BRANDEIS UNIVERSITY SUMMER INSTITUTE
LECTURES IN THEORETICAL PHYSICS
K. W. Ford, ***Editor***

1960. Lectures

C. Miller • P. T. Matthews • J. Schwinger • N. Fukuda • J. J. Sakurai

1961. Lectures

Vol. 1 R. J. Eden • J. C. Polkinghorne • G. Källén • J. J.Sakurai

Vol. 2 M. E. Rose • E. C. G. Sudarshan

1962. Lectures

Vol. 1—Elementary Particle Physics and Field Theory T. Fulton • G. Kallen • J. D. Jackson • C. Fronsdal

Vol. 2—Astrophysics and ihe Many-Uody Problem E. N. Parker • J. S. Goldstein • A. A. Maiadudin • V. Anibcgaokar

Vol. 3—Statistical Physics G. E. Uhlenbeck • N. Rosenzwcig • A. J. F Siegert • E. T. Jaynes • S. Fujita

Brandeis Summer Institute 1962
STATISTICAL
PHYSICS
3
G. E. Uhlenbeck N. Hosenzweig A. J, F. Siegert E. T. Jaynes S. Fujita
W. A. BENJAMIN, INC.

STATISTICAL PHYSICS

1962 Brandeis Lectures in Theoretical Physics, Volume 3

G. E. Uhlenbeck, N. Rosenzweig, A. J. F. Siegert, E. T. Jaynes, and C. Fujita.

In his course on SELECTED TOPICS IN STATISTICAL MECHANICS, Professor G. E. Uhlenbeck begins with an exposition of some diagrammatic methods used to calculate virial coefficients and the equation of state. He then gives a detailed analysis of the mathematics of phase transition with a soluble one-dimensional model.

The second set of lectures, by Dr. N. Rosenzweig, STATISTICAL MECHANICS OF EQUALLY LIKELY QUANTUM SYSTEMS, is a discussion of the statistical properties of energy levels and eigenfunctions for heavy nuclei and complex atoms, stressing the role of time reversal and other symmetries.

Professor A. J. F. Siegert lectures on FUNCTIONAL INTEGRALS IN STATISTICAL MECHANICS, demonstrating the utility of new techniques by analysis of the partition function of the Ising model with long range interactions.

Information theory has provided the long hoped for algorithm analogous to the partition sum of equilibrium theory, for calculation of irreversible processes. The lectures of Professor E. T. Jaynes, INFORMATION THEORY AND STATISTICAL MECHANICS, provide an introduction to this subject.

In the final set of lectures, Professor C. Fujita reviews and compares the independent achievements of Van Hove and Prigogine and their schools, in their progress toward better understanding the APPROACH TO EQUILIBRIUM OF A MANY-PARTICLE SYSTEM.

Foreword

It is now an established tradition of the Brandeis Summer Institute in Theoretical Physics to have lecturers who present a systematic account of recent research in various fields of theoretical physics. The lecture notes have also become a part of tins tradition, and, although these are sometimes but a first approximation to the spoken lecture, they may serve to bring these much needed expositions to the wider audience of physicists who may aspire to contribute to these fields.

I should like to take this opportunity to thank all those whose participation in the Institute during the summer of 1962 helped maintain these traditions. Particular words of appreciation are due the National Science Foundation, for its indispensable financial support, and Professor Kenneth Ford, who graciously carried the responsibility for getting the notes ready for publication.

In this volume, the notes of Professor Jaynes and Professor Fujita have been prepared by the lecturers; Professor Uhlenbeck, Dr. Rosenzweig, and Professor Siegert have kindly checked over the notes based on their lectures.

David L. Falkoff Co-Director of the ***1962*** Institute

Information Theory and Statistical Mechanics

E. T. Jaynes
Washington University

July 1962

Notes by the lecturer

Contents

1 Introduction

At the beginning of every problem in probability theory, there arises a need to assign some initial probability distribution; or what is the same thing, to "set up an ensemble." This is a problem which cannot be evaded, and for which the laws of physics give us no help. For example, the laws of physics tell us that a density matrix $\rho(t)$ must vary with time according to $\dot{\rho} = [H, \rho]$, but they do not tell us what function $\rho(0)$ should be put in at the start. Assignment of $\rho(0)$ is, of course, a matter of free choice on our part—it is for us to say which problem we want to solve.

The assignment of initial probabilities must, in order to be useful, agree with the initial information we have (i.e., the results of measurements of certain parameters). For example, we might know that at time $t = 0$, a nuclear spin system having total (measured) magnetic moment $M(0)$, is placed in a magnetic field H, and the problem is to predict the subsequent variation $M(t)$, which presumably tends to an equilibrium value $M(\infty) = x_0 H$ after a long time. What initial density matrix for the spin system $\rho(0)$, should we use? Evidently, we shall want it to satisfy, at the very least,

$$\mathrm{Tr}\left(\rho(0) M_{o_p}\right) = M(0) \tag{1}$$

where M_{o_p} is the operator corresponding to total magnetic moment. But Eq. (1) is very far from uniquely specifying $\rho(0)$. Out of the infinite number of density matrices satisfying (1), which should we choose as the starting point of our calculation to predict $M(t)$?

Conventional quantum theory has provided an answer to the problem of setting up initial state descriptions only in the limiting case where measurements of a "complete set of commuting observables" have been made, the density matrix $\rho(0)$ then reducing to the projection operator onto a pure state $\psi(0)$ which is the appropriate simultaneous eigenstate of all the measured quantities. But there is almost no experimental situation in which we really have all this information, and before we have a theory able to treat actual experimental situations, existing quantum theory must be supplemented with some principle that tells us how to translate, or encode, the results of measurements into a definite state description $\rho(0)$. Note that the problem is not to find the $\rho(0)$ which correctly describes the "true physical situation." That is unknown, and always remains so, because of incomplete information. In order to have a usable theory we must ask the much more modest question: "What $\rho(0)$ best describes our *state of knowledge* about the physical situation?"

In order to emphasize that this problem really has nothing to do with the laws of physics (and, as a corollary, that its solution will have applications outside the field of physics), consider the following problem. A die has been tossed a very large number N of times, and we are told that the *average* number of spots up per toss was not 3.5, as we might expect from an honest die, but 4.5. Translate this information into a probability assignment $P_n, n = 1,\ 2,\ \ldots, 6$, for the nth face to come up on the next toss.

To explain more fully what is meant by this, note that we are *not* asking for an estimate of the fraction (i.e., the relative frequency) of tosses which give n spots There is, indeed, a connection between the probability

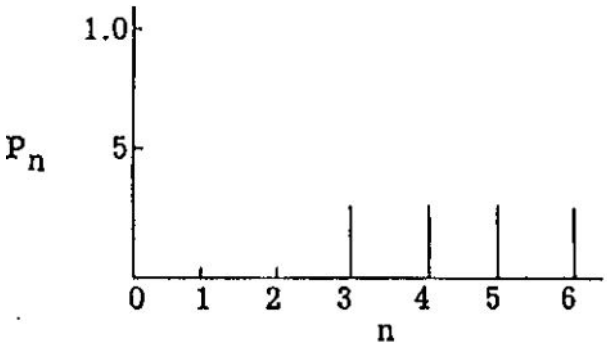

Figure 1:

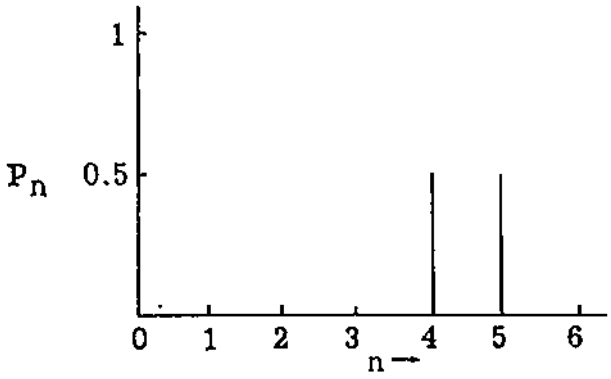

Figure 2:

and the frequency, which we will derive later. But the problem stated is to reason as best we can about the *individual* case. The probability P_n must therefore be interpreted in the so-called "subjective" sense; it is only a means of describing how strongly we *believe* that the n-th face will come up in the next toss.

To state the problem more drastically, imagine that we are offered several bets, at various odds, on various values of n, and we are compelled to accept one of these bets. The probabilities P_n are the basic raw material from which we decide which one to accept. This is typical of many practical problems faced by the scientist, the engineer, the statistician, the politician; and indeed all of us. We are continually faced with situations where some definite decision must be made *now*, even though we do not have all the information we might like.

