Lecture 10

DISCRETE PRIOR PROBABILITIES--THE ENTROPY PRINCIPLE

I would like to return to the job of designing this robot. We've got part of his brain designed, and we have seen how he would reason in a few simple problems of hypothesis testing and estimation. But he is still not a very versatile reasoning machine, because he has only one means by which he can translate raw information into numerical values of probability; the "principle of indifference," Rule 4. Consistency requires him to recognize the relevance of prior information, and so in almost every problem he is faced at the outset with the problem of assigning prior probabilities. He can use Rule 4 for this if he 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 he has. But often he will have prior information that does give him some reason for preferring one possibility to another. What do we do in this case?

10.1 A New Kind of Prior Information.

Let's imagine a certain class of problems in which the robot's prior information consists of average values of certain things. Suppose, for example, we tell him that statistics were collected in a recent earthquake and that out of 100 windows broken, there were 1,000 pieces found. We will state this in the form: "the average window is broken into 10 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. Let's imagine some other problems where the same situation would arise. Here's a fairly elaborate one.

Suppose I have a table which I cover with a black cloth, and I have some dice, which I am going to toss onto this table, but for reasons that will be clear in a minute, let's make these dice black with white spots. I toss a die onto the black table. Above I have a camera. Every time I toss it, I take a snapshot. The camera will record only the white spots. Now I 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 density of the film, 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 on top turned out to be 4 1/2 instead of the 3 1/2 that we might expect from an honest die. What probability should our robot assign to the n'th face coming up?

To give still another example of a problem where the information available consists of average values, suppose that we have a string of 1,000 cars, bumper to bumper, and they occupy the full length of say three miles. We know the total length of this string of cars, and as they drive onto a rather large ferry boat, the distance that it sinks into the water tells us their total weight. So we know the average length and the average weight of the 1,000 cars. 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 number of cars of each make that were in the cluster? That is a problem where we have two average values given to us.

Now, it is not at all obvious how our robot should handle problems of this sort. So let's think about how we would want him to behave in this situation. We would not want him to jump to conclusions which are not warranted by the evidence he has. He should always frankly admit the full extent of his ignorance. 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 subsequent information 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 him 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 as "spread-out" as possible. In particular, the robot must not ignore any possibility--he must not assign zero probability to any situation unless his information really rules out that situation.

So, 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 his available information. Let's approach this in the way all 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 this distribution is would be its variance. Would it make sense if we build into the robot the property that whenever he is given information about average values, he will assign probabilities in such a way that the variance is maximized subject to that information? Well, consider the distribution of maximum variance for a given $\overline{m}$ if the values of m are unlimited, as in the broken window problem. Then the maximum variance

solution would be just the one where we assign a very large probability for no breakage at all, and an enormously small probability for 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 10. 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 you come out with the right average. So that, evidently, is not the way we would want our robot to behave; if he used the principle of maximum variance, he would be assigning zero probability to many cases which were not at all impossible on the information we gave him.

10.2. 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. I 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 by writing

$$\delta\left[\sum_m p_m^2 - \lambda \sum_m m p_m - \mu \sum_m p_m\right]$$

$$= \sum_m (2p_m - \lambda m - \mu)\, \delta p_m = 0 \qquad (10\text{-}1)$$

where λ and μ are Lagrange multipliers. So p_m will always be a linear function of m:

$$2p_m - \lambda m - \mu = 0.$$

Now, μ and λ are found from

$$\sum_m p_m = 1 \quad , \qquad \sum_m m p_m = \bar{m} \quad , \tag{10-2}$$

where $\bar{m}$ is the average value of m.

Let's investigate this and draw the graph for a simple version. Let's say that m can take on only the values 1, 2, and 3. Then we easily find that the formal solution for minimum $\sum_m p_m^2$ is

$$\begin{aligned} p_1 &= \frac{4}{3} - \frac{\bar{m}}{2} \\ p_2 &= \frac{1}{3} \\ p_3 &= \frac{\bar{m}}{2} - \frac{2}{3} \end{aligned} \tag{10-3}$$

In Figure (10.1) these results are plotted. This shows that p_1 and p_3 become negative. In these regions let's say we will replace the negative values by zero and then adjust the other probabilities to agree with the given value of $\bar{m}$. If we do this the results are shown in Figure (10.2).

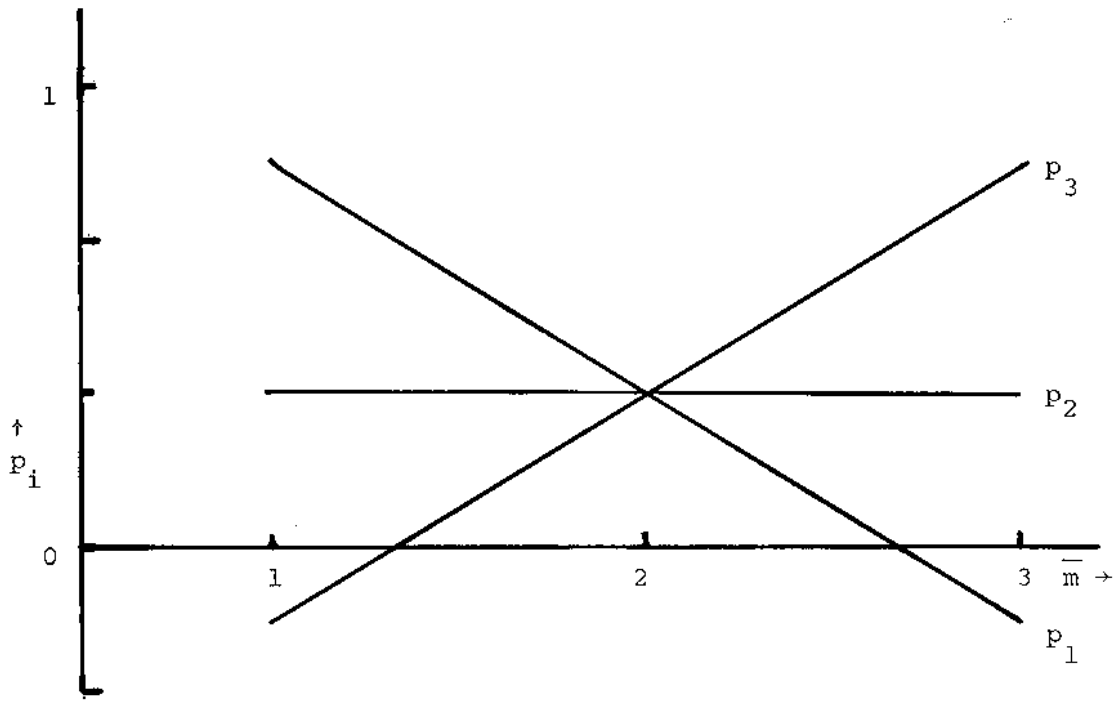

Figure 10.1. Formal solution for minimum $\sum p^2$.

