CHAPTER 30

MAXIMUM ENTROPY: MATRIX FORMULATION.

Back in Chapter 11, we saw how the principle of maximum entropy leads us to a general means of assigning probabilities. From a theoretical standpoint, that discussion contained a fairly complete treatment of the general formal properties of MAXENT distributions. But from a pragmatic viewpoint, Chapter 11 was left incomplete, in a way that appears as soon as we try to apply the formalism to real, nontrivial problems; we need more powerful mathematical tools.

There is a second point to be made: our MAXENT formalism contained as a special case the standard Gibbs formalism of equilibrium statistical mechanics, via arguments very much shorter and simpler than the "ergodic" approach of antiquity. In statistical mechanics, therefore, the MAXENT principle is, at the very least, a useful pedagogical device, by which known results may be derived more quickly. But, of course, the real test of any new principle in science is not its ability to re–derive known results, but its ability to give new results, which could not be (or at least, had not been) derived without it. But since we agree with standard formalism in all equilibrium problems, the only place where new results are possible is in the extension to nonequilibrium problems, where previously no general theory existed.† How is this extension to be made in our formalism?

It is one of the most satisfying things about this approach that both these needs – finding a mathematical technique for complicated problems, and setting up a general formalism for nonequilibrium problems – are met by a single mathematical development. The basic mathematical facts to be explained here were found long ago by John von Neumann [Göttinger Nachrichten, 1927], but their full significance could be seen only after the MAXENT principle had been recognized, and the re–interpretation of probability theory as extended logic had been developed.

Density Matrix Formulation

First, let us consider Statistical Mechanics in quantum theory. In Chapter 11 we have developed a formalism in which the enumeration of the possible "states of nature" could take place simply by listing all the stationary quantum states. In other words, quantities that are constants of the motion are the only things that we have allowed ourselves to specify so far. Evidently, if we are ever going to get to non–equilibrium theory, we have to generalize this to the case where we put in information about things which are not constants of the motion, so something can happen when we let the equations of motion take over. If we started out with the initial canonical probability assignments of Chapter 11 and then solved the Schrödinger equation for the time development, we would find nothing at all happening. It would just sit there. Of course, that is as it ought to be for the equilibrium case; but for the non–equilibrium case, we need a bit more.

Also, as just noted, even in the equilibrium case, we need to generalize this before we can actually do the calculation for nontrivial physical problems, because in practice we don't have the kind of information assumed above. The theory given so far presupposes an enumeration of the exact energy levels in our system to start with. But in a realistic problem, we can't calculate these.

† The fundamental postulate of ergodic theory was that ensemble averages are equal to time averages. It would follow that, in equilibrium problems where there is no time dependence, ensemble averages are also equal to experimental values. Obviously, such a theory is helpless to deal with the time–dependent nonequilibrium problems, where the very facts to be explained are that ensemble averages are *not* equal to time averages; but they are still equal to experimental values.

What we know is a Hamiltonian operator which, in the cases we can actually solve, can be split into a term H_0 which is big but simple and another term H_1 which is complicated but small,

$$H = H_0 + H_1 \tag{30–1}$$

Then we have to do some kind of perturbation theory in order to find approximate values for the energy levels defined by the entire Hamiltonian. To find them exactly is a problem that we haven't solved. It will happen in all nontrivial problems that H_1 does not commute with H_0. So we have to learn how to generalize the mathematical machine so we can put in information about quantities which don't commute with each other. We can't enumerate states of nature simply by citing energy levels; in fact we don't even know the representation in which this would be possible. For this reason, in any representation we can find, the relative phase of these quantum states has to get into the picture even for equilibrium problems. The way to do this is to restate this theory in terms of the density matrix.