Conventional probability theory does not provide any principle for assigning the probabilities P_n; so let us think about it a little. We must evidently, choose P_n such that

$$\sum_{n=1}^{6} P_n = 1 \tag{2}$$

$$\sum_{n=1}^{6} nP_n = 4.5 \tag{3}$$

where (3) is analogous to (1). A possible solution of (2) and (3) is indicated in Fig. 1; we could take $P_4 = P_5 = 1/2$, all other $P_n = 0$, This agrees with all the given data. But our common sense tells us it is not a *reasonable* assignment. The assignment of Fig. 2 is evidently a more honest description of what we know. But even this is not reasonable—nothing in the data tells us that n = 1, 2 are impossible events. In Fig. 2, we are still jumping to conclusions not warranted by the available evidence. Evidently, it is unreasonable to assign probability zero to *any* situation unless our data really rules out that case. If we assign $P_1 > 0$, $P_2 > 0$, then in order to keep the average at 4.5, we shall have to give some increased weight to the cases

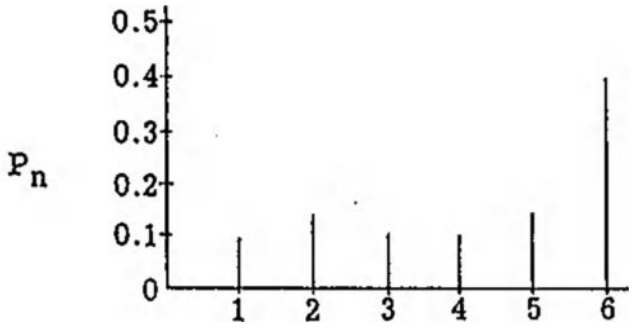

Figure 3:

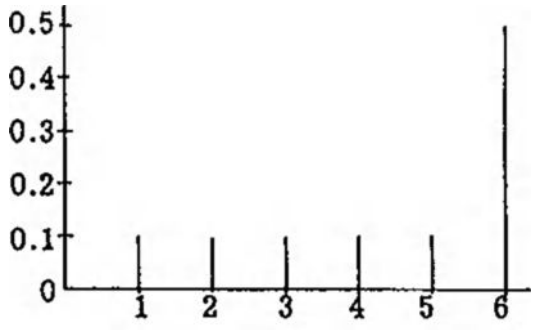

Figure 4:

$n = 5,\ 6$. Figure 3 shows an assignment that agrees with the data and does not ignore any possibility. But it still seems unreasonable to give the case $n = 6$ such exceptional treatment. Figure 4 represents what we should probably call a backward step—nothing in the data of the problem indicates any reason for such an uneven treatment. A reasonable assignment P_n must not only agree with the data and must not ignore any possibility—but it must also not give undue emphasis to any possibility. The P_n should vary as smoothly as possible, in some sense. One criterion of "smoothness" might be that adjacent differences $P_{n+1} - P_n$ should be constant; and, indeed, there is a solution with that property. It is given by $P_n = (12n - 7)/210$ and is shown in Fig. 5. This is evidently the most reasonable probability assignment so far. But there is a limit to how high an average you can get with this linear variation of P_n. If we took the extreme case, $P_n = (\text{const.})(n - 1)$, we should again violate one of our principles because $P_1 = 0$, and the average would be only $\sum_n P_n = 70/15 = 4.67$. Suppose the data of the problem had been changed so that the average is to be 4.7 instead of 4.5. Then there is *no* straight-line solution satisfying $P_n \geq 0$. The P_n must lie on some concave curve, as in Fig. 6. But the principles by which we reason surely are the same whether the data specify 4.5 or 4.7; so it appears that a result qualitatively such as Fig. 6 should be used also when $n = 4.5$.

This is about as far as qualitative reasoning can take us, and I have carried the argument through on that basis in order to show how ordinary common sense leads us to a result that has all the important features of the quantitative solution given below. The probability assignment P_n which most honestly describes what we know is the one that is as smooth and "spread out" as possible subject to the data. It is the most conservative assignment in the sense that it does not permit one to draw any conclusions not warranted by the data.

This suggests that the problem is a variational one; we need a measure of the "spread" of a probability distribution which we can maximize, subject to constraints which represent the available information. It is by now amply demonstrated by many workers that the "information measure" introduced by Shannon [1] has special properties of consistency and uniqueness which make it *the* correct measure of "amount of uncertainty" in a probability distribution. This is, of course, the expression

$$S_I = -\sum_i p_i \log p_i \tag{4}$$

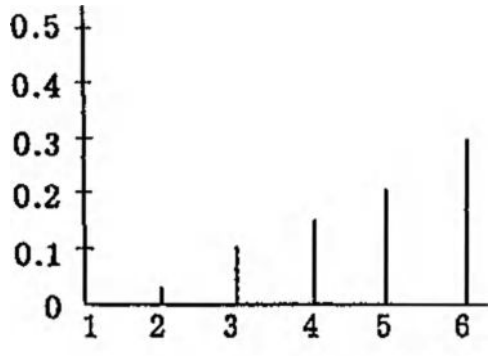

Figure 5:

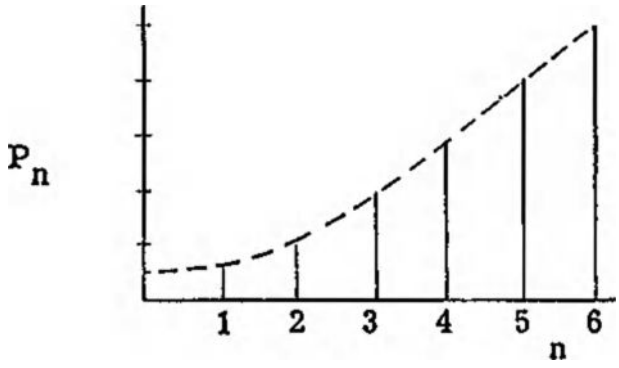

Figure 6:

which, *for some distributions* and *in some physical situations*, has long been recognized as representing entropy. However, we have to emphasize that "information-theory entropy" S_I and the experimental thermodynamic entropy S_e are entirely different concepts. Our job cannot be to *postulate* any relation between them; it is rather to *deduce* whatever relations we can from known mathematical and physical facts. Confusion about the relation between entropy and probability has been one of the main stumbling blocks in developing a general theory of irreversibility.

2 The General Maximum-Entropy Formalism

To generalize the above problem somewhat, suppose that the quantity x can take on the values $(x_1, x_2, \ldots x_n)$ where n can be finite or infinite, and that the average values of several functions $(f_1(x), f_2(x), \ldots f_m(x)$ are given, where $m < n$. The problem is to find the probability assignment $p_i = p(x_i)$ which satisfies the given data: $p_i \geq 0$,

$$\sum_{i=1}^{n} p_i = 1 \tag{5}$$

$$\sum_{i=1}^{n} p_i f_k\left(x_i\right) = \langle f_k(x) \rangle = F_k \quad k = 1, 2, \ldots, m \tag{6}$$

and, subject to (5) and (6), maximizes the entropy

$$S_I = -\sum_{i=1}^{n} p_i \log p_1 \tag{7}$$

The solution to this mathematical problem can be found immediately by the method of Lagrangian multipliers, and special cases are given in every statistical mechanics textbook. This method has the merit that it

leads immediately to the answer, but the weakness that it does not make it obvious whether one obtains a true absolute maximum of S_I. The following argument establishes this important result more rigorously.

Let $(p_1 \dots p_n)$ and $(u_1 \dots u_n)$ be any two possible probability distributions over the x_1; i.e., $p_i \geq 0, u_i \geq 0, i = 1, 2, \dots$ n and

$$\sum_{i=1}^{n} p_1 = \sum_{i=1}^{n} u_i = 1 \tag{8}$$

Then, by using the fact that $\log x \geq \left(1 - x^{-1}\right)$, with equality if and only if $x = 1$, we find the following:

Lemma

$$\sum_{i=1}^{n} p_i \log \frac{p_i}{u_i} \geq \sum_{i=1}^{n} p_i \left(1 - \frac{u_i}{p_i}\right) = 0 \tag{9}$$

with equality if and only if $p_i = u_i, i = 1, 2, \dots n$. Now make the choice

$$u_i = \frac{1}{Z(\lambda_1 \dots \lambda_m)} \exp\left(-\lambda_1 f_1(x_i) - \dots - \lambda_m f_m(x_i)\right) \tag{10}$$

where $\lambda_1 \dots \lambda_m$ are fixed constants, and

$$Z(\lambda_1 \dots \lambda_m) \equiv \sum_{i=1}^{n} \exp\left(-\lambda_1 f_1(x_1) - \dots - \lambda_m f_m(x_i)\right) \tag{11}$$

will be called the "partition function." Substituting (10) into (9) results in the inequality

$$\sum_{i=1}^{n} p_i \log p_i \geq \sum_{i=1}^{n} p_i \log u_i = -\sum_{i=1}^{n} p_i \left[\lambda_1 f_1(x_i) + \dots + \lambda_m f_m(x_i)\right] - \log Z(\lambda_1 \dots \lambda_m)$$

or

$$S_I \leq \log Z(\lambda_1 \dots \lambda_m) + \sum_{k=1}^{m} \lambda_k \langle f_k \rangle \tag{12}$$

Now let the distribution p_i vary over the class of all possible distributions that satisfy (6). The right-hand side of (12) remains fixed, and (12) shows that S_I attains its maximum possible value

$$(S_I)_{\max} = \log Z + \sum_{k=1}^{m} \lambda_k \langle f_k \rangle \tag{13}$$

if and only if p_i is taken as the generalized canonical distribution (10). It only remains to choose the unspecified constants λ_k so that (6) is satisfied. This is the case, as one readily verifies, if the λ_k are determined in terms of the given data $F_k = \langle f_k \rangle$ by

$$\langle f_k \rangle = -\frac{\partial}{\partial \lambda_k} \log Z(\lambda_1 \dots \lambda_m) \qquad k = 1, 2, \dots, m \tag{14}$$

We now survey rapidly the main formal properties of the distribution found. The maximum attainable entropy (13) is some function of the given data:

$$(S_I)_{\max} = S(\langle f_1 \rangle, \dots \langle f_m \rangle) \tag{15}$$

and, by using (13) and (14), we find

$$\frac{\partial S}{\partial \langle f_k \rangle} = \lambda_k \quad k = 1, 2, \dots, m \tag{16}$$

Regarding, in (14), the $\langle f_k \rangle$ expressed as functions of $(\lambda_1 \dots \lambda_m)$ we find, on differentiating, the reciprocity law

$$\frac{\partial \langle f_k \rangle}{\partial \lambda_j} = \frac{\partial \langle f_j \rangle}{\partial \lambda_k} = -\frac{\partial^2}{\partial \lambda_k \partial \lambda_j} \log Z = A_{jk} \tag{17}$$

while by the same argument, if we regard λ_k in (16) expressed as a function of $\langle f_1 \rangle \dots \langle f_m \rangle$, we find a corresponding law

$$\frac{\partial \lambda_k}{\partial \langle f_j \rangle} = \frac{\partial \lambda_j}{\partial \langle f_k \rangle} = \frac{\partial^2 S}{\partial \langle f_j \rangle \partial \langle f_k \rangle} = B_{jk} \tag{18}$$

Comparing (17) and (18) and remembering the chain rule for differentiating,

$$\frac{\partial \langle f_j \rangle}{\partial \langle f_k \rangle} = \sum_\ell \frac{\partial \langle f_j \rangle}{\partial \lambda_\ell} \frac{\partial \lambda_\ell}{\partial \langle f_k \rangle} = \delta_{jk}$$

we see that the second derivatives of S and of $\log Z$ yield inverse matrices:

$$A = B^{-1} \tag{19}$$

The functions $\log Z(\lambda_1 \dots \lambda_n)$ and $S(\langle f_1 \rangle \dots \langle f_n \rangle)$ are equivalent in the sense that each gives full information about the probability distribution; indeed (13) is just the Legendre transformation that takes us from one representative function to the other.