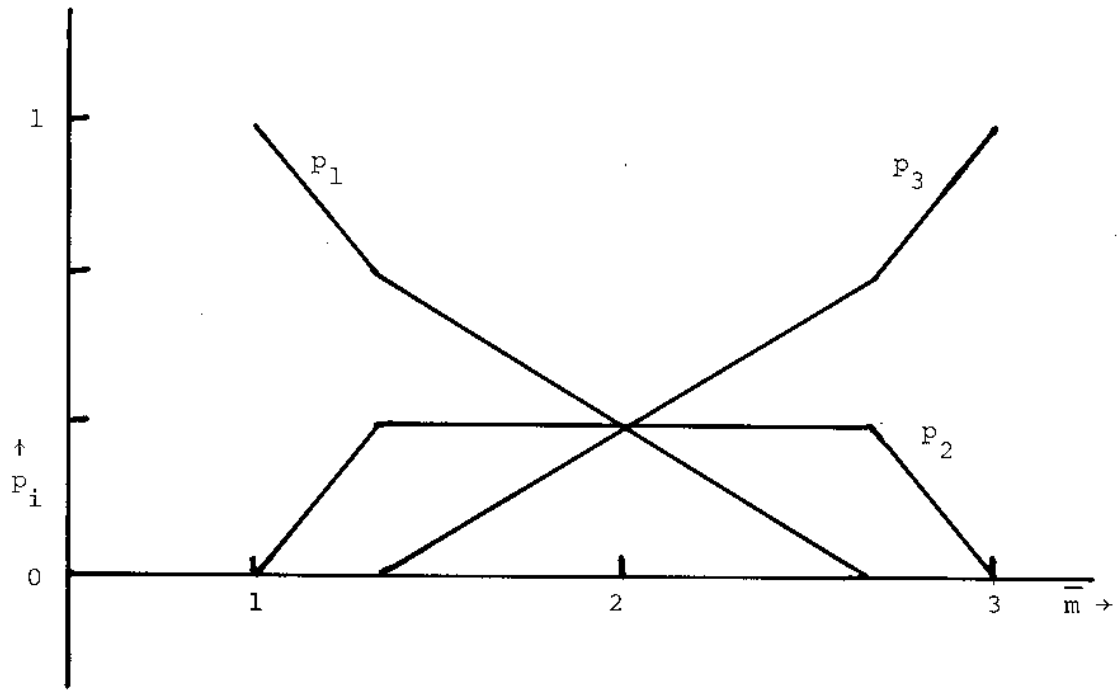

Figure 10.2. Corrected formal solution.

All right, so that's what this criterion will give to us. Now, is the robot behaving in a reasonable way if we build this behavior pattern into him? This is certainly a big improvement over maximum variance, but he is still, in certain ranges of $\overline{m}$, assigning zero probability to one of the possibilities, and there is nothing in the data we gave him which said one was impossible. So he is still jumping to unjustified conclusions. But the idea behind it 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 it forces the robot to be completely honest about what he knows, and in particular it does not permit the robot to draw any conclusions unless those conclusions are really justified by the evidence he has.

10.3. Entropy: Shannon's Theorem.

Well, at this stage we turn to the most quoted theorem in Shannon's work on information theory (Shannon, 1948; Shannon and Weaver, 1949). This is the theorem. 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. I am going to state them in a way which will remind you of the arguments we gave in Lecture 3; in fact, this is really a continuation of the basic development of probability theory. Here is the line of reasoning:

(1) We assume that some numerical measure $H_n(p_1, p_2, \cdots, p_n)$ 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

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

shall be a monotonic increasing function of n. This merely 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've got to get the same answer for every possible way.

Previously, our conditions of consistency took the form of the functional equations (3-2), (3-7), (3-11). 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 he assigns probabilities p_1 and $q = 1 - p_1$. 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 he 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, he decides whether the first possibility is or is not true; his uncertainty for 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

$$H_3(p_1,p_2,p_3) = H_2(p_1,q) + qH_2\left(\frac{p_2}{q},\frac{p_3}{q}\right) \tag{10-4}$$

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 (10-4) 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. The most general equation of consistency would be a functional equation which is satisfied by any monotonic increasing function of H_n, but I don't know how to write it.

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, let's find the most general form of the composition law (10-4)

for the case that there are n mutually exclusive propositions $(A_1,\ldots,A_n)$ to consider, to which we assign probabilities $(p_1,\ldots,p_n)$ respectively. Instead of giving the probabilities of the $(A_1,\ldots,A_n)$ directly, we might first group the first k of them together as the proposition denoted by $(A_1+A_2+\ldots+A_k)$ in Boolean algebra, and give its probability which by Eq. (3-21) is equal to $w_1 = (p_1+\ldots+p_k)$; then the next m propositions are combined into $(A_{k+1}+\ldots+A_{k+m})$, for which we give the probability $w_2 = (p_{k+1}+\ldots+p_{k+m})$, 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 $(A_1+\ldots+A_k)$ is true. The additional uncertainty $H(p_1/w_1,\ldots p_k/w_1)$ is then encountered with probability w_1. Carrying this out for the other composite propositions $(A_{k+1}+\ldots+A_{k+m})$, etc., we arrive ultimately at the same state of knowledge as if the $(p_1,\ldots,p_n)$ had been given directly; so if our measure of "amount of uncertainty" is to be consistent, we must obtain the same ultimate uncertainty no matter how the choices were broken down in this way. Thus we must have

$$H(p_1\ldots p_n) = H(w_1\ldots w_r) + w_1 H(p_1/w_1,\ldots,p_k/w_1) + w_2 H(p_{k+1}/w_2,\ldots,p_{k+m}/w_2) + \ldots \qquad (10\text{-}5)$$

which is the general form of the functional equation (10-4). 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 be sufficient to determine it for all rational values

$$p_i = \frac{n_i}{\Sigma n_i} \qquad (10\text{-}6)$$

with n_i integers. But then (10-5) determines the function H already in terms of the quantities $h(n) \equiv H(1/n,\ldots,1/n)$ which measure "amount of uncertainty" in the case of n equally likely alternatives. For we can regard a choice of

one of the alternatives $(A_1,\ldots,A_n)$ 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 (10-5) 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, (10-5) reduces to

$$h(\Sigma n_i) = H(p_1\ldots p_n) + \sum_i p_i h(n_i) \qquad (10\text{-}7)$$

Now we can choose all $n_i = m$; whereupon (10-7) collapses to

$$h(mn) = h(m) + h(n) \qquad (10\text{-}8)$$

Evidently, this is solved by setting

$$h(n) = k \log n \qquad (10\text{-}9)$$

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 (10-8), and you have proved that (10-9) is the only solution. But in our case, (10-8) 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 (10-9) is no longer unique; in fact, (10-8) 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 (10-8) 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 $h(q_i)$ <u>arbitrarily</u> for the prime numbers q_i, whereupon (10-8) 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 (10-8) may 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