First, let's recall our basic definition of the density matrix. This is perfectly standard material which is in a hundred textbooks on quantum theory and statistical mechanics. Suppose we have a state of knowledge about a system; and for the time being, don't worry about how we got this state of knowledge. We just want to describe it. Our system contains n moving particles with coordinates $\{x_1(t), \cdots x_n(t)\}$, and in quantum theory we describe our state of knowledge about them by a wave function, or "state vector"

$$\Psi(x_1 \cdots x_n)$$

But this describes the *maximum* amount of information permitted by quantum theory. In most cases there are various states $\Psi_1, \Psi_2, \ldots$, in which the system might be, and we don't know which one it is. All we know is described by assigning some probability w_1 to it being in state Ψ_1. Now, if we knew the system was in a definite quantum state Ψ_i and we wanted to predict the value of some physical quantity F like momentum or magnetization, we represent this by some Hermitian operator F_{op}, whereupon the expectation of F in state Ψ is, according to quantum theory,

$$\langle F \rangle_i = \int \Psi_i^* F_{op} \Psi_i d\tau \tag{30–2}$$

where $\int d\tau$ stands for an integration over all particle co–ordinates x_i, and, if there are spin indices s_i in the problem, for summation over all those. Now the Ψ_i functions that we started with are not necessarily orthogonal functions. They could be any old set of conceivable states of the system. But each of them could be expanded in a complete orthogonal set. Let's say that u_k are a complete orthonormal set of functions in which we can expand any state of this system. For the moment, it doesn't matter what states they are; just any set that we know is complete. We could expand Ψ_i in terms of those, getting some expansion coefficients $a_k^{(i)}$:

$$\Psi_i = \sum_k u_k a_k^{(i)} \tag{30–3}$$

and then write

$$\langle F \rangle_i = \int \left(\sum_k u_k a_k^{(i)} \right)^* F_{op} \left(\sum_j u_j a_j^{(i)} \right) d\tau. \tag{30–4}$$

Now the a_k^* and a_j are constants which can be taken outside the integral,

$$\langle F\rangle_i = \sum_{kj} a_k^{*(i)} a_j^{(i)} \int u_k^* F_{op} u_j \, d\tau \tag{30–5}$$

and the integral (or sum)

$$\int u_k^* F_{op} u_j \, d\tau \equiv F_{kj} \tag{30–6}$$

defines the *matrix elements* F_{kj} of F in the u_k representation, so that

$$\langle F\rangle_i = \sum_{kj} F_{kj} a_k^{(i)*} a_j^{(i)}. \tag{30–7}$$

The expectation of any quantity, if we are given the wave function Ψ_i, is a quadratic form in these matrix elements F_{kj}.

Now if we're in this fix where we don't know what the state is, the best expectation value we can give you is not just one of these, but we have to average it also over these w_i which represent our uncertainty as to what the actual state is,

$$\langle F\rangle = \sum_i w_i \langle F\rangle_i = \sum_i w_i \sum_{jk} F_{kj} a_k^{(i)*} a_j^{(i)}. \tag{30–8}$$

Our expectation values are now double averages. Even if we know the exact quantum state, there are still statistical things in quantum theory (or, to put it more cautiously, in the current "Copenhagen" interpretation of that theory), which would allow us to give only expectations in general. We're not even that well off. We don't even know what the right state is, so we have to average over the ignorance (w_i) also.

When you have a thing like (30–8), the only thing you can possibly do with it is change the order of summations and see what happens. Let us do that;

$$\langle F\rangle = \sum_{jk} F_{kj} \sum_i w_i a_k^{(i)*} a_j^{(i)}$$

Now, define a matrix ρ by

$$\sum_i w_i a_k^{(i)*} a_j^{(i)} = \rho_{jk} \tag{30–9}$$

then

$$\langle F\rangle = \sum_{jk} F_{kj} \rho_{jk}. \tag{30–10}$$

The summation over j builds the matrix product $F\rho$; and then the summation over k is the sum of the diagonal elements, which we call the trace. Or, we could have written the sum with ρ and F interchanged. In this case we would now say the summation over k builds us the matrix product ρF, and then the summation over j gives the trace, so we could write this equally well as

$$\langle F\rangle = \sum_{jk} F_{kj} \rho_{jk} = \mathrm{Tr}(F\rho) = \mathrm{Tr}(\rho F). \tag{30–11}$$