The reciprocity law (17) acquires a deeper meaning when we consider the "fluctuations" in our probability distribution. Using the distribution (10), a short calculation shows that the second central moments of the distribution of the $f_k(x)$ are given by

$$\begin{aligned} \langle (f_k - \langle f_k \rangle)(f_\ell - \langle f_\ell \rangle) \rangle &= \langle f_k f_\ell \rangle - \langle f_k \rangle \langle f_\ell \rangle \\ &= \frac{\partial^2}{\partial \lambda_k \partial \lambda_\ell} \log Z \end{aligned} \tag{20}$$

and so, comparing with (17) , there is a universal relation between the "fluctuations" of the f_k and the "compliance coefficients" $\partial \langle f_k \rangle / \partial \lambda_\ell$:

$$\langle f_k f_l \rangle - \langle f_k \rangle \langle f_\ell \rangle = -\frac{\partial \langle f_k \rangle}{\partial \lambda_\ell} = -\frac{\partial \langle f_\ell \rangle}{\partial \lambda_k} \tag{21}$$

Likewise, higher derivatives of $\log Z(\lambda_1 \dots \lambda_n)$ yield higher central moments of the f_k, in a manner analogous to (20), and a hierarchy of fluctuation laws similar to (21).

In addition to their dependence on x, the functions f_k may depend on another parameter, α. The partition function will then also have an explicit dependence on α:

$$Z(\lambda_1 \dots \lambda_m; \alpha) \equiv \sum_{i=1}^{n} \exp\left(-\lambda_1 f_1(x_i; \alpha) - \dots - \lambda_m f_m(x_i; \alpha)\right) \tag{22}$$

and a short calculation shows that the expected derivatives

$$\left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle$$

satisfy the relations

$$\sum_{k=1}^{m} \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{\partial}{\partial \alpha} \log Z = -\frac{\partial S}{\partial \alpha} \tag{23}$$

If several parameters $\alpha_1 \dots \alpha_r$ are present, a relation of this form will hold for each of them.

Finally, we note an important variational property which generalizes (16) to the case where we have also variations in the parameters $\alpha_1 \dots \alpha_r$. Let $Z = Z(\lambda_1 \dots \lambda_m; \alpha_1 \dots \alpha_r)$, and consider an arbitrary small

change in the problem, where the given data $\langle f_k \rangle$ and the parameters α_j are changed by small amounts $\delta \langle f_k \rangle / \delta \alpha_j$ This will lead to a change $\delta \lambda_k$ in λ_k. From (13), the maximum attainable entropy is changed by

$$\begin{aligned} \delta S = & \sum_{k=1}^{m} \frac{\partial \log Z}{\partial \lambda_k} \delta \lambda_k + \sum_{j=1}^{r} \frac{\partial \log Z}{\partial \alpha_j} \delta \alpha_j \\ & + \sum_{k=1}^{m} \langle f_k \rangle \, \delta \lambda_k + \sum_{k=1}^{m} \lambda_k \delta \langle f_k \rangle \end{aligned} \tag{24}$$

The first and third terms cancel by virtue of (14). Then, using (23), we have

$$\delta S = - \sum_{j=1}^{r} \sum_{k=1}^{m} \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha_j} \right\rangle \delta \alpha_j + \sum_{k=1}^{m} \lambda_k \delta \langle f_k \rangle \tag{25}$$

Now we can write

$$\sum_{j=1}^{r} \left\langle \frac{\partial f_k}{\partial \alpha_j} \right\rangle \delta \alpha_j = \left\langle \sum_{j=1}^{r} \frac{\partial f_k}{\partial \alpha_j} \delta \alpha_j \right\rangle = \langle \delta f_k \rangle \tag{26}$$

and so finally

$$\delta S = \sum_{k=1}^{m} \lambda_k \left[\delta \left(f_k \right\rangle - \langle \delta f_k \rangle \right] \tag{27}$$

or

$$\delta S = \sum_{k=1}^{m} \lambda_k \delta Q_k \tag{28}$$

where

$$\delta Q_k \equiv \delta \langle f_k \rangle - \langle \delta f_k \rangle \tag{29}$$

In general δQ_k is not an exact differential; i.e., there is no function $Q_k \left(\lambda_1 \ldots \lambda_m ; \alpha_1 \ldots \alpha_r \right)$ which yields δQ_k by differentiation. But (28) shows that λ_k is an integrating factor such that $\sum_k \lambda_k \delta Q_k$ is the exact differential of some "state function" $S \left(\lambda_1 \cdots \lambda_m ; \alpha_2 \cdots \alpha_r \right)$.

All the above relations, (10) to (29), are elementary consequences of maximizing the information theory entropy subject to constraints on average values of certain quantities. Although they bear a strong formal resemblance to the rules of calculation provided by statistical mechanics, they make no reference to physics, and, therefore, they must apply equally well to any problem, in or out of physics, where the situation can be described by (1) enumerating a discrete set of possibilities and by (2) specifying average values of various quantities. The above formalism has been applied also to problems in engineering [2] and economics. [3]

In most problems, interest centers on making the best possible predictions for a *specific* situation, and we are not really interested in properties of any ensemble, real or imaginary. (For example, we want to predict the magnetization $M(t)$ of the *particular* spin system that exists in the laboratory.) In this case, as already emphasized, the maximum-entropy probability assignment p_i cannot be regarded as describing any objectively existing state of affairs; it is only a means of describing a state of knowledge in a way that. is "maximally noncommital" by a certain criterion. The above equations then represent simply the best predictions we are able to make on the given information. We are not entitled to assert that the predictions must be "right," only that to make any better ones, we should need more information than was given. However, in cases where it makes sense to imagine x_i as being the result of some random experiment which can be repeated many times, a somewhat more "objective" interpretation of this formalism is possible, which in its essentials was given already by Boltzmann. We are given the same average values $\langle f_k(x) \rangle$ as before, but we are now asked a different question. If the random experiment is repeated N times, the result x_i will be obtained m_i times, $1 = 1, 2, \ldots, n$, We are to make the best estimates of the numbers m_i on the basis of this much information. The knowledge of average values tells us that

$$\sum_{i=1}^{n} \frac{m_i}{N} f_k \left(x_i \right) = \langle f_k \rangle \;\; k = 1, 2, \ldots, m \tag{30}$$

and, of course,

$$\sum_{i=1}^{n} \frac{m_i}{N} = 1 \tag{31}$$

Equations (30) and (31) do not uniquely determine the m_i if $m < n-1$, and so again it is necessary to introduce some additional principle, which now amounts to stating what we mean by the "best" estimate. The following criterion seems reasonable. In N repetitions of the random experiment, there are *a priori* n^N conceivable results, since each trial could give independently any of the results $(x_1, x_2, \ldots, x_n)$. But for given m_i, there are only W of these possible, where

$$W \equiv \frac{N!}{m_1! \ldots m_n!} = \frac{N!}{(Ng_1)!\,(Ng_2)! \ldots (Ng_n)!} \tag{32}$$

and

$$g_i = \frac{m_i}{N} \;\; i = 1, 2, \ldots, n \tag{33}$$

is the *relative frequency* with which the result x_i is obtained. Which cholce of the g_i can happen in the greatest number of ways? If we have to guess the frequencies on the basis of no more information than (30) it seems that a reasonable criterion is to ask what choice will maximize (32) while agreeing with (30). Now in the limit of large N, we have by the Stirling formula,

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \log W &= \lim_{N\to\infty} \frac{1}{N} \log \left[\frac{N!}{(Ng_1)! \ldots (Ng_n)!} \right] \\ &= -\sum_{i=1}^{n} g_i \log g_i \end{aligned} \tag{34}$$

and so, if we are to estimate limiting frequencies in an indefinitely large number of trials, we have in (30) and (34) formulated exactly the same mathematical problem as in (6) and (7). The same solution (10) and formal properties, Eqs, (11) to (29), follow immediately, and we have an alternative interpretation of the maximum-entropy formalism: the probability p_i which information theory assigns to the event x_i at a *single* trial is numerically equal to an estimate of the relative frequency g_i of this result in an indefinitely large number of trials, obtained by enumerating all cases consistent with our knowledge, and placing our bets on the situation that can happen in the greatest number of ways. Thus, for example, the fluctuation laws (21) describe, on the one hand, our uncertainty as to the unknown true values of $f_k(x)$ in a specific instance; on the other hand, they give the best estimates we can make of the *average* departures from $\langle f_k \rangle$ in many repetitions of the experiment, by the criterion of placing our bets on the situation that can happen in the greatest number of ways. Two points about these interpretations should be noted:

1. In most practical problems, repeated repetition of the experiment is either impossible or not relevant to the real problem, which is to do the best we can with the *individual* case. Thus if one were to insist, as has sometimes been done, that only the second interpretation is valid, the result would be to deny ourselves the use of this formalism in most of the problems where it is helpful.