$$h(n^k) = k\, h(n) \tag{10-10}$$

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

$$\frac{m}{n} \leq \frac{\log t}{\log s} < \frac{m+1}{n} \tag{10-11}$$

or,

$$s^m \leq t^n < s^{m+1}$$

Since h is monotonic increasing,

$$h(s^m) \leq h(t^n) \leq h(s^{m+1})$$

or from (10-10),

$$m\, h(s) \leq n\, h(t) \leq (m+1)\, h(s)$$

which can be written as

$$\frac{m}{n} \leq \frac{h(t)}{h(s)} \leq \frac{m+1}{n} \tag{10-12}$$

Comparing (10-11), (10-12), we see that

$$\left|\frac{h(t)}{h(s)} - \frac{\log t}{\log s}\right| \leq \frac{1}{n}$$

or

$$\left|\frac{h(t)}{\log t} - \frac{h(s)}{\log s}\right| \leq \varepsilon \tag{10-13}$$

where

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

is arbitrarily small. Thus h(t)/log t must be a constant, and the uniqueness

of (10-9) is proved.

Now different choices of k 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 (10-7), 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

$$H(p_1,\ldots,p_n) = -\sum_{i=1}^{n} p_i \log p_i \qquad (10\text{-}14)$$

Accepting this interpretation, it follows that the distribution $(p_1\ldots p_n)$ which maximizes (10-14) subject to constraints imposed by the available information, will represent the "most honest" description of what we know 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.

The function H will be called the <u>entropy</u> of the distribution $(p_1,\ldots,p_n)$ from now on. It is a new measure of how uniform a probability distribution is--any change in the direction of equalizing the different probabilities will increase the entropy.

I don't think that either this demonstration or the ones we gave in the third lecture are anywhere near in satisfactory form yet. In particular, the functional equation (10-4) does not seem quite so intuitively compelling as our previous ones were. You might ask why the factor q must appear in the last term, and the only answer I can give is that if you leave it out, the solution of the functional equation will collapse to $H_n(p_1\ldots p_n) = (n-1)$, independently of the p_i, and you will lose everything we had hoped to get from this argument. In this case, I think the trouble is just that neither I nor any other writer known to me has yet learned how to verbalize the argument leading to (10-4) in a fully convincing manner. Perhaps this will inspire

you to try your hand at improving the verbiage that I used just before writing (10-4).

For this reason, it is comforting to know that there are several other possible arguments which will also lead to the same conclusion (10-14). Khinchin (1957) has given a slightly different set of conditions. They are:

(1) For given n, $H_n(p_1 \ldots p_n)$ attains its maximum value when $p_k = (1/n)$, $k = 1, 2, \ldots, n$.

(2) If we include in our enumeration a new situation which is, however, known to be impossible, our state of uncertainty is not really changed. Therefore, we should have $H_{n+1}(p_1 \ldots p_n, 0) = H_n(p_1 \ldots p_n)$.

(3) A composition law essentially equivalent to (10-4) although stated in slightly different terms.

Khinchin shows that the only continuous function satisfying these requirements is the entropy expression (10-14).

10.4. The Wallis Derivation.

Another, and quite amusing, way of deriving the maximum-entropy principle resulted from a suggestion made to me 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 $\{p_1 \ldots p_m\}$ 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 laws 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 $n >> m$, 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 a "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, the second n_2, etc., we will say that the random experiment has generated the probability assignment

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

The probability that this will happen is

$$m^{-n} \frac{n!}{n_1! \ldots n_m!}$$

Now imagine that a blindfolded friend repeatedly scatters the n quanta at random among the m possibilities. 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? It is the one which maximizes

$$W \equiv \frac{n!}{n_1! \ldots n_m!} \tag{10-15}$$

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

$$\log n! = n \log n - n + \sqrt{2\pi n} + \frac{1}{12n} + O\left(\frac{1}{n^2}\right) \tag{10-16}$$

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 easily find that as $n\to\infty$, $n_i\to\infty$, in such a way that $n_i/n \to p_i = \text{const.}$,

$$\frac{1}{n}\log W = \frac{1}{n}\left[\log n! - \sum_{i=1}^{n}\log (np_i)!\right]$$

$$\to \log n - 1 - \frac{1}{n}\sum_{i=1}^{n}\left[np_i \log (np_i) - np_i\right]$$

Since $\Sigma p_i = 1$, several terms cancel, and we are left with

$$\frac{1}{n}\log W \to -\sum_{i=1}^{n} p_i \log p_i = H(p_1 \ldots p_n) \tag{10-17}$$

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. I leave it for you 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.

This derivation is, in several respects, the best one yet produced. It is entirely independent of Shannon's functional equation (10-5); it does not require any postulates about connections between probability and frequency; nor does it suppose that the different possibilities {1 ... 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." Let me stress this point. It is a big mistake to try to read too much philosophical signifi-

cance into theorems which lead to equation (10-14). 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 really prophetic insight into the reception his work would get, tried to play it down by pointing out immediately after stating his 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." However, while granting that this is perfectly true, I would like now to try 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, Rule 4, and many frequency connections of probability are special cases of a single principle, and that statistical mechanics and communication theory are both instances of a single method of reasoning.

10.5. An Example.

First, let's test this principle and see how it would work out if we ask the robot to assign probabilities in such a way that the entropy (10-14) is maximized subject to the available information, in the simple example discussed in Sec. 10.2, 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; i.e., as in (10-1),

$$\delta\left[H(p_1 \ldots p_3) - \lambda \sum_{m=1}^{3} m p_m - \mu \sum_{m=1}^{3} p_m \right] =$$

$$= \sum_{m=1}^{3} \left[\frac{\partial H}{\partial p_m} - \lambda m - \mu \right] \delta p_m = 0.$$

Now,

$$\frac{\partial H}{\partial p_m} = - \log p_m - 1 \tag{10-18}$$

so our solution is

$$p_m = e^{-\lambda_o - \lambda m} \tag{10-19}$$

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

So the distribution which has maximum entropy, subject to a given average value, will always be in exponential form, and we have to fit the constants λ_o and λ by forcing this to agree with the constraints that the sum of the p's must be one and that the average value must be equal to the average $\overline{m}$ that we assigned. Well, the mathematics that you have to go through in order to do this is very straightforward and comes out very beautifully if you define a function

$$Z(\lambda) \equiv \sum_{m=1}^{3} e^{-\lambda m} \tag{10-20}$$

which we call the <u>partition function</u>. The equations (10-2) which fix our Lagrange multipliers then take the form

$$\lambda_o = \log Z(\lambda) \tag{10-21}$$

and

$$\overline{m} = - \frac{\partial}{\partial \lambda} \log Z(\lambda) \tag{10-22}$$

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

$$p_k = \frac{\exp[(2-k)\lambda]}{1 + 2 \cosh \lambda}, \quad k = 1, 2, 3. \tag{10-23}$$

$$\overline{m} = \frac{e^{2\lambda} + 2e^{\lambda} + 3}{e^{2\lambda} + e^{\lambda} + 1}. \tag{10-24}$$

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 algebra-

ically, leading to the explicit solution

$$p_1 = \frac{3 - \overline{m} - p_2}{2}$$

$$p_2 = \frac{1}{3}\left[\sqrt{4 - 3(\overline{m}-2)^2} - 1\right] \qquad (10\text{-}25)$$

$$p_3 = \frac{\overline{m} - 1 - p_2}{2}$$

These results are plotted in Figure (10.3). p_2 is the arc of an ellipse which comes in with unit slope at the ends. p_1 and p_3 are also arcs of ellipses, but slanted one way and the other.