This matrix ρ is, of course, called the *density matrix*, and you see that it is a Hermitian matrix, $\rho_{kj}^* = \rho_{jk}$, or in matrix notation

$$\rho^\dagger = \rho. \tag{30–12}$$

The neat way to develop our quantum statistics, so the phases are taken into account automatically, is in terms of the density matrix. From now on we will express expectations of any quantities we want to talk about in the form (30–11). We started out with a problem of how you set up a probability assignment which describes a certain state of knowledge; now we have the problem of setting up a density matrix which describes a certain state of knowledge.

Take a specific case; suppose somebody measures the total magnetic moment of some spin system and they give us a number M. We want to find a density matrix which describes what we know about the spin system when we have just this number; or rather these three numbers, the three components $\{M_x, M_y, M_z\}$. At the very least we want the density matrix to satisfy

$$\vec{M} = \langle \vec{M}_{op} \rangle = \mathrm{Tr}\left(\rho \vec{M}_{op}\right). \tag{30–13}$$

In other words, if we give this density matrix to anybody else, and he tries to predict the moments from the density matrix, he should be able to get back the numbers that were given to us, by following the usual rule for prediction in statistical mechanics. If he couldn't do that, then it wouldn't make sense to say that the density matrix "contained" the given information $\{M_x, M_y, M_z\}$.†

In general, there are an infinite number of density matrices which would all do this. Again, we are faced with the problem of making a free choice of a density matrix, which is "honest" in the sense that it doesn't assume things that we don't know, and spreads out the probability as evenly as possible over all possibilities allowed by what we do know. We do this by maximizing an entropy; but what is the appropriate entropy now? We started out in Chapter 11 with the information entropy

$$S_I = -\sum_i p_i \log p_i$$

so, suppose we now take

$$S_A = -\sum_i w_i \log w_i \tag{30–14}$$

and we might choose the density matrix which makes S_A a maximum. But if we took that as our measure of amount of uncertainty, we would be in big trouble. A sort of Gibbs paradox would show up, as a consequence of the fact that the initial states Ψ_i are not necessarily orthogonal to each other. We can have Ψ_1 and give it a probability w_1; and to the state Ψ_2 we give probability w_2. Now, let's make a continuous change in the problem such that $\Psi_1 \to \Psi_2$; our state of knowledge shades continuously into: Ψ_1 with a probability $(w_1 + w_2)$. But nothing like that happens to S_A. In S_A as $\Psi_2 \to \Psi_1$ the term $w_1 \log w_1 + w_2 \log w_2$ would have to be replaced suddenly by

$$(w_1 + w_2)\log(w_1 + w_2).$$

If we took this quantity S_A as the measure of uncertainty about the system, then you would have this phenomenon of sudden discontinuities in our uncertainty when two wave functions became exactly equal. But our intuitive state of knowledge has no discontinuity when we do that. It goes

† This is all we are doing when we choose ρ to satisfy (30–13); but for reasons we do not understand, this step seems to cause major conceptual hangups for some, who think that we are "measuring an expectation value". Of course, that just does not make sense; we are *choosing* a density matrix so that its expectation agrees with the measurement.

continuously from one case to another. That's one thing that would be wrong if we tried to use this S_A as a measure of uncertainty.

There's another thing that would be even worse, and perhaps easier to see. For a given density matrix, there's no upper limit to the S_A that we could get. If S_A is going to be the thing that counts, let's say we have 26 different states, Ψ_a to Ψ_z. They all happen to be equal to Ψ_1 but we assign probabilities w_a to w_z to them. Now, of course, the summation

$$-\sum_{a=1}^{26} w_a \log w_a$$

over the alphabet (this notation is not quite consistent, but I think you see the point) – the summation over all these terms could be a very large number. We can introduce thousands of them. There would be no upper limit to the $-\sum w \log w$ we could get if we used this S_A.