2. The argument leading from the averages (30) to the estimate of frequencies g_i was not deductive reasoning, but only plausible reasoning. Consequently, we are not entitled to assert that the estimates g_i *must* be right; only that, in order to make any better estimates, we should need more information. Thus the apparently greater "objectivity" of the second interpretation is to a large extent illusory.

3 Application to Equilibrium Thermodynamics

We apply the formalism of the preceding section to the following situation: $m = 1$, $f_1(x_j, \alpha) = E_i(V)$. The parameter V (volume) and the expectation value of the energy of the system $\langle E \rangle$ are given. The partition function is

$$Z(\lambda, V) \equiv \sum_{i=1}^{\infty} e^{-\lambda E_i(V)} \tag{35}$$

Then, by (14), λ is determined from

$$\langle E\rangle = -\frac{\partial}{\partial\lambda}\log Z \tag{36}$$

and, as a special case of (23) we have

$$\lambda\left\langle\frac{\partial E}{\partial V}\right\rangle = -\frac{\partial}{\partial V}\log Z \tag{37}$$

But $-\langle\partial E/\partial V\rangle = \langle P\rangle$ is the maximum-entropy estimate of pressure, and so the predicted equation of state is

$$\langle P\rangle = \frac{1}{\lambda}\frac{\partial}{\partial V}\log Z \tag{38}$$

To identify the temperature and entropy, we use the general variational property (28). A small change δV in volume will change the energy levels by $\delta E_i = (\partial E_i/\partial V)\,\delta V$, and if this is carried out infinitely slowly (i.e., reversibly), the "adiabatic theorem" of quantum mechanics tells us that the probabilities p_i will not be changed. So, the maximum-entropy estimate of the work done is

$$\delta W = -\langle\delta E\rangle \tag{39}$$

Of course, the given $\langle E\rangle$ is interpreted as the thermodynamic energy function U. In addition to the change δV, we allow a small reversible heat flow δQ, and by the first law, the net change in energy is $\delta U = \delta Q - \delta W$, or

Equation Missing from Original (40)

Thus, if f_k is the energy, then the δQ_k defined by (29) is the predicted heat flow in the ordinary sense. Equation (28) shows that for *any* quantity f_k, there is a quantity δQ_k formally analogous to heat.

In the present case (28) reduces to

$$\delta S(\langle E\rangle, V) = \lambda\delta Q \tag{41}$$

Now the Kelvin temperature is defined by the condition that $(1/T)$ is the integrating factor for infinitesimal reversible heat in closed systems and the experimental entropy S_e is defined as the resulting state function. So from (41) the predicted temperature T' and experimental entropy S'_e are given by

$$\lambda = \frac{1}{kT'} \tag{42}$$

$$S'_e = kS(\langle E\rangle, V) = k\,(S_I)_{\max} \tag{43}$$

The presence of Boltzmann's constant k merely indicates the particular practical units in which we choose to measure temperature and entropy. For theoretical discussions, we may as well adopt units such that $k = 1$.

All that we have shown so far is that the general maximum entropy formalism leads automatically to definitions of quantities *analogous* to those of thermodynamics. This is, of course, as far as any mathematical theory can go; no amount of mathematics can prove anything about experimental facts. To put it differently, before we can establish any connection between our theoretical entropy S'_e and the experimentally measured quantity S_e, we have to introduce some physical assumption about what the result of an experiment would in fact be:

Physical assumption The equilibrium thermodynamic properties of a system, as measured experimentally, agree with the results calculated by the usual methods of statistical mechanics; i.e., from the canonical or grand canonical ensemble appropriate to the problem.

(44)

This assumption has proved correct in every case where one has succeeded in carrying out the calculations, and its universal validity is taken so much for granted nowadays that authors of textbooks no longer list it as an assumption. But strictly speaking, all we can prove here is that systems conforming to this assumption will also conform to various other statements made below.

If we accept (44), then the identification of entropy is complete, and connection between information theory entropy and experimental entropy for the present problem can be stated as a theorem.

Theorem: Let $p_i \equiv \text{prob}\,(E_i)$ be any probability assignment which conforms to the data in the sense that $\langle E\rangle = \sum_i p_i E_i$ is the measured energy. Let $S_I \equiv -\sum p_i \log p_i$ be the corresponding information theory entropy, and S_e be the experimentally measured entropy for the system. The additive constant is chosen so that at zero temperature $S_e = \log n$, where n is the degeneracy of the ground state, and let S_e be expressed in units such that Boltzmann's constant $k \equiv 1$. Then

$$S_I \leq S_e \tag{45}$$

with equality if and only if p_i is chosen as the canonical distribution

$$p_i = \frac{1}{Z}\exp\left(-\lambda E_i(V)\right) \tag{46}$$

This is the physical meaning, for the present problem, of the general inequality (12). Obviously, the above statement can be greatly generalized; we can introduce more degrees of freedom in addition to V, we can consider open systems, where the number of molecules can change, and we can use the grand canonical ensemble, etc. The corresponding statement will still hold; over all probability assignments that agree with the data in the aforementioned sense, the information theory entropy attains an absolute maximum, equal to the experimental entropy, if and only if p_i is taken as the appropriate canonical or grand canonical distribution.

Remarks:

1. We have taken $\langle E\rangle$ as the given quantity. In practice, it is usually the temperature that is measured. To treat the temperature as the observable, one must regard the system of interest to be in contact with a heat reservoir, with which it may exchange energy and which acts as a thermometer. Detailed analysis of the resulting system (given in reference [4]) leads to the same probability assignments as we have found with $\langle E\rangle$ as the given datum.

2. If not only $\langle E\rangle$ is known, but also the accuracy of the measurement, as given for example by $\langle E^2\rangle$, then this information may be incorporated into the problem by taking $f_1\,(x_1,\alpha) = E_1(V)$, $f_2\,(x_1,\alpha) = E_i^2(V)$. The partition function (11) becomes

$$Z\,(\lambda_1,\lambda_2,V) = \sum_i \exp\left[-\lambda_1 E_i(V) - \lambda_2 E_i^2(V)\right] \tag{47}$$

and from (14),

$$\langle E\rangle = -\frac{\partial}{\partial\lambda_1}\log Z \quad \langle E^2\rangle = -\frac{\partial}{\partial\lambda_2}\log Z \tag{48}$$

The fluctuation theorem (21) then gives the relation

$$\langle E^3\rangle - \langle E\rangle\langle E^2\rangle = -\frac{\partial\langle E\rangle}{\partial\lambda_2} = -\frac{\partial\langle E^2\rangle}{\partial\lambda_1} \tag{49}$$

In principle, whenever information of this sort is available, it should be incorporated into the problem. In practice, however, we find that for the macroscopic systems that exhibit reproducible thermodynamic properties, the variance $\langle E^2\rangle - \langle E\rangle^2$ as calculated from (46) is already very small compared to any reasonable mean-square experimental error, and so the additional information about accuracy of the measurement did not lead to any difference in the predictions. This is, of course, the basic reason for the success of the Gibbs canonical ensemble formalism.

3. The theory as developed here has, in principle, an additional freedom of choice not present in conventional statistical mechanics. The statement that a system has a definite, reproducible equation of state means, for example, that if we fix *experimentally* any two of the parameters P, V, T, then the third is determined. Correspondingly, in the theory it should be true that *information* about any two of these quantities should suffice to enable us to *predict* the third; there is no basic reason for constructing our ensembles always in terms of energy rather than any other measurable quantities. Use of energy has the mathematical convenience that energy is a constant of the motion, and so the statement that the

system is in equilibrium (i.e., measurable parameters are not time-dependent) requires no additional constraint. With an ensemble based on some quantity, such as pressure or magnetization, which is not an intrinsic constant of the motion, if we wish to predict equilibrium properties we need to incorporate into the theory an additional statement, involving the equations of motion, which specifies that these quantities are constant. To do this requires no new principles of reasoning beyond those given above; we merely include the values of such a quantity $f(t_i)$ at many different times (or in the limit, at all times) into the set of quantities f_k whose expectation values are given. In the limit, the partition function thus becomes a partition functional:

$$Z[\lambda(t)] = \sum_i \exp\left[-\int \lambda(t) f(x_i, t)\, \mathrm{d}t\right] \tag{50}$$

and the relations (14) determining the λ's go into the corresponding functional derivative relations

$$\langle f(t)\rangle = -\frac{\delta}{\delta\lambda(t)} \log Z[\lambda(t)] \tag{51}$$

which determine the function $\lambda(t)$.

We have not found any general proof that the predicted equation of state is independent of the type of information used, but a special case is proved in the 1961 Stanford thesis of Dr. Douglas Scalapino. There it is shown that the same equation of state of a paramagnetic substance with spin-spin interaction is obtained whatever the input information. We conjecture that this is true for any system that exhibits an experimentally reproducible equation of state.

It is doubtful whether this new degree of freedom in applying the theory will prove useful in calculations pertaining to the equilibrium state, since it is more complicated than the usual procedure. However, it is just this extra freedom that makes it possible to develop a general formalism for irreversible processes; indeed, prediction of time-dependent phenomena is obviously impossible as long as our probability distributions depend only on constants of the motion, Equations (50) and (51) form the starting point for a general theory of the nonequilibrium steady state, the Scalapino thesis providing an example of the calculation of transport coefficients from them.

4 Generalization

For most applications of interest, the foregoing formalism needs to be generalized to the case of (a) systems described by a density matrix or (b) continuous probability distributions as occur in classical theory. We indicate briefly how this is done.