Let's just notice that we have finally arrived here at a solution which meets the objections we had to the first two criteria. The maximum entropy distribution automatically has the property $p_m \geq 0$ 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

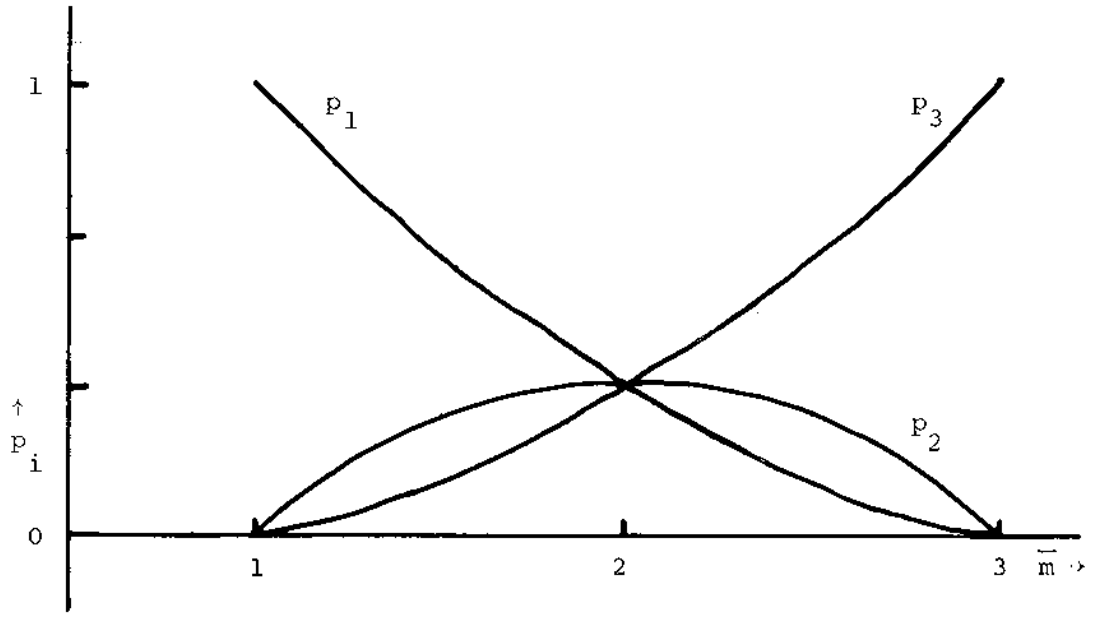

Figure 10.3. Maximum-Entropy solution.

place where a probability goes to zero is in the limit where the $\overline{m}$ is exactly one or exactly three. But of course, in that case, some probabilities did have to be zero.

We see the comparison between these two criteria is very interesting. The criterion that

$$\sum_m p_m^2 = \text{minimum}$$

gives [Fig. (10.2)] the same value and the same slope as the maximum entropy solution, at the end points and at the middle. It represents, in a sense, the best straight-line approximation you could have made to the maximum entropy solution.

10.6. 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

$$f_k(x) \quad , \qquad 1 \leq k \leq m, \quad m < n, \tag{10-26}$$

for which we know the mean values. What probabilities $(p_1 \ldots p_n)$ will the robot assign to the possibilities $(x_1 \ldots x_n)$? The average of $f_k(x)$ is supposed known for each of the possible values of k, i.e.,

$$F_k \equiv \langle f_k(x) \rangle = \sum_{i=1}^{n} p_i \, f_k(x_i) \quad , \tag{10-27}$$

and the robot will find the set of p_i's which has maximum entropy subject to all these constraints simultaneously. Let's see what he'll come out with. We just have to introduce as many Lagrange multipliers as there are constraints imposed on the problem.

$$\delta [H(p_1 \ldots p_n) - (\lambda_o - 1) \sum_i p_i - \lambda_1 \sum_i p_i \, f_1(x_i) - \ldots - \lambda_m \sum_i p_i \, f_m(x_i)]$$

$$= \sum_i \left[\frac{\partial H}{\partial p_i} - (\lambda_o - 1) - \lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i) \right] \delta p_i = 0$$

and so from (10-18) our solution is the following:

$$p_i = e^{-\lambda_o - \lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)} . \qquad (10\text{-}28)$$

That's the form of the distribution, and we still have to find how he is going to evaluate these constants. In the first place, the sum of all probabilities will have to be unity, i.e.,

$$1 = \sum_i p_i = e^{-\lambda_o} \sum_i e^{-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)} . \qquad (10\text{-}29)$$

If we now define a partition function as

$$Z(\lambda_1 \ldots \lambda_m) \equiv \sum_{i=1}^{n} e^{-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)} \qquad (10\text{-}30)$$

then (10-29) reduces to

$$\lambda_o = \log Z(\lambda_1 \ldots \lambda_m) \qquad (10\text{-}31)$$

The average value (10-27) of $f_k(x)$ is then

$$F_k = e^{-\lambda_o} \sum_i e^{-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)} f_k(x_i) ,$$

or,

$$F_k = - \frac{\partial}{\partial \lambda_k} \log Z \qquad (10\text{-}32)$$

What is the maximum value of the entropy that we get from this probability distribution? After an entropy has been maximized, I will call it S, the way physicists do, instead of H, the way information theory people do:

$$S \equiv (H)_{max} = \left(- \sum_{i=1}^{n} p_i \log p_i \right)_{max} \qquad (10\text{-}33)$$

From (10-28) we find that

$$S = \lambda_o + \lambda_1 F_1 + \ldots + \lambda_m F_m \qquad (10\text{-}34)$$

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 per cent 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 <u>is</u> the answer. Suppose we had a function like the one in Fig (10.4), and our job was to locate the maximum of it. Well, if we state it as a variational problem and set the derivative equal to 0, we'll get solutions at A, B, C, etc. And, of course, we could investigate these separately and see which one is really a minimum, which one is a maximum. But after we prove that A is a local maximum, still we have doubt as to whether it's an absolute maximum. Maybe there is some other point that is still higher. Even after we've proved that we have the highest value that can be reached by variational methods, it is still possible that the function reaches a still higher value at some cusp E 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.