On the other hand, there's one property that is unique. S_A has no upper bound, but it does have a lower bound. S_A for a given density matrix has an absolute minimum given by

$$S_A \geq -\mathrm{Tr}(\rho \log \rho). \tag{30–15}$$

There's one and only one way, in general, of setting up these states Ψ_i and corresponding probabilities w_i so that this lower bound is reached. When we say "in general," we mean if there are no degeneracies in the eigenvalues of ρ. The simple proof is given in many places, for example Jaynes (1957b), but the reader should be able to work it out for himself.

Well, now what does $\log \rho$ mean? There's a theorem in matrix theory that says: if ρ commutes with its Hermitian conjugate $[\rho\rho^\dagger = \rho^\dagger\rho]$, there is a matrix S such that the eigenvalues $\{\rho_1, \rho_2, \cdots\}$ of ρ are displayed explicitly:

$$S\rho S^{-1} = \begin{pmatrix} \rho_1 & & & \\ & \rho_2 & & \\ & & \ddots & \\ & & & \rho_n \end{pmatrix} \tag{30–16}$$

Since ρ is itself Hermitian, this necessary and sufficient condition is met, so we can always find some similarity transformation which would have made it diagonal. Now, in the representation where ρ is diagonal, then by $\log \rho$ we mean the diagonal matrix

$$\log \rho = \begin{pmatrix} \log \rho_1 & & & \\ & \log \rho_2 & & \\ & & \ddots & \\ & & & \log \rho_n \end{pmatrix} \tag{30–17}$$

If we choose for our basis u_k the particular set of functions Ψ_i for which S_A does reach its absolute minimum value, then the diagonal elements of ρ are just the probabilities w_i assigned to these states. In other words, the choice of possible states Ψ_i which makes S_A a minimum for a given ρ, is the one for which the probabilities w_i are the eigenvalues of this matrix ρ.

If states Ψ_1 and Ψ_2 are not orthogonal and you tell me the system is in state Ψ_1, then, of course, the present Copenhagen interpretation says: the probability that, if I did a measurement, I would actually find it in Ψ_2, is not zero. It's the scalar product squared, $|(\Psi_1, \Psi_2)|^2$; sometimes called a transition probability from one state to another. We are not writing down the probabilities of mutually exclusive events unless we choose our states Ψ_i to be orthogonal, and that's just what we do by making the choice that minimizes S_A. I'm going to say now that the von Neumann information entropy S_I for a density matrix is this unique minimum value of S_A:

$$S_I \equiv (S_A)_{min} = -\mathrm{Tr}(\rho \log \rho) = -\sum \rho_i \log \rho_i \tag{30–18}$$

which is just the Shannon entropy that we used in Chapter 11, now based on the eigenvalues ρ_i of ρ. For a system described by the density matrix ρ, (30–18) is the quantity that measures the effective number of microstates in which the system might be. There are a number of other arguments why you choose (30–18) rather than some other expressions that you could think of, and they are also given in this previously mentioned paper.

Generality of the Formalism. This makes another point evident; we have been thinking in terms of quantum theory, where the density matrix is a virtual necessity for any nontrivial calculation. But since the entropy expressions are really the same, we can equally well consider any problem with discrete probabilities $\{p_1 \cdots p_n\}$ which has nothing to do with quantum theory (they might refer to a problem in economics), and define a matrix with the p_i down the main diagonal:

$$\rho = \begin{pmatrix} p_1 & & & \\ & p_2 & & \\ & & \ddots & \\ & & & p_n \end{pmatrix} \tag{30–19}$$