4.1 Density Matrix

The expectation value of an operator F_k of a system described by the density matrix ρ is

$$\langle F_k\rangle = \mathrm{Tr}\,(\rho F_k) \tag{52}$$

where Tr stands for the trace. The information theory entropy corresponding to ρ is

$$S_I = -\,\mathrm{Tr}(\rho \log \rho) \tag{53}$$

(See reference[5] for the arguments that lead to this definition of S_I and discussion of other expressions which have been proposed.) Maximizing S_1 subject to the constraints imposed by knowledge of the $\langle F_k\rangle$ yields

$$\rho = \frac{1}{Z(\lambda_1 \ldots \lambda_m} \exp\left(-\lambda_1 F_1 - \ldots - \lambda_m F_m\right) \tag{54}$$

where

$$Z(\lambda_1 \ldots \lambda_m) \equiv \mathrm{Tr} \exp\left(-\lambda_1 F_1 - \ldots - \lambda_m F_m\right. \tag{55}$$

To prove (54), use the lemma

$$\mathrm{Tr}(\rho \log \rho) \geq \mathrm{Tr}(\rho \log \sigma) \tag{56}$$

analogous to (9) Here ρ is any density matrix satisfying (52), and σ is the canonical density matrix (54). All the formal relations (12) to (29) still hold, except that when the F_k do not all commute, the fluctuation law (21) must be generalized to where

$$\begin{aligned} -\frac{\partial \langle F_k \rangle}{\partial \lambda_j} = -\frac{\partial \langle F_j \rangle}{\partial \lambda_k} &= \int_0^1 \left\langle e^{xA} F_k e^{-xA} F_j \right\rangle \mathrm{d}x \\ &= \langle F_k \rangle \langle F_j \rangle \end{aligned} \tag{57}$$

$$A \equiv \sum_{k=1}^{m} \lambda_k F_k \tag{58}$$

For all ρ that agree with the data in the sense of (52), we have $S_Y(\rho) \leq S_e$, with equality if and only if ρ is the canonical matrix (54).

4.2 Continuous Distributions

Shannon's fundamental uniqueness theorem (reference,[1] theorem 3) which establishes $-\sum p_i \log p_i$ as the correct information measure, goes through only for discrete probability distributions. At the present time, the only criterion we have for finding the analogous expression for the continuous case is to pass to the limit from a discrete one; presumably, future study will give a more elegant approach. The following argument can be made as rigorous as we please, but at considerable sacrifice of clarity. In the discrete entropy expression

$$S_I^{(\mathrm{d})} = -\sum_{i=1}^{n} p_i \log p_i \tag{59}$$

we suppose that the discrete points $x_i, i = 1, 2, \ldots, n$, become more and more numerous, in such a way that, in the limit $n \to \infty$ the density of points approaches a definite function $m(x)$:

$$\lim_{n\to\infty} \frac{1}{n}(\text{ number of points in } a < x < b) = \int_a^b m(x)\mathrm{d}x \tag{60}$$

If this passage to the limit is sufficiently well behaved, it will also be true that adjacent differences $(x_{i+1} - x_i)$ in the neighborhood of any particular value of x will tend to zero so that

$$\lim_{n\to\infty} \left[n\left(x_{i+1} - x_i\right)\right] = \left[m\left(x_i\right)\right]^{-1} \tag{61}$$

The discrete probability distribution p_i will go over into a continuous probability density $w(x)$, according to the limiting form of

$$p_i = w\left(x_i\right)\left(x_{i+1} - x_i\right)$$

or, from (61),

$$p_i \to w\left(x_i\right)\left[nm\left(x_i\right)\right]^{-1} \tag{62}$$

Consequently, the discrete entropy (59) goes over into the integral

$$S_I^{(\mathrm{d})} - \int w(x)dx \log\left[\frac{w(x)}{nm(x)}\right].$$

In the limit, this contains an infinite term $\log n$; but if we subtract this off, the difference will, in the cases of interest, approach a definite limit which we take as the continuous information measure:

$$S_I^{(c)} \equiv \lim\left[S_I^{(\mathrm{d})} - \log n\right] = -\int w(x) \log\left[\frac{w(x)}{m(x)}\right] \mathrm{d}x. \tag{63}$$

The expression (63) is invariant under parameter changes; i.e., instead of x another quantity $y(x)$ could be used as the independent variable. The probability density and measure function $m(x)$ transform as

$$w_1(y)\mathrm{d}y = w(x)\mathrm{d}x$$
$$m_1(y)\mathrm{d}y = m(x)\mathrm{d}x$$

so that (63) goes into

$$S_I^{(c)} = -\int w_1(y)\mathrm{d}y \log\left[\frac{w_1(y)}{m_1(y)}\right]. \tag{64}$$

To achieve this invariance it is necessary that the "measure" $m(x)$ be introduced. I stress this point because one still finds, in the literature, statements to the effect that the entropy of a continuous probability distribution is *not* an invariant. This is due to the historical accident that in his original papers, Shannon [1] assumed, without calculating, that the analog of $\sum p_1 \log p_i$ was $\int w \log w dx$, and got into trouble for lack of invariance. Only recently have we realized that mathematical deduction from the uniqueness theorem, instead of guesswork, yields the invariant information measure (63).

In many cases it is more natural to pass from the discrete distribution to a continuous distribution of several variables, $x_1 \ldots x_r$; in this case the results readily generalize to

$$S_I^{(c)} = -\int \ldots \int w\left(x_1 \ldots x_r\right) \log\left[\frac{w\left(x_1 \ldots x_r\right)}{m\left(x_1 \ldots x_r\right)}\right] dx_1 \ldots dx_r. \tag{65}$$

We apply this to the Liouville function of classical mechanics. For a system of N particles, $W_N\left(x_1 p_1 \ldots x_2 p_x; t\right) \mathrm{d}^3x_1 \ldots \mathrm{d}^3p_i$ is the probability that at time t the system is in the element $d^3x_1 \ldots \mathrm{d}^3p_N$ of $6N$ -dimensional phase space. Before we can set up the information measure for this case, we must decide on a basic measure $m\left(x_1 \ldots p_N\right)$ for phase space. In classical statistical mechanics, one has always taken uniform measure: $m = \text{const.}$, largely because one couldn't think of anything else to do. However, the more careful writers have all stressed the fact that *within the context of classical theory*, no real justification of this has ever been produced. For the present, I propose to dodge this issue by regarding classical statistical mechanics merely as a limiting form of the (presumably more fundamental) discrete quantum statistical mechanics. In other words, the well-known proposition that each discrete quantum state corresponds to a volume h^{3N} of classical phase space, will determine our uniform measure as resulting from equal weighting of all orthogonal quantum states, and passing to the limit $h \to 0$. Thus, apart from an irrelevant additive constant which we drop, our information measure will be just the negative of the Gibbs H-function, H_G:

$$-S_I = H_G = \int W_N \log W_N \mathrm{d}\tau \tag{66}$$

where $d\tau = \mathrm{d}^3x_1 \ldots \mathrm{d}^3p_N$.

With this continuous probability distribution, we are able to incorporate into the theory a more detailed kind of macroscopic information than we have considered up till now. Suppose we are given the macroscopic density $\rho(x)$ as a function of position. We interpret this as specifying at each point of space, the expectation value of a certain quantity:

$$\left\langle f_1\left(x_1 p_1 \ldots x_N p_N; x\right)\right\rangle = \int W_N f_2 \mathrm{d}\tau = \rho(x) \tag{67}$$

where the phase function f_1 is given by

$$f_1\left(x_1 p_1 \ldots x_N p_N; x\right) = \sum_{i=1}^{N} m\delta\left(x_i - x\right) \tag{68}$$

The position x now plays the same role as the index k in the elementary version of the formalism, Eqs. (10) to (29) and so in place of the sum $\sum_k \lambda_k, f_k\left(x_i\right)$ in the exponent of the probability distribution

$$\int \lambda(x) f_1 \mathrm{d}^3x$$

into the exponent of W_N. The partition function then becomes a partition functional of the function $\lambda(x)$.

In general, we might have several phase functions of this kind, whose expectation values are given at each point of space:

$$\begin{aligned} \langle f_1 (x_1 \dots p_N; x) \rangle &= \int W_N f_1 \mathrm{d}\tau \\ &\dots\dots\dots \\ \langle f_m (x_1 \dots p_N; x) \rangle &= \int W_N f_m \mathrm{d}\tau \end{aligned} \tag{69}$$

Maximization of S_I subject to these constraints gives the partition functional

$$\begin{aligned} Z [\lambda_1(x), \dots, \lambda_m(x)] = \int \mathrm{d}\tau \exp \Bigg\{ - \sum_{k=1}^{m} \int \lambda_k(x) \\ \times f_k (x_1 \dots p_N; x) \, \mathrm{d}^3 x \Bigg\} \end{aligned} \tag{70}$$

The Lagrange multiplier functions $\lambda_k(x)$ are determined by relations analogous to (14), but now involving the functional derivatives:

$$\langle f_k (x_1 \dots p_N; x) \rangle = - \frac{\delta}{\delta \lambda_k(x)} \log Z [\lambda_1(x), \dots, \lambda_m(x)] \tag{71}$$

and the other properties, Eqs. (16) to (29), are likewise easily generalized.