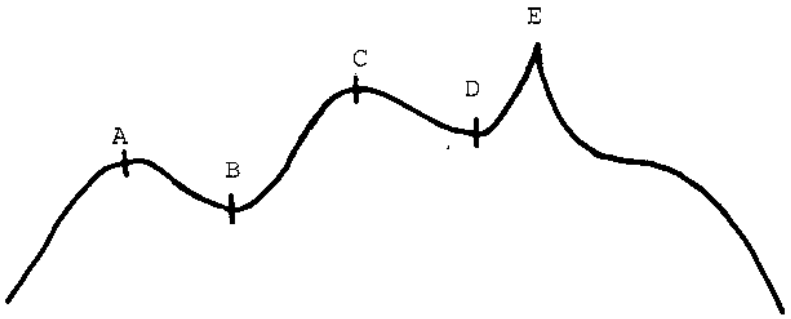

Figure 10.4.

So, I would like now to give you an entirely different derivation which is strong just where the variational argument is weak. For this I want a lemma. Let p_i be any set of numbers which could be a possible probability distribution; in other words, they add up to one and they are not negative,

$$\sum_{i=1}^{n} p_i = 1 \quad , \qquad p_i \geq 0 \tag{10-35}$$

and let u_i be another possible probability distribution,

$$\sum_{i=1}^{n} u_i = 1 \quad , \qquad u_i \geq 0. \tag{10-36}$$

Now let's think for a moment about the function log x. The graph of log x looks like this, Fig. 10.5.

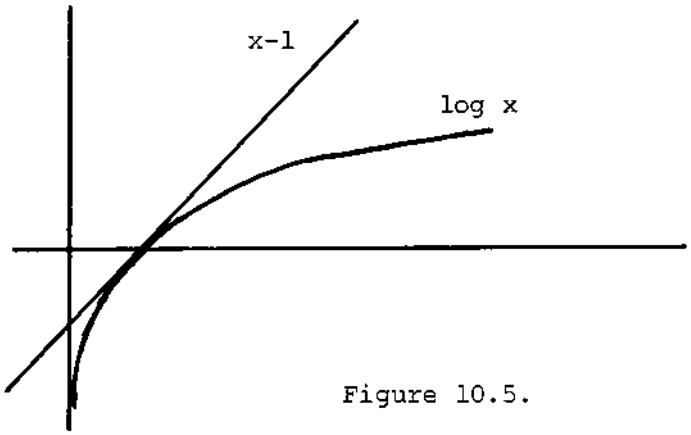

Figure 10.5.

It passes through the point (1,0) with unit slope. So if I draw a tangent to it at this point, the straight line has the equation y = x - 1. You see that log x always has curvature downward and so it stays below the tangent; therefore,

$$\log x \leq (x - 1) \quad , \qquad 0 < x < \infty \tag{10-37}$$

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

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

or,

$$H(p_1 \cdots p_n) \leq \sum_{i=1}^{n} p_i \log \left(\frac{1}{u_i}\right) \tag{10-38}$$

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

I'm going to simply pull a distribution u_i out of the hat;

$$u_i \equiv \frac{1}{Z(\lambda_1 \cdots \lambda_m)} \exp\{-\lambda_1 f_1(x_i) - \cdots - \lambda_m f_m(x_i)\} \; . \tag{10-39}$$

where $Z(\lambda_1 \cdots \lambda_m)$ is defined by (10-30). Never mind why I chose u_i this

particular way; we'll see why in a minute. But now let's play with the inequality (10-38). We can now write it as

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

or

$$H \leq \log Z + \lambda_1 \langle f_1(x) \rangle + \ldots + \lambda_m \langle f_m(x) \rangle \ . \qquad (10\text{-}40)$$

Now, let the p_i vary over the class of all possible probability distributions that satisfy the constraints (10-27) of the problem. The right-hand side of (10-40) stays constant. Our lemma now says that H attains its absolute maximum, making (10-40) an equality, if and only if the p_i are chosen as the canonical distribution (10-39).

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 I just had to pull the answer out of a hat in writing (10-39). I had to know the answer before I could prove it. If you have both arguments side by side, then you have the whole story.

10.7. Formal Properties of Maximum-Entropy Distributions.

Now I want to put down a list of the general formal properties of this canonical distribution (10-39). This is a bad way of doing it in one sense; it sounds very abstract and you don't see the connection to any physical 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 I'll go into specific physical problems and we'll see that every one of these formal relations turns out to have many different useful physical 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,

$$(H)_{max} = S(F_1 \ldots F_m) = \log Z + \sum_{k=1}^{m} \lambda_k F_k \qquad (10\text{-}41)$$

H itself we can regard as a measure of the "amount of the uncertainty" in any probability distribution. After I have maximized it, it becomes a function of the definite physical data of the problem, and I'll call it S. 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 data of the problem, and not on anybody's personality.

If S is to be only a function of $(F_1 \ldots F_m)$, then in (10-41) 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 flapping around loose, but eventually we have to find what they are. If I choose different λ_i, I am writing down different probability distributions (10-39); and we saw in (10-32) that the averages over this distribution agree with the given averages F_k if

$$F_k = \langle f_k \rangle = -\frac{\partial}{\partial \lambda_k} (\log Z) \quad , \quad k = 1, 2, \ldots, m \qquad (10\text{-}42)$$

So we are now to regard (10-42) as a set of m simultaneous equations which are to be solved for the λ_i in terms of the given data F_k; at least one would like to dream about this. Generally, when you get to non-trivial problems, this is so involved that you have to leave the λ_i where they are, and express things in parametric form. If you've got more than about two λ_i in the problem, it is generally impractical to solve for them explicitly. Actually, this isn't such a tragedy, because the λ_i usually turn out to have such important physical meanings that we are quite happy to use them as the independent variables. However, I would like to show you that if we can evaluate the function $S(F_1 \ldots F_m)$, then we can give the λ_i as explicit functions of the

given data.

Suppose I take S and differentiate it. I make a small change in one of the values F_k that we fed into the problem; how does this change the maximum attainable H? We have from (10-41),

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

which, thanks to (10-42), collapses to

$$\lambda_k = \frac{\partial S}{\partial F_k} \tag{10-43}$$

in which λ_k is given explicitly.

Compare this equation with (10-42); one gives F_k explicitly in terms of the λ_k, the othergives the λ_k explicitly in terms of the F_k. If I specify log Z as a function of the λ_k; or if I 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 (10-41) 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. Let me differentiate (10-42) with respect to λ_j:

$$\frac{\partial F_k}{\partial \lambda_j} = \frac{\partial^2}{\partial \lambda_j \partial \lambda_k} (\log Z) = \frac{\partial F_j}{\partial \lambda_k} \tag{10-44}$$

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 I differentiate (10-43) a second time, I'll have

$$\frac{\partial \lambda_k}{\partial F_j} = \frac{\partial^2 S}{\partial F_j \partial F_k} = \frac{\partial \lambda_j}{\partial F_k} \tag{10-45}$$

another reciprocity law, which is however not independent of (10-44), 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 easily see 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 non-obvious physical meanings.