Then everything we can do with the probabilities $\{p_1 \cdots p_n\}$ we can do as well with the matrix ρ. If it is a help for any calculation, we are free to carry out similarity transformations and work with

$$\rho' \equiv S \rho S^{-1} \tag{30–20}$$

which has off–diagonal elements. Thus all the following formalism, developed originally for quantum theory, can be used as well for any problem with discrete probabilities. The expectation of any quantity $\{q_1, q_2, \cdots\}$ which we wrote before as $\langle q \rangle = \sum p_i q_i$, can now be written equally well as $\langle q \rangle = \mathrm{Tr}\rho q$, where q is a vector with components q_i. The only difference is that in quantum theory it is generally the matrix ρ' that we meet with first, and it may be a difficult problem to find ρ from it. In practice, we must resort usually to some approximation method, of which a perturbation expansion is probably the best developed example.

So, from this point on we may interpret the equations either as referring to quantum theory, or to general problems with discrete probabilities p_i

Setting up the Formalism: Now, we are back at the same problem that we studied in Chapter 11, but the F_k are matrices, and the constraints are

$$\langle F \rangle_k = \mathrm{Tr}(\rho F_k), \qquad k = 1, 2, \ldots, m. \tag{30–21}$$

We are to find the density matrix that maximizes $S_I \equiv -\mathrm{Tr}\rho \log \rho$ while agreeing with the conditions (30–21). Now, the formal solution of this goes through in exactly the same way as we did in Chapter 11. You recall that our proof back then was based on the fact that when we have an ordinary discrete probability distribution

$$\sum_{i=1}^{n} p_i \log p_i \geq \sum_{i=1}^{n} p_i \log u_i \tag{30–22}$$

The inequality, given by J. Willard Gibbs (1902), becomes an equality if, and only if, $p_i = u_i$. Now, we have a precisely similar situation here. You can prove that if ρ and σ are any two density matrices, there is an inequality

$$\mathrm{Tr}\,(\rho \log \rho) \geq \mathrm{Tr}\,(\rho \log \sigma) \tag{30–23}$$

with equality if and only if $\rho = \sigma$. We'll leave this as an "exercise for the reader" to prove. The argument goes through in much the same way that we did it before. The density matrix that maximizes S_I subject to these constraints is again given by

$$\rho = \frac{1}{Z(\lambda_1 \dots \lambda_m)} \exp\{-\lambda_1 F_1 - \cdots - \lambda_m F_m\} \tag{30–24}$$

One would guess, of course, that it generalizes in some such way as this, but intuition would not tell us whether the proper generalization was exactly this form. All the formal properties that we noted in Chapter 11 follow from this distribution in just the same way that we gave before – with one exception, arising from the fact that the F_k do not necessarily commute, which we'll get to after we've developed our mathematics a little bit more.

Of course the number one must have expectation value of one

$$\langle 1 \rangle = \mathrm{Tr}(\rho 1) = \mathrm{Tr}(\rho). \tag{30–25}$$

This is one more condition just like the one that $\sum p_i$ had to be equal to one. The normalizing factor which will guarantee this, is evidently

$$Z(\lambda_1 \dots \lambda_m) = \mathrm{Tr} \exp\{-\lambda_1 F_1 - \cdots - \lambda_m F_m\} \tag{30–26}$$

which is the *partition function* that we used already in Chapter 9 to solve combinatorial problems.

Perhaps we ought to say a word about what is meant by the exponential of a matrix. If we have a function of an ordinary number x that we can expand in a power series,

$$f(x) = \sum_{n=0}^{\infty} a_n x^n, \tag{30–27}$$

of course, there is nothing to stop us from defining the same function of a matrix M by the same power series,

$$f(M) \equiv \sum_{n=0}^{\infty} a_n M^n. \tag{30–28}$$

Then the question arises; does this converge to a definite matrix and if so does the resulting matrix function $f(M)$ have any useful properties? There is a theorem: if the original power series converged for x equal to each of the eigenvalues of the matrix M, then the matrix power series is guaranteed to converge to a definite matrix $f(M)$. This is obvious from (30–17) if M can be diagonalized; but it remains true for any square matrix. Now in particular the exponential function,

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \tag{30–29}$$

converges so well that it has infinite radius of convergence and, therefore, the exponential of a square matrix with finite elements is guaranteed to exist and to be a well defined matrix.