Example: Suppose the macroscopic density of mass, momentum, and kinetic energy are given at the initial time. This corresponds to expectation values of (68) and

$$\langle f_2 (x_1 \dots p_N; x) \rangle = \left\langle \sum_{i=1}^{N} p_i \delta (x_i - x) \right\rangle = p(x) \tag{72}$$

$$\langle f_3 (x_1 \dots p_N; x) \rangle = \left\langle \sum_{i=1}^{N} \frac{p_i^2}{2m} \delta (x_i - x) \right\rangle = K(x) \tag{73}$$

Since all the given data are formed additively from contributions of each particle, the maximum-entropy Liouville function factors:

$$W_N = w_1 (x_i, p_i) \tag{74}$$

(this would not be the case if the given information concerned mutual properties of different particles, such as the potential energy), and the exponential in the partition functional (70) reduces to

$$\begin{aligned} &- \int \mathrm{d}^3 x \left[\lambda_1(x) \sum_i m \delta (x_i - x) + \lambda_2(x) \cdot \sum_i p_i \delta (x_i - x) \right. \\ &\left. + \lambda_3(x) \sum_i \frac{p_i^2}{2m} \delta (x_i - x) \right] \\ &= - \sum_{i=1}^{N} \left[m \lambda_1 (x_i) + p_i \cdot \lambda_2 (x_i) + \frac{p_i^2}{2m} \lambda_3 (x_i) \right] \end{aligned}$$

so that

$$\begin{aligned} \log Z = & N \log \int \left(\exp \left[- m \lambda_1(x) - p \cdot \lambda_2(x) \right. \right. \\ & \left. \left. - \frac{p^2}{2m} \lambda_3(x) \right] \right) \mathrm{d}^3 x \mathrm{d}^3 p \end{aligned} \tag{75}$$

Application of (71) now yields the physical meaning of the Lagrange multipliers: defining the "mass velocity" $u(x)$ by $P(x) = \rho(x)u(x)$ and the "local temperature" $T(x)$ by the mean-square velocity as seen by an observer moving at velocity $u(x)$, we find

$$\begin{aligned} \lambda_3(x) &= \frac{1}{kT(x)} = \beta(x) \\ \lambda_2(x) &= \beta(x)u(x) \\ m\lambda_1(x) &= 1/2mu^2(x)\beta(x) - 3/2\log\beta(x) - \log\rho(x) + (\text{const.}) \end{aligned} \tag{76}$$

and the single-particle distribution function w_1 of (74) reduces to

$$w_1(x,p) = \frac{\rho(x)}{mN\left[2\pi mkT(x)\right]^{3/2}} \exp\left\{-\frac{[p-mu(x)]^2}{2mkT(x)}\right\} \tag{77}$$

In this rather trivial example we merely recover a well-known result; but from a different viewpoint than the usual one, which leads us to interpret (77) differently, and regard it as a very special case. The method used enables us to translate other kinds of macroscopic information into definite probability distributions. In other words, we suggest that the maximum-entropy formalism provides the general solution to the problem of "setting up an ensemble" to describe an arbitrary macroscopic situation, equilibrium or nonequilibrium.

The distributions found in the above way, of course, describe he situation only at the initial time for which the macroscopic information is given. For predictions referring to other times, one should, in principle, solve the equations of motion, or Liouville equation,

$$\dot{W}_N + [W_N, H] = 0 \tag{78}$$

where H is the Hamiltonian and $[W_N, H]$, the Poisson bracket. In practice, the history of irreversible statistical mechanics has been one of unceasing efforts to replace this impossibly difficult calculation by a simpler one, in which we try to reduce 78 to an 'irreversible' equation variously termed Boltzmann equation, rate equation, or master equation. Although considerable progress has been made in this direction in recent years, we are still far from really bridging the gap between these two methods of description.

As a preliminary step in this direction, it is necessary that we understand clearly the physical meaning of the Liouville function W_N and the various reduced distribution functions derived from it. The following section surveys these questions.

5 Distribution Functions

A recent review of transport theory by Dresden[6] (hereafter referred to as MD) illustrates that attempts to bridge the gap between phenomenological rate equations and fundamentals (equations of Liouville and Gibbs) have been largely frustrated because basic conceptual difficulties, dating from the time of Boltzmann, are still unresolved. This section is intended as a supplement to the discussion of these problems given to, MD, Sec. I. B.

Early attempts to base transport theory on the BBGKY hierarchy of distribution functions made no distinction between the Boltzmann distribution function $f(x,p,t)$ and the single-particle function $w_1(xpt)$ of the hierarchy. In MD this distinction is pointed out without, however, stating any precise relation between them. To do this requires, first of all, precise definitions of f and the Liouville function W_N. Boltzmann originally defined f as giving the *actual number* of particles in various cells of six-dimensional phase space; thus if R is the set of phase points comprising a cell, the number of particles in R is

$$n_R = \int_R f(x,p,t)\mathrm{d}^3x\mathrm{d}^3p \tag{79}$$

The well-known paradoxes involving the H-theorem led to a feeling that this definition should be modified; but the exact way seems never to have been stated. Here we retain the definition (79), which has at least the merit of being a precise statement, and accept the consequence that the Boltzmann collision equation cannot be strictly correct, for reasons given by Zermelo and Loschmidt.

From (79) it is immediately clear that-Boltzmann's f is not a probability distribution at all, but a "random variable." In other words, instead of saying that f gives the probability of various conditions, we should ask, "What is the probability that f takes on various values?"

Establishment of a precise connection between Boltzmann's f and the single-particle function of the hierarchy,

$$w_2\left(x_1, p_1, t\right) = \int w_N \mathrm{d}^3 x_2 \ldots \mathrm{d}^3 p_N \tag{80}$$

requires no coarse-graining, time-smoothing, or any other mutilation of the hierarchy. If we agree that a particle will be considered "in R" if its center of gravity is in R, and that the Liouville function W_N is symmetric under permutations of particle labels, then from (79) and (80) the exact connection between them is simply

$$\langle f \rangle = N w_1 \tag{81}$$

where the angular brackets denote an average over the Liouville function W_N. The only "statistical notion" which needs to be adjoined to it is the usual one that $W_N \mathrm{d}\tau$ shall be interpreted as the probability that the *individual system* is in the phase region dt. To say that W_N refers to number density in a fictitious ensemble is only to say the same thing in different words; this cannot be emphasized too strongly. Indeed, the notlon of an ensemble 1s merely a device that enables us to speak of probabilities on the Gibbs, or global level, as if they were frequencies in some larger system which is defined for just that purpose.

The reason why it was felt necessary to introduce the notion of an ensemble is that the development of equilibrium statistical mechanics took place entirely in a period when the frequency theory of probability was the only one considered respectable. It has been taken for granted that any probability distributions used must be, in principle, empirically measurable frequencies, *and that the fundamental problem of statistical mechanics is to justify these distributions in the frequency sense.*

The statistical practice of physicists has tended to lag about 20 years behind current developments in the field of basic probability and statistics. I hope to shorten that gap to about 10 years by pointing out that a revolution in statistical thought has recently taken place, brought about largely by the development of statistical decision theory. Two brief summaries of these developments have been published [3;7] and a detailed analysis of the present situation [8] will soon be available. The net result is a vindication of the viewpoint of Laplace, and of Jeffreys,[9] that probability theory is properly regarded as an extension of logic to the case of inductive, or plausible, reasoning, the probabilities denoting basically a "degree of reasonable belief," rather than limiting frequencies. This does not mean that there are no longer any connections between probability and frequency; the situation is rather that every connection between probability and frequency which is actually used in applications is deducible as a mathematical consequence of the "inductive logic" theory of probability.[8] Equation (81), and others given below, provide examples of the kind of connections that exist.

Use of probability in this "modern" (actually the original) sense is, of course, essential to the maximum-entropy formalism; for the *frequencies* with which different microscopic states are occupied are manifestly not given, in general, by a distribution canonical in the observed quantities; indeed, for a time-dependent problem the notion of occupation frequency is meaningless. Nevertheless, in a problem where frequencies are meaningful, if our job is to estimate those frequencies, our best estimate on the basis of the information available will be numerically equal to the probabilities. One example of this was given in the "objective" interpretation of the maximum-entropy formalism in Sec. 2, and we now give another example which clarifies the meaning of the distribution functions.

From Eqs. (79) and (81) one sees that the single-particle function w_1 does *not* contain full information about the distribution of particles in six-dimensional phase space. Integrating (81) over the cell R, we see that it determines only the expectation value of particle occupation numbers:

$$\langle n_R \rangle = N \int_R w_1(x, p, t) \mathrm{d}^3 x \mathrm{d}^3 p \tag{82}$$

In words: the integral in (82) represents the probability that any *specified* particle is in the phase cell R. This is not the same as the fraction of particles in that cell but represents only the expectation value of that fraction, over the Liouville distribution W_N Before we are justified in the usual interpretation which identifies (82) with the actual number of particles in R, it must be shown that the variance of the n_R distribution is

small:

$$\frac{\left\langle n_R^2 \right\rangle - \left\langle n_R \right\rangle^2}{\left\langle n_R \right\rangle^2} \ll 1 \tag{83}$$

Unless (83) is satisfied, the Liouville function is making no definite prediction about the number of particles in R. But *we are not allowed to postulate (83) on the grounds of any "law of large numbers" even for a cell R of macroscopic size*, because the two-particle distribution function of the hierarchy,

$$w_2\left(x_1 p_1, x_2 p_2, t\right) = \int w_N \mathrm{d}^3 x_3 \ldots \mathrm{d}^3 p_N \tag{84}$$

completely determines whether (83) is or is not satisfied. To see this, introduce the characteristic function of the set R:

$$M(x,p) \equiv \left\{ \begin{array}{l} 1, x, p \text{ in } R \\ 0, \text{ otherwise} \end{array} \right\} \tag{85}$$

Then

$$\left\langle n_R^2 \right\rangle = \sum_{i,j=1}^{N} \left\langle M\left(x_i, p_i\right) M\left(x_j, p_j\right) \right\rangle = N I_1 + N(N-1) I_2 \tag{86}$$

where

$$I_1 \equiv \int_R w_1(x,p) \mathrm{d}^3 x \mathrm{d}^3 p \tag{87}$$

$$I_2 \equiv \int_R \mathrm{d}^3 x_1 \mathrm{d}^3 p_1 \int_R \mathrm{d}^3 x_2 \mathrm{d}^3 p_2 w_2\left(x_1 p_1, x_2 p_2\right) \tag{88}$$

The measure of dispersion (83) then reduces to

$$\frac{I_2 - I_1^2}{I_1^2} + \frac{I_1 - I_2}{N I_1^2} \tag{89}$$

Thus, when $N \gg 1$ and $\langle n_R \rangle \gg 1$, the necessary and sufficient condition for validity of ($\hat{8}3$) becomes

$$\left| \frac{I_2}{I_1^2} - 1 \right| \ll 1 \tag{90}$$

Usually one omits gravitational forces from the Hamiltonian and chooses a Liouville function which makes w_1 independent of position. If we then describe thermal equilibrium by $W_N \sim \exp(-\beta H)$ and choose a cell R consisting of all of momentum space, and a region V_R of ordinary space of macroscopic size, Eq. (90) becomes the necessary and sufficient condition that the Liouville function makes a sharp prediction of the density of the fluid; i.e., it predicts that only one phase is present in V_R. Thus the condition for condensation, or more precisely for the coexistence of more than one phase, is that (90) fails to hold. Equation (82) then gives only a weighted average of the density of the various possible phases.