Now let's consider the possibility that one of these functions $f_k(x)$ has an extra parameter α in it 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 this represents the energy of the i'th state of the spin system and α represents the magnetic field that's applied. Very often we want to make a prediction of how certain quantities change as I change α. I want to calculate the pressure; or the susceptibility. By the criterion of minimum mean square error, the best estimate I can make of that derivative would be the mean value over the probability distribution. If I write it out, it will be

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

which reduces to

$$\left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{1}{\lambda_k Z} \frac{\partial}{\partial \alpha} \sum_i \exp\{\quad\}$$

$$= -\frac{1}{\lambda_k} \frac{\partial}{\partial \alpha} \log Z \ . \qquad (10\text{-}46)$$

In this derivation, I supposed that this parameter α only shows up in one function f_k. If the same parameter shows up in several different f_k, then I'll leave it for you to verify that this generalizes to

$$\sum_{k=1}^{m} \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{\partial}{\partial \alpha} \log Z \ . \qquad (10\text{-}47)$$

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

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

$$-\frac{\partial S}{\partial \alpha} = \sum_{k=1}^{m} \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{\partial}{\partial \alpha} \log Z \qquad (10\text{-}48)$$

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 (10-48) 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 (10-41). Evidently, if there are several different parameters $\{\alpha_1, \alpha_2, \ldots, \alpha_r\}$ in the problem, a relation of the form (10-48) 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 symbols F_k, $\langle f_k \rangle$ to stand for the same <u>number</u>. They are equal because I 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 I want to emphasize that these quantities are expectation values over the canonical distribution (10-39), I'll use the notation $\langle f_k \rangle$. When I want to emphasize that they are the given data, I'll call them F_k. At the moment, I want to do the former, and so the reciprocity law (10-44) can be written equally well as

$$\frac{\partial \langle f_k \rangle}{\partial \lambda_j} = \frac{\partial \langle f_j \rangle}{\partial \lambda_k} = \frac{\partial^2}{\partial \lambda_j \partial \lambda_k} \log Z \qquad (10\text{-}48)$$

In varying the λ's here, we're changing from one canonical distribution (10-39) 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 (10-48).

But now I want to show that the quantities in (10-48) also have an important meaning with reference to a single maximum entropy problem. In the canonical distribution (10-39), 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:

$$\langle (f_j - \langle f_j \rangle)(f_k - \langle f_k \rangle) \rangle$$

$$= \langle [f_j f_k - f_j \langle f_k \rangle - \langle f_j \rangle f_k + \langle f_j \rangle \langle f_k \rangle] \rangle$$

$$= \langle f_j f_k \rangle - \langle f_j \rangle \langle f_k \rangle \qquad (10\text{-}49)$$

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:

$$\langle (f_k - \langle f_k \rangle)^2 \rangle = \langle f_k^2 \rangle - \langle f_k \rangle^2 \geq 0 \ . \qquad (10\text{-}50)$$

To calculate these quantities directly from the canonical distribution (10-39), we can first find

$$\langle f_j f_k \rangle = \frac{1}{Z(\lambda_1 \ldots \lambda_m)} \sum_{i=1}^{n} f_j(x_i) f_k(x_i) \exp\{-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)\}$$

$$= \frac{1}{Z} \sum_{i=1}^{n} \frac{\partial^2}{\partial \lambda_j \partial \lambda_k} \exp\{-\lambda_1 f_1(x_i) - \ldots - \lambda_m f_m(x_i)\}$$

$$= \frac{1}{Z}\frac{\partial^2 Z}{\partial\lambda_j\partial\lambda_k} \tag{10-51}$$

Then, using (10-42), the covariance becomes

$$\langle f_j f_k\rangle - \langle f_j\rangle\langle f_k\rangle = \frac{1}{Z}\frac{\partial^2 Z}{\partial\lambda_j\partial\lambda_k} - \frac{1}{Z^2}\frac{\partial Z}{\partial\lambda_j}\frac{\partial Z}{\partial\lambda_k}$$

$$= \frac{\partial^2}{\partial\lambda_j\partial\lambda_k}\log Z \tag{10-52}$$

But this is just the quantity (10-48); therefore the reciprocity law takes on a bigger meaning,

$$\langle f_j f_k\rangle - \langle f_j\rangle\langle f_k\rangle = -\frac{\partial\langle f_j\rangle}{\partial\lambda_k} = -\frac{\partial\langle f_k\rangle}{\partial\lambda_j} \tag{10-53}$$

That second derivative of log Z which gave us the reciprocity law also gives us the covariance of f_j and f_k in our distribution.

Note that (10-53) 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 easily verify,

$$\langle q f_k\rangle - \langle q\rangle\langle f_k\rangle = -\frac{\partial\langle q\rangle}{\partial\lambda_k} \quad , \tag{10-54}$$

a relation that I hadn't noticed in several years of using this formalism, until it was pointed out to me by my former student, Dr. Baldwin Robertson.

From comparing (10-42), (10-48), (10-53) we might expect that still higher derivatives of log Z would correspond to higher moments of the distribution (10-39). This is easily checked; for the third central moments of the f_k we have

$$\langle (f_j - \langle f_j\rangle)(f_k - \langle f_k\rangle)(f_r - \langle f_r\rangle)\rangle$$

$$= \langle f_j f_k f_r\rangle - \langle f_j\rangle\langle f_k f_r\rangle - \langle f_k\rangle\langle f_j f_r\rangle - \langle f_r\rangle\langle f_j f_k\rangle + 2\langle f_j\rangle\langle f_k\rangle\langle f_r\rangle$$

$$= -\frac{\partial^3}{\partial\lambda_j\partial\lambda_k\partial\lambda_r}\log Z \tag{10-55}$$

and in general, all the central moments are given by

$$\langle (f_i - \langle f_i \rangle)^{m_i} (f_j - \langle f_j \rangle)^{m_j} \ldots \rangle$$

$$= (-)^{m_i+m_j+\cdots} \left(\frac{\partial^{m_i}}{\partial \lambda_i^{m_i}} \frac{\partial^{m_j}}{\partial \lambda_j^{m_j}} \cdots \right) \log Z \qquad (10\text{-}56)$$

For noncentral moments, it is customary to define a <u>moment generating function</u>

$$\phi(\beta_1 \ldots \beta_m) \equiv \langle \exp[\beta_1 f_1 + \ldots + \beta_m f_m] \rangle \qquad (10\text{-}57)$$

which evidently has the property

$$\langle f_i^{m_i} f_j^{m_j} \ldots \rangle = \left(\frac{\partial^{m_i}}{\partial \beta_i^{m_i}} \frac{\partial^{m_j}}{\partial \beta_j^{m_j}} \cdots \right) \phi(\beta_1 \ldots \beta_m) \Big|_{\beta_k \neq 0} \qquad (10\text{-}58)$$

However, we find from (10-57)

$$\phi(\beta_1 \ldots \beta_m) = \frac{Z[(\lambda_1 - \beta_1), \ldots, (\lambda_m - \beta_m)]}{Z(\lambda_1 \ldots \lambda_m)} \qquad (10\text{-}59)$$

so that the partition function Z serves this purpose; instead of (10-58) we may write equally well,

$$\langle f_i^{m_i} f_j^{m_j} \ldots \rangle = \frac{1}{Z} \left(\frac{\partial^{m_i}}{\partial \lambda_i^{m_i}} \frac{\partial^{m_j}}{\partial \lambda_j^{m_j}} \cdots \right) Z \qquad (10\text{-}60)$$

which is the generalization of (10-51).