The choosing of the λ_k is again something which we do in order to make the expectation values agree with the given data. Again it's going to turn out that same formal relations hold when we are talking matrices. Again we have to solve

$$\langle F_k \rangle = -\frac{\partial}{\partial \lambda_k} \log Z \tag{30–30}$$

for the λ_k. But to prove that this is right, we have to give a mathematical argument that is a little more involved than that needed to prove (11–43), because the different F_k need not commute with each other. It turns out that this argument is also fundamental to everything that we want to discuss from now on, so let's take time out for it now.

Heims Perturbation Theory

We want to develop a general perturbation theory in which if there's a complicated problem we can break it down into a simple problem plus a small change. We want to expand this density matrix in powers of some small perturbation, and the perturbation theory we get for equilibrium will also be exactly the one we need for our irreversible theory.

This was worked out in about 1959 by the writers', former student, Steve Heims. It appears in his doctoral thesis (Stanford, 1962) and we published a condensed account of it in the appendix to a paper on gyromagnetic effects [*Revs. Mod. Physics* **34,** 143 (1962)]. You see we have always the problem of evaluating exponentials of matrices. First, I would like to work out the well–known perturbation expansion of this, then convert it into the Heims expansion for expectations. We have a matrix A, and the matrix e^A is something that we can calculate. That is simple; but the thing we really want to calculate is

$$\exp\,(A\, +\, \text{something else})$$

or

$$e^{A+\epsilon B} = e^A \left[1 + \sum_{n=1}^{\infty} \epsilon^n S_n \right]. \tag{30–31}$$

We indicate that this something else is small by putting ϵ in it and expanding in powers of ϵ. You see this is the typical situation we would have if we tried to evaluate a density matrix

$$\rho = \frac{1}{Z} \exp\left\{ -\lambda_1 F_1 - \cdots - \lambda_m F_m \right\}. \tag{30–32}$$

Some of these operators might be simple so we could evaluate their exponentials; then some others might be complicated and not commute with the others, and they would mess up the whole problem. At that point we would resort to approximations. To put it in general form, let's talk just A and B for a while. Form a quantity

$$e^{-xA}\; e^{x(A+\epsilon B)}$$

where x is an ordinary number and by xA we mean the matrix in which every element is multiplied by x. If $\epsilon \to 0$, this goes into the unit matrix. But it isn't quite the unit matrix, if $\epsilon > 0$. But how does it vary with x? Well, by starting at this power series definition of the exponential function, you can convince yourself very quickly that the same rule of differentiating an exponential function works even if a matrix is in the exponent. We have the option of writing it either way:

$$\frac{d}{dx}e^{-xA} = -Ae^{-xA} = -e^{-xA}A. \tag{30–33}$$

Therefore,

$$\frac{d}{dx}\left[e^{-xA}e^{x(A+\epsilon B)}\right] = -e^{-xA}Ae^{x(A+\epsilon B)} + e^{-xA}(A+\epsilon B)e^{x(A+\epsilon B)} \tag{30–34}$$

Now two terms cancel, and ϵ is just a number, so

$$\frac{d}{dx}\left[e^{-xA}\,e^{x(A+\epsilon B)}\right] = \epsilon e^{-xA}\,B\,e^{x(A+\epsilon B)}. \tag{30–35}$$