Similarly, in the problem of deriving the laws of hydrodynamics from the Liouville equation, one needs to find the predicted momentum density. In terms of the Boltzmann. distribution function, the total momentum in any phase cell R is

$$P = \int_R p f(x,p,t) \mathrm{d}^3 x \mathrm{d}^3 p \tag{91}$$

and we choose R to consist of all momentum space plus a cell S' of ordinary space that is "microscopically large but macroscopically small." Again, the single-particle function gives only the expectation value,

$$\langle P \rangle = N \int_R p w_2(x,p,t) \mathrm{d}^3 x \mathrm{d}^3 p \tag{92}$$

but w_1 gives no information at all as to whether this is a *reliable* prediction, To answer this, we must appeal to the two-particle function:

$$\begin{aligned}\left\langle P^2\right\rangle =& N\int_R p^2 w_1 \mathrm{d}x\mathrm{d}p + N(N-1)\int_R \mathrm{d}x\mathrm{d}p\int_R \mathrm{d}x'\mathrm{d}p' \\ & \times p\cdot p' w_2\left(x,p,x',p'\right)\end{aligned} \tag{93}$$

If the variance of P is everywhere small, then the Liouville function is making a definite prediction of a flow pattern; i.e., it predicts laminar flow. But if the last term of (93) is large, the single-particle function gives only a weighted average of several possible flows. In this case, the information put into the Liouville function was not sufficient to determine any definite mass motion of the fluid. But if we incorporated into W_N all the information about the experimentally imposed conditions, the theory is now telling us that under these conditions the flow will not be experimentally reproducible. In other words, the theory is predicting turbulent flow.

These examples show that the proper physical interpretation of the distributions (i.e., their exact relation to physical quantities) is not an obscure philosophical point. Failure to distinguish between w_1 and f as given in (79) means failure to distinguish between expectation values and actual values, and amounts to the same thing as simply *postulating* that ensemble averages are equal to observed values of physical quantities. This is not only unjustified because of the probability nature of W_N; it would mean loss of the correct criterion for phase changes and of the criterion which distinguishes between laminar and turbulent flow.

On the other land, we can see no basis for any distinction between equilibrium and nonequilibrium situations here. One of the most elementary theorems of probability theory assures us that, for any phase function Q and any probability assignment W_N whatsoever, the expectation value $\langle Q\rangle$, denoted by Q_{Obs} in MD, is the best estimate of Q in the sense that it minimizes the expected square of the error. Whether the information put into W_N permits an *accurate* estimate (i.e., whether the expected square of the error is small), can be neither postulated nor denied arbitrarily; it is determined by W_N. In all cases, equilibrium or otherwise, the test is to calculate $\left\langle Q^2\right\rangle = \int Q^2 W_N \mathrm{d}v$, and see whether it is sufficiently close to $\langle Q\rangle^2$ in the sense of (83). If calculation of $\langle Q\rangle$ requires knowledge of the function w_s of the hierarchy, but not w_{S+1}, and $2s < N$, then information about the reliability of the ensemble average (Q) as an estimate of Q appears for the first time in the function w_{2S}, and is, of course, retained in all higher order functions.

Any system of "kinetic equations," such as the Boltzmann or Bogoliubov scheme, which attempts to write the higher-order functions in terms of w_1, throws away information about the reliability of the predictions. This, however, may represent a net advantage if it simplifies the mathematics without greatly affecting the actual predictions; consequently the search for such kinetic equations is a major objective of current theoretical effort. If the particles move under the influence of a potential energy function $V\left(x_1\ldots x_N\right)$, the exact differential equation satisfied by $w_2\left(x_1,p_1,t\right)$ may be written compactly

$$\frac{\partial w_1}{\partial t} + \frac{p_{1\alpha}}{m}\frac{\partial w_1}{\partial x_\alpha} + \frac{\partial}{\partial p_{1\alpha}}\left[\left\langle F_\alpha\right\rangle, w_1\right] = 0 \tag{94}$$

where

$$\left\langle F_\alpha\right\rangle = -\int \frac{\partial V}{\partial x_{1\alpha}}\left(x_2\ldots p_N \mid x_1 p_1\right)\mathrm{d}^3x_2\ldots\mathrm{d}^3p_N \tag{95}$$

is the conditional expectation value of the force seen by particle 1 , given that it has position and momentum (x_1,p_1). Here $\left(x_2\ldots p_N \mid\ x_1 p_2\ \right)$ is the conditional probability density for the other particles, defined by $W_N\left(x_1\ldots p_N\right) = \left(x_2\ldots p_N \mid x_1p_1\right) w_1\left(x_1p_2\right)$.

Although direct calculation of $\left\langle F_\alpha\right\rangle$ would be very difficult, the form of (94) should prove useful in two respects. In the first place, it shows that, although the basic ideas may be stated in entirely different terms, any proposed equation for w_1, such as the Boltzmann, the Fokker-Planck, or the Bogoliubov equation, is equivalent to some assumption about the expected force $\left\langle F_\alpha\right\rangle$. The physical reasonableness of any proposed equation may, therefore, be judged by comparing it to (94), and seeing what explicit assumption it makes about $\left\langle F_\alpha\right\rangle$. Second, (94) shows that all the complications of this subject reduce ultimately to the determination of one quantity, $\left\langle F_\alpha\right\rangle$. Therefore, a phenomenological theory should be feasible in which $\left\langle F_\alpha\right\rangle$ is determined from appropriate experiments. In situations close to equilibrium, one finds in this way that

in first approximation $\langle F_\alpha \rangle$ is proportional to the density gradient, and independent of p_1. The condition for condensation, which is a particular kind of hydrodynamic instability, is then that this proportionality coefficient exceeds a certain critical value.

6 Entropy and Probability

Now we turn to what is perhaps the most serious confusion of all in current irreversible statistical mechanics-the interpretation of entropy in terms of probability distributions. As recent literature gives ample testimony, even the issue of Boltzmann's versus Gibbs' H functions to represent entropy has not been resolved in any commonly agreed way. For example, in MD it is stated that the Boltzmann H,

$$H_B = \int f \log f \mathrm{d}^3x \mathrm{d}^3p \tag{96}$$

is "directly related" to the entropy, whereas the Gibbs expression

$$H_G = \frac{1}{N} \int W_N \log W_N dv \tag{97}$$

is rejected with the statement: "There is, however, no possibility of identifying or relating H_G to the macroscopic entropy, for one proves directly from (23) and (18) that H_G is constant in time, whereas the macroscopic entropy always increases in a nonequilibrium situation." Similar statements appeared in the Ehrenfest [10] review article of 1912, when the work of Gibbs had not yet been understood. From the frequency with which this objection to Gibbs' H has been repeated in the literature since then, it is clear that the nature of Gibbs' contribution has not been fully appreciated to this day.

We wish to point out that the mathematical relations proved by Gibbs, plus one physical assumption which is universally accepted today (although it had hardly been formulated at the time of the Ehrenfest article) are sufficient to prove, on the contrary, the following four statements:

I. The Gibbs H has a simple and universally valid connection with the entropy; for all probability assignments that agree with the measured thermodynamic parameters we have $S \geq -kH_G$, with equality if and only if H_G is computed from the appropriate canonical or grand canonical probability assignment.

II. The Boltzmann H is related to the entropy in only one case, the nonexistent ideal Boltzmann (i.e., not Bose or Fermi) gas. In general, $H_B \leq H_G$, and the entropy can be either greater or less than $-kH_B$.

III. The constancy of Gibbs' H, far from *conflicting* with the increase of entropy, is the sole dynamical property needed to *demonstrate* that increase.

IV. The Gibbs H provides a generalized definition of entropy for nonequilibrium cases, in such a way that the usual statement of the second law remains valid. It gives, therefore, a new rule telling which *nonequilibrium* states are accessible from others in adiabatic processes.

The fourth statement is a nontrivial extension of the second law which is capable of being tested experimentally, and whose finding required only a careful reading of Gibbs. Since the second law is a statement of experimental fact, it cannot be "proved" mathematically without some assumption about what the result of an experiment would be. The assumption we need is just the statement (44) which we appealed to before.

Before turning to the proofs, some preliminary remarks are needed. We are still faced with the ambiguity in the definition of f. The function defined by (79) is singular in such a way that the integral (96) diverges; thus before we can introduce a Boltzmann H at all, we have to abandon Boltzmann's definition of f in favor of some other, unspecified one. In MD it is stated that f gives an "average" occupation number, and that this can be made more precise by reference to an equation which is indeed an average over an undefined probability distribution P. If we suppose that, in going to fundamentals, this would eventually become an average over the Liouville function W_N, we have a definition of H_B for which exact relations can be proved. In other words, we mean to use the single-particle function w_1 of the hierarchy to define a Boltzmann H:

$$H_B = \int w_1 \log w_1 \mathrm{d}^3x \mathrm{d}^3p \tag{98}$$

There is really no other way of doing it if we are ever to prove precise statements about Boltzmann's H, because eventually this will have to depend on precise properties of the dynamics, and the Liouville hierarchy is just the precise expression of the dynamics.