Now, we might ask, what are the covariances of the derivatives of f_k with respect to a parameter α? Let's define

$$g_k \equiv \frac{\partial f_k}{\partial \alpha} \quad , \qquad (10\text{-}61)$$

if f_k is the energy and α is the volume then $-g_k$ is the pressure. The law

for the fluctuation of these is, by a similar derivation that I'll leave for you to work out,

$$\sum_{j=1}^{m} \lambda_j [\langle g_j g_k \rangle - \langle g_j \rangle \langle g_k \rangle] = \left\langle \frac{\partial g_k}{\partial \alpha} \right\rangle - \frac{\partial \langle g_k \rangle}{\partial \alpha} \tag{10-62}$$

a very interesting thing. I had found and used special cases of this for some time, before I finally realized it's actually completely general.

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 (10-62) is

$$\frac{\partial^2 \log Z}{\partial \alpha^2} = \sum_{jk} \lambda_j \lambda_k [\langle g_j g_k \rangle - \langle g_j \rangle \langle g_k \rangle] - \sum_k \lambda_k \left\langle \frac{\partial g_k}{\partial \alpha} \right\rangle \tag{10-63}$$

The cross-derivatives give us a simple and useful relation

$$\frac{\partial^2 \log Z}{\partial \alpha \partial \lambda_k} = -\frac{\partial \langle f_k \rangle}{\partial \alpha}$$

$$= \sum_j \lambda_j [\langle f_k g_j \rangle - \langle f_k \rangle \langle g_j \rangle] - \langle g_k \rangle \tag{10-64}$$

which also follows from (10-48) and (10-54); and by taking further derivatives an infinite hierarchy of similar moment relations is obtained. As we will see later, the above theorems have many applications in calculating the fluctuations in pressure of a gas or liquid, the voltage fluctuations, or "noise" generated by a reversible electric cell, etc.

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

$$\frac{\partial^2 \log Z}{\partial \alpha_1 \partial \alpha_2} = \sum_k \lambda_k \left\langle \frac{\partial^2 f_k}{\partial \alpha_1 \partial \alpha_2} \right\rangle - \sum_{kj} \lambda_j \lambda_k \left[\left\langle \frac{\partial f_k}{\partial \alpha_1} \frac{\partial f_j}{\partial \alpha_2} \right\rangle - \left\langle \frac{\partial f_k}{\partial \alpha_1} \right\rangle \left\langle \frac{\partial f_j}{\partial \alpha_2} \right\rangle \right] \tag{10-65}$$

will appear.

Well, these moment theorems are quite numerous, but easy to derive. Because of the relation (10-41) between $\log Z(\lambda_1 \ldots \lambda_m; \alpha_1 \ldots \alpha_m)$ and

$S(\langle f_1\rangle \ldots \langle f_m\rangle ; \alpha_1 \ldots \alpha_r)$, you can see 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 that I want to develop.

In (10-43) we supposed that the definitions of the functions $f_k(x)$ were fixed once and for all, the variation in $\langle f_k\rangle$ 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

$$\delta \log Z = \frac{1}{Z} \sum_{i=1}^{n} \Big\{ \sum_{k=1}^{m} [-\lambda_k \delta f_k(x_i) - \delta\lambda_k f_k(x_i)] \cdot \exp\Big[-\sum_{j=1}^{m} \lambda_j f_j(x_i)\Big] \Big\}$$

$$= -\sum_{k=1}^{m} [\lambda_k \langle \delta f_k\rangle + \delta\lambda_k \langle f_k\rangle] \qquad (10\text{-}66)$$

and thus from the Legendre transformation (10-41)

$$\delta S = -\sum_k \lambda_k [\delta\langle f_k\rangle - \langle \delta f_k\rangle]$$

or,

$$\delta S = \sum_k \lambda_k Q_k \qquad (10\text{-}67)$$

where

$$\delta Q_k \equiv \delta\langle f_k\rangle - \langle \delta f_k\rangle$$

$$= \sum_{i=1}^{n} f_k(x_i)\ \delta p_i \qquad (10\text{-}68)$$

This result, which generalizes (10-43), shows that the entropy S is stationary not only in the sense of the maximization property which led to the canonical distribution (10-39); 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 (10-67), suppose that the functions f_k contain parameters $\{\alpha_1 \ldots \alpha_r\}$ as in (10-65), which generate the $\delta f_k(x_i)$ by

$$\delta f_k(x_i,\alpha_j) = \sum_{j=1}^{r} \frac{\partial f_k(x_i,\alpha)}{\partial \alpha_j}\ \delta\alpha_j \quad . \qquad (10\text{-}69)$$

While δQ_k is not in general the exact differential of any function $Q_k(\langle f_1\rangle \ldots \langle f_m\rangle;\alpha_1 \ldots \alpha_r)$, Eq. (10-67) shows that λ_k is an integrating factor such that $\sum \lambda_k \delta Q_k$ is the exact differential of a "state function" $S(\langle f_1\rangle \ldots \langle f_m\rangle;\alpha_1 \ldots \alpha_r)$. At this point, perhaps all this is beginning to sound vaguely familiar.

Finally, I leave it for you to prove from (10-67) that

$$\sum_{k=1}^{m} \langle f_k\rangle \frac{\partial \lambda_k}{\partial \alpha} = 0 \qquad (10\text{-}70)$$

where $\langle f_1\rangle \ldots \langle f_m\rangle$ 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 he can solve in rather a wholesale way. He first evaluates this partition function Z, or better still, log Z. Then just by differentiating that with respect to everything in sight, he obtains all sorts of predictions in the form of mean values. This is quite a neat mathematical procedure, and, of course, you recognize what we have been doing here. These equations are all just the standard equations of statistical mechanics, in a disembodied form with all the physics removed. In the next lecture, we'll examine that application; but from the way we derived it, it's already clear that this same mathematics also has a lot of other applications outside of physics.

10.8. 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, I 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 (10-39) 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 cannot lead to any definite physical results because different people have different information, which would lead to different distributions--so the results are basically arbitrary. (D) 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 non-zero probabilities to values of $f_k \neq F_k$.