We can't pull that B outside because in general it doesn't commute with what is either to the left of it or to the right of it. Now that we have differentiated this, let us integrate with respect to x and get it back again:

$$\begin{aligned}\int_0^x \frac{d}{dx_1}\left[e^{-x_1 A}e^{x_1(A+\epsilon B)}\right]dx_1 &= e^{-xA}e^{x(A+\epsilon B)} - 1 \\ &= \epsilon\int_0^x e^{-x_1 A}Be^{x_1(A+\epsilon B)}dx_1.\end{aligned} \tag{30–36}$$

To clean this up, multiply both sides by e^{xA} from the left. We find

$$e^{x(a+\epsilon B)} = e^{xA}\left[1+\epsilon\int_0^x e^{-x_1 A}Be^{x_1(A+\epsilon B)}dx_1\right]. \tag{30–37}$$

This is an integral equation which $e^{x(A+\epsilon B)}$ satisfies. Well now, if you have an integral equation, you grind out perturbation solutions of it simply by iteration; that is, substituting the equation into itself over and over again. The first iteration gives

$$\begin{aligned}e^{x(A+\epsilon B)} &= e^{xA}\left\{1+\epsilon\int_0^x dx_1 e^{-x_1 A}Be^{x_1 A}\left[1+\epsilon\int_0^{x_1} dx_2 e^{-x_2 A}Be^{x_2(A+\epsilon B)}\right]\right\} \\ &= e^{xA}\left\{1+\epsilon\int_0^x dx_1 e^{-x_1 A}Be^{x_1 A} + \epsilon^2\int_0^x dx_1\int_0^{x_1} dx_2 e^{-x_1 A}Be^{(x_1-x_2)A}Be^{x_2(A+\epsilon B)}\right\},\end{aligned}$$

and by repeated substitution we get

$$\begin{aligned}e^{A+\epsilon B} = e^{A}\Bigg[&1+\epsilon\int_0^1 e^{-xA}Be^{xA}dx \\ &+\epsilon^2\int_0^1 dx_1\int_0^{x_1} dx_2 e^{-x_1 A}Be^{(x_1-x_2)A}Be^{x_2 A} \\ &+\epsilon^3\int_0^1 dx_1\int_0^{x_2} dx_2\int_0^{x_2} dx_3 e^{-x_1 A}Be^{(x_1-x_2)A}Be^{(x_2-x_3)A}Be^{x_3 A} \\ &+\cdots\qquad\Bigg].\end{aligned} \tag{30–38}$$

We can keep playing this game as long as we please, and so this generates an infinite series in powers of ϵ. Or, we can terminate (30–38) at any finite number of terms, replace A by $A+\epsilon B$ in the last exponent, and it is an exact equation. The exponential of any matrix is a well–behaved thing, so we can put in any ϵ we please – large or small – and the infinite series is guaranteed to converge to the right thing. Of course, if we have to take more than about two terms of the series, then we'll be wound up in another bad calculation and this whole method will not be too useful.

Let's summarize this: we have found the power series expansion

$$e^{A+\epsilon B} = e^A \left[1 + \sum_{n=1}^{\infty} \epsilon^n S_n\right] \tag{30–39}$$

in which

$$S_1 \equiv \int_0^1 e^{-xA} B e^{xA} dx \tag{30–40}$$

$$S_2 \equiv \int_0^1 dx_1 \int_0^{x_1} dx_2 e^{-x_1 A} B e^{(x_1 - x_2)A} B e^{x_2 A} \tag{30–41}$$

and if we write

$$B(x) \equiv e^{-xA} B e^{xA} \tag{30–42}$$

the general order term is

$$S_n \equiv \int_0^1 dx_1 \int_0^{x_1} dx_2 \cdots \int_0^{x_{n-1}} dx_n B(x_1) B(x_2) \cdots B(x_n). \tag{30–43}$$