Another point is that, strictly speaking, all this should be restated in terms of quantum theory using the density matrix formalism. This will introduce the $N!$ permutation factor, a natural zero for entropy, alteration of numerical values if discreteness of energy levels becomes comparable to kT, etc. But there seems to be complete agreement as to how this transcription is to be made, and it will affect the Boltzmann and Gibbs expressions in the same way. We shall first attempt to define the Boltzmann H as $H' = \mathrm{Tr}(\sigma \log \sigma)$, where σ is the "molecular" density matrix operating in the Hilbert space of a single molecule and gives occupation numbers. The Gibbs H will become $H'_G = N^{-1} \mathrm{Tr}(\rho \log \rho)$, where ρ is the "global" density matrix with an enormously greater number of rows and columns, operating in the entire Hilbert space of the system. On closer examination, we shall wonder whether the diagonal elements of σ are to represent the actual values, probable values, average values, etc. of the occupation numbers, and H will peter out in ambiguities until we note that, if it is to have any precisely provable properties, it must be precisely related to the dynamics; i. e., out of all possible definitions of σ, we decide to use ρ_1, the projection of ρ onto the subspace of a single molecule, as defined in reference,[5] Sec. 11. Its diagonal elements are expectation values, over the global density matrix ρ, of occupation fractions. Then with H'_G and $H'_B = \mathrm{Tr}\,(\rho_1 \log \rho_1)$ we can prove exactly the same inequalities as for the classical case. Thus, the issue of Boltzmann versus Gibbs entropy expressions does not involve quantum theory, and we continue to use classical terminology for brevity.

Statement (I) is now just the theorem (45) already proved, if one grants the physical assumption (44), for the quantum theory case.

Statement (II) quotes a well-known mathematical theorem,

$$H_G \geq H_B \tag{99}$$

with equality if and only if the Liouville function factors "almost everywhere"

$$W_N\,(x_1 \ldots p_N) = \prod_{i=1}^{N} w_A\,(x_i, p_i) \tag{100}$$

which corresponds, in quantum theory, to the condition that the global density matrix is a direct product [5]

$$\rho = \rho_1 \times \rho_2 \times \ldots \times \rho_N \tag{101}$$

where ρ_1 is the projection of ρ onto the Hilbert space of the i-th molecule. The final part of statement II then follows from the fact that the canonical distribution $W_N \sim \exp(-\beta H)$ has the factorized form (100) only in the case of an ideal Boltzmann gas. In this case the "Boltzmann entropy," $S_B = -kH_B$, is equal to the experimental entropy; in all other cases, if w_1 is constructed from the appropriate canonical distribution W_N, we shall have $S_B > S_e$

Statement III is likewise an immediate consequence of statement I and the well-known fact that H_G is, in consequence of the equations of motion, constant in time in either classical or quantum theory. To make this clearer, consider the following experiment. At time $t = 0$, we measure the values of various parameters $X_1 \ldots X_n$ adequate to determine the state of a thermodynamic system of n degrees of freedom. The experimental entropy is, of course, some function $S_e\,(X_1 \ldots X_n)$ of the measured quantities; and not primarily related to any probability distribution. But we have shown that the maximum attainable information theory entropy S_I corresponding to the appropriate canonical distribution based on the values of $X_1 \ldots X_n$, is equal to S_e. At some later time t, a new measurement of the thermodynamic state yields different values, $X'_1, \ldots, X'_n$, and a different experimental entropy $S_e\,(X'_1 \ldots X'_n)$. But the inequality $S_I \leq S_e$ still holds; and so the statement that S_I (or what is the same thing, H_G) is constant, then gives us $S'_e \geq S_e$.

There is still an apparent paradox hiding here; for suppose we choose t negative. It looks as if this argument then says that the experimental entropy in the past was greater than at the time of the measurements $X_1 \ldots X_n$. Actually, the explanation of this paradox has been given before.[5] We have, of course, assumed in the above that forward integration of the equations of motion does, in fact, yield the correct predictions at time t; i.e., the measured X_i are equal to ensemble averages calculated from the time-developed Liouville function obeying (78), or the time-developed global density matrix obeying $r\dot{h}o = [H, \rho]$. In reference,[5] it is

shown that this is the case *if the observed change $X_i - X_i'$ is an experimentally reproducible one.* But we know that many past macroscopic states X_i'' would all relax into the same state X_i at time $t = 0$ Thus, we suggest that the correct statement of the second law is that spontaneous decreases in the experimental entropy, although not absolutely prohibited by the laws of physics, *cannot occur in an experimentally reproducible process.*

Statement IV now follows from the fact that nothing in the above reasoning restricts us to equilibrium states. In conventional thermodynamics, the experimental entropy is defined only for equilibrium states; however, our definition $S_e \equiv [\max S_I$ over all probability distributions that agree with the data in the sense of (52)] defines a function $S_e\,(X_1 \ldots X_n)$ of the experimentally measured parameters for the equilibrium or nonequilibrium case, which by the above arguments cannot spontaneously decrease in an experimentally reproducible process. It can no longer be found by numerical integration of dQ/T over a reversible path; but the content of statement IV is that a function S_e still *exists*, such that the usual statement of the second law remains valid. it requires a great deal more analysis, to be given elsewhere, before we can reduce this to a suggestion of a definite experiment that could test statement IV; I am trying here only to point out in the briefest terms why it is that an extension of the second law is predicted by theory as soon as we have understood everything revealed by Gibbs about the connection between entropy and probability.

Finally, we note that the Boltzmann H-theorem, whether correct or not, cannot have any real relevance to the second law. For, summarizing the above inequalities,

$$-kH_B \geq -kH_G \leq S_e \tag{102}$$

where the first inequality becomes an equality if and only if there are no interparticle correlations (i.e., ideal Boltzmann gas), the second if and only if H_G is computed from the appropriate canonical distribution. Obviously, whether H_B increases or decreases allows us to infer nothing about S_e. The situation is even worse than that; for the Boltzmann H-theorem was based on incorrect equations of motion, and whether H_B increases or decreases depends on the form of the distribution and the force law, To see this, note that from (98) and the exact equation of motion (94), the exact rate of change of H_B is just the negative of the expected divergence in momentum space of the molecular force $\langle F_\alpha \rangle$:

$$\dot{H}_B = -\left\langle \frac{\partial \langle F_\alpha \rangle}{\partial p_\alpha} \right\rangle \tag{103}$$

and this can have either sign. For example, if $\langle F_\alpha \rangle$ is dominated by a "dragging" term as in the Langevin equation: $\langle F_\alpha \rangle = -Kp_\alpha + \ldots$ then we find that the exact equations give us an "anti-H-theorem," $\dot{H}_B > 0$

7 Conclusion

We have seen that the principle of maximum entropy leads immediately to the same final rules of calculation that conventional statistical mechanics had provided only after long and inconclusive discussion of phase space, ergodicity, metric transitivity, etc.; and then only for the equilibrium case. The viewpoint advocated here thus represents, from the pedagogical standpoint, a considerable simplification of the subject. But this agreement also means that, from a pragmatic standpoint, if there is any new content in this principle, we must look for it in the extension to the statistical mechanics of irreversible processes, where there does not exist at present any general formal theory, and ask whether the principle of maximum entropy provides such a basis. Over the past several years, my students and I have verified that all the commonly accepted principles of irreversible statistical mechanics can be derived from this formalism; that is, of course, a minimum requirement that any proposed new theory must pass. The real test of these ideas can come only through their application to problems that have resisted solution by older methods. Although a few results along this line are now in,[11] and a few others have been hinted at in these talks, a final settlement of the questions raised still lies rather far in the future.

References

[1] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x.

[2] Edwin T. Jaynes. Note on unique decipherability. *IRE Transactions on Information Theory*, 5(3): 98–102, 1959. doi: 10.1109/TIT.1959.1057500.

[3] Edwin T. Jaynes. New engineering applications of information theory. In J. Bogdanoff and F. Kozin, editors, *Engineering Uses of Random Function Theory and Probability*, pages 163–203. Wiley, New York, 1963.

[4] Edwin T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957. doi: 10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620.

[5] Edwin T. Jaynes. Information theory and statistical mechanics. II. *Phys. Rev.*, 108:171–190, Oct 1957. doi: 10.1103/PhysRev.108.171. URL https://link.aps.org/doi/10.1103/PhysRev.108.171.

[6] Max Dresden. Recent developments in the quantum theory of transport and galvanomagnetic phenomena. *Rev. Mod. Phys.*, 33:265–342, Apr 1961. doi: 10.1103/RevModPhys.33.265. URL https://link.aps.org/doi/10.1103/RevModPhys.33.265.

[7] Edwin T. Jaynes. Review of the algebra of probable inference by Richard T. cox. *American Journal of Physics*, 31(1):66–67, 1963. doi: 10.1119/1.1969248. URL https://doi.org/10.1119/1.1969248.

[8] Edwin T. Jaynes. *Probability Theory in Science and Engineering*. McGraw Hill, 1962. in press.

[9] Harold Jeffreys. *The Theory of Probability*. Oxford Classic Texts in the Physical Sciences. 1939. ISBN 978-0-19-850368-2, 978-0-19-853193-7.

[10] Paul Ehrenfest and T. Ehrenfest. Begriffliche Grundlagen der statistischen Auffassung in der Mechanik. *Encyklopädie der mathematischen Wissenschafte*, IV(32):1–90, 1911.

[11] S. P. Heims and Edwin T. Jaynes. Theory of gyromagnetic effects and some related magnetic phenomena. *Rev. Mod. Phys.*, 34:143–165, Apr 1962. doi: 10.1103/RevModPhys.34.143. URL https://link.aps.org/doi/10.1103/RevModPhys.34.143.