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 for honestly recognizing 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 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 $\theta_o(1 \pm 10^{-6})$ 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 tells us just what conclusions we are entitled to draw about θ, on the basis of the information which was put into the equations.

Now to answer objection (B), I want to show that the situation is vastly more subtle than that. The principle of maximum entropy has, fundamentally, nothing to do with any "random experiment," and some of the most important applications are to cases where the probabilities p_i in (10-39) have no frequency connection for just that reason--the x_i are simply an enumeration of the possibilities, and there are no "random variables" in the problem. However, nothing prevents us from applying the principle of maximum entropy also to those cases where the x_i may be regarded as produced by some random

experiment; and in this case, the question of the relation between maximum-entropy probabilities and observable frequencies is capable of mathematical analysis.

I want to give you this analysis now, and 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 empirical frequency distribution does not agree with the maximum-entropy probability distribution.

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, $i = 1, 2, \ldots, n$. 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 will occur in the long run?

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 this consists of average values of serveral 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 result x_1 will be obtained n_1 times, x_2 will be obtained n_2 times, etc. Of course,

$$\sum_{i=1}^{n} n_i = N \tag{10-71}$$

and if the specified mean values F_k are in fact obtained, we have the additional relations

$$\sum_{i=1}^{n} n_i f_k(x_i) = NF_k \quad , \qquad k = 1, 2, \ldots, m \tag{10-72}$$

If $m < n-1$, the relations (10-71), (10-72) are insufficient to determine the relative frequencies $f_i = n_i/N$. Nevertheless, we do have good and strong 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

$$W = \frac{N!}{n_1!n_2!\ldots n_n!} = \frac{N!}{(Nf_1)!(Nf_2)!\ldots(Nf_n)!} \tag{10-73}$$

The set of frequencies $\{f_1 \ldots f_n\}$ which can be realized in the greatest number of ways is therefore the one which maximizes W subject to the constraints (10-71), (10-72). Now you see it coming--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 already saw in (10-17),

$$\frac{1}{N} \log W \to -\sum_{i=1}^{n} f_i \log f_i = H_f \tag{10-74}$$

So you see that, in (10-71), (10-72), (10-74) we have formulated exactly the same mathematical problem as in the maximum-entropy derivation of Sec. (10.6), so the two problems will have the same solution. This derivation is mathematically very reminiscent of the Wallis derivation that I gave you a few minutes ago, but of course the equations now have an entirely different meaning.

You also see 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 value--and I want to say more about that in

a moment--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 [i.e. a set which satisfies the constraints (10-71), (10-72)] 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 (10-74), as

$$\frac{W}{W'} \to \exp\{N(H_f - H_{f'})\} \tag{10-75}$$

and passes all bounds as $N \to \infty$. Therefore, the distribution predicted by maximum entropy can be realized experimentally in overwhelmingly more ways than can any other.

We have here another precise and quite general connection between probability and frequency; once again, it had nothing to do with the definition of probability, but emerged as a mathematical consequence of probability theory, interpreted as the "calculus of inductive reasoning." Two more kinds of connection between probability and frequency, whose precise mathematical statements are different in form, but which have the same practical consequences, will appear later, in lectures 12 and 17.

Now let's turn to objection (C) and analyze the situation there. Does this connection between probability and frequency justify our predicting that the maximum-entropy distribution will in fact be observed in a real random experiment? Clearly not, in the sense of deductive proof; for just as objection (C) points out, we have to concede that different people may

have different amounts of information, which will lead them to writing down different distributions, and they can't all be right. But let's look at this more closely. Consider a specific case: Mr. A knows the mean values $\langle f_1(x)\rangle$, $\langle f_2(x)\rangle$. Mr. B knows in addition $\langle f_3(x)\rangle$. 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

$$H_B \leq H_A \tag{10-76}$$

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's 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 (10-76).

From this example 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.

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 with this property. How

does our robot tell us this? Mathematically, you will then find that the equations

$$F_k = -\frac{\partial}{\partial \lambda_k} \log Z(\lambda_1,\lambda_2,\lambda_3) \tag{10-77}$$

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

$$\begin{aligned} Z(\lambda_1,\lambda_2,\lambda_3) &= \sum_{i=1}^{n} \exp[-\lambda_1 f_1(x_i) - \lambda_2 f_2(x_i) - \lambda_3 f_3(x_i)] \\ &= \sum_{i=1}^{n} \exp[-(\lambda_1+\lambda_3) f_1(x_i) - (\lambda_2+2\lambda_3) f_2(x_i)] \end{aligned} \tag{10-78}$$

and so

$$\frac{\partial Z}{\partial \lambda_3} = \frac{\partial Z}{\partial \lambda_1} + 2\,\frac{\partial Z}{\partial \lambda_2} \tag{10-79}$$

and so (10-77) 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, giving us no 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 (10-76), 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 (10-75) we have

$$\frac{W_A}{W_B} \sim \exp\{N(H_A - H_B)\} \tag{10-80}$$

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 random 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 $f_4(x)$, $f_5(x)$, ..., etc. operating in the experiment but 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 operative in the random experiment, then the distribution predicted by maximum entropy is overwhelmingly the most likely to be observed experimentally, because it can be realized in overwhelmingly the greatest number of ways.

Conversely, if the experiment fails to confirm the maximum-entropy prediction, and this disagreement persists on indefinite repetition of the experiment, then we will conclude that the physical mechanism of the experiment must contain additional constraints which were not taken into account in the maximum-entropy calculation. The observed deviations then provide a clue as to the nature of these new constraints. In this way, Mr. A can discover empirically that his information was incomplete.

Now 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 classical phase volume, based on a specified mean value of the energy) failed to predict thermodynamic properties (heat capacities, equations of state, equilibrium constants, etc.) 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 he drew the conclusion that the microscopic laws of physics must involve additional

constraints not contained in the laws of classical mechanics. Unfortunately, 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 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.

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

$$w(q_1 \cdots q_n; p_1 \cdots p_n) = \exp[\eta(q_1 \cdots p_n)] \qquad (10\text{-}81)$$

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 $\bar{\eta}$ ...." This is, of course, just what we would describe today as maximizing the entropy subject to constraints.

Unfortunately, Gibbs did not give any clear explanation, and we can only conjecture whether he possessed one, as to why this particular function is to be minimized on the average, in preference to all others. Consequently, his procedure appeared arbitrary to many, and for sixty years there was

controversy over the validity and justification of Gibbs' method. In spite of its enormous practical success when adapted to quantum statistics, few attempts were made to extend it beyond problems of thermal equilibrium.

It was not until the work of Shannon in our own time that the full significance of Gibbs' method could be appreciated. Once we had Shannon's theorem establishing the uniqueness of entropy as an "information measure," it was clear that Gibbs' procedure was an example of a general method for inductive inference, whose applicability is in no way restricted to equilibrium thermodynamics or to physics.