibbs derived as the maximum–entropy distribution over the classical state space, or phase volume, based on a specified mean value of the energy) failed to predict some thermodynamic properties (heat capacities, equation of state) correctly. Analysis of the data showed that the entropy of a real physical system was always less than the value predicted. At that time, therefore, Gibbs was in just the position of Mr. A in the scenario, and the conclusion was that the microscopic laws of physics must involve some additional constraint not contained in the laws of classical mechanics.

But Gibbs died in 1903 and it was left to others to find the nature of this constraint; first by Planck in the case of radiation, then by Einstein and Debye for solids, and finally by Bohr for isolated atoms. The constraint consisted in the discreteness of the possible energy values, thenceforth called energy levels. By 1927, the mathematical theory by which these could be calculated from first principles had been developed by Heisenberg and Schrödinger.

Thus it is an historical fact that the first clues indicating the need for the quantum theory, and indicating some necessary features of the new theory, were uncovered by a seemingly "unsuccessful" application of the principle of maximum entropy. We may expect that such things will happen again in the future, and this is the basis of the remark that the principle of maximum entropy is most useful to us in just those cases where it fails to predict the correct experimental facts. This illustrates the real nature, function, and value of inductive reasoning in science; an observation that was stressed also by Jeffreys (1957).

Gibbs (1902) wrote his probability density in phase space in the form

$$w(q_1, \ldots, q_n; p_1, \ldots, p_n) = \exp[\eta(q_1 \ldots q_n)] \qquad (11\text{–}82)$$

and called the function η the "index of probability of phase." He derived his canonical and grand canonical ensembles from constraints on average energy, and average energy and particle numbers, respectively, as (loc. cit., p. 143) "the distribution in phase which without violating this condition gives the least value of the average index of probability of phase $\overline{\eta} \ldots$" This is, of course, just what we would describe today as maximizing the entropy subject to constraints.

Unfortunately, Gibbs' work was left unfinished due to failing health. He did not give any clear explanation, and we can only conjecture whether he possessed one, as to why this particular function is to be maximized in preference to all others. Consequently, his procedure appeared arbitrary to many, and for sixty years there was confusion and controversy over the justification for Gibbs' methods; they were rejected summarily by some writers on statistical mechanics, and treated with the greatest caution by others. Only with the work of Shannon (1948) could one see the way to new thinking on a fundamental level. These historical matters are discussed in more detail in Jaynes (1967) and Jaynes (1992b).