Now we have an "unperturbed" density matrix

$$\rho_0 = \frac{e^A}{\mathrm{Tr}\,(e^A)} \tag{30–44}$$

and a "perturbed" one in which some kind of additional information is put in:

$$\rho = \frac{e^{A+\epsilon B}}{\mathrm{Tr}\,[e^{A+\epsilon B}]} \tag{30–45}$$

How did this additional information affect our prediction of some quantity C? In the unperturbed ensemble, any operator C has the expectation

$$\langle C \rangle_0 = \mathrm{Tr}(\rho_0 C) \tag{30–46}$$

and in the perturbed ensemble, it will be instead,

$$\langle C \rangle = \mathrm{Tr}(\rho C). \tag{30–47}$$

And what we really want is a power series expansion of $\langle C \rangle$. So let's write out the expansion we would like to get; using (30–39),

$$\langle C \rangle = \frac{\mathrm{Tr}\,\left[e^{A+\epsilon B} C\right]}{\mathrm{Tr}\,[e^{A+\epsilon B}]} = \frac{\mathrm{Tr}\,(e^A C) + \sum_{n=1}^{\infty} \epsilon^n \mathrm{Tr}\,\left(e^A S_n C\right)}{\mathrm{Tr}\,(e^A) + \sum_{n=1}^{\infty} \epsilon^n \mathrm{Tr}\,(e^A S_n)}$$

and divide by $\mathrm{Tr}\,\left(e^A\right)$ to get, from (30–46),

$$\langle C\rangle = \frac{\langle C\rangle_0 + \sum_{n=1}^{\infty}\epsilon^n\langle S_n C\rangle_0}{1+\sum_{n=1}^{\infty}\epsilon^n\langle S_n\rangle_0} \tag{30–48}$$

We now have everything reduced to expectations over the unperturbed distribution, which we assumed was something simple that we could calculate. But still this is in a little messy form. We have the ratio of two infinite series, which we know are well–behaved. Both the numerator and denominator series have infinite radius of convergence. But, we would like to write this as a single series over ϵ and get rid of this denominator. If we can invert the power series for this denominator; that is, find the coefficients a_n in

$$\frac{1}{1+\sum_{n=1}^{\infty}\epsilon^n\langle S_n\rangle_0} = 1-\sum_{n=1}^{\infty} a_n\epsilon^n,$$

then we'll have it. This equation is the same as

$$1 = \left(1+\sum_{n=1}^{\infty}\epsilon^n\langle S_n\rangle_0\right)\left(1-\sum_{m=1}^{\infty}\epsilon^m a_m\right)$$

or, after careful manipulation of indices in the double sum,

$$1 = 1+\sum_{n=1}^{\infty}\epsilon^n\left[\langle S_n\rangle_0 - a_n - \sum_{k=1}^{n-1}\langle S_k\rangle_0 a_{n-k}\right].$$

Now since different powers of ϵ are linearly independent functions, if a power series in ϵ is to vanish identically (*i.e.*, for all ϵ), the coefficients of each term must be zero separately. So, the problem is: choose the a_n so that

$$\langle S_n\rangle_0 = a_n + \sum_{k=1}^{n-1}\langle S_k\rangle_o a_{n-k}. \tag{30–49}$$

This is a discrete version of a Volterra integral equation, and is solved as follows. Define a sequence of operators Q_n,

$$Q_1 \equiv S_1 \tag{30–50}$$

$$Q_2 \equiv S_2 - S_1\langle Q_1\rangle_0 \tag{30–51}$$

$$Q_n \equiv S_n - \sum_{k=1}^{n-1} S_k\langle Q_{n-k}\rangle_0, \qquad n>1 \tag{30–52}$$

Taking the expectation of (30–52) and comparing with (30–49), we see that the desired solution is just

$$a_n = \langle Q_n\rangle_0, \qquad n\ge 1 \tag{30–53}$$

Now, returning to (30–48) with this result, we have

$$\langle C\rangle = \left[\langle C\rangle_0 + \sum_{k=1}^{\infty}\epsilon^k\langle S_k C\rangle_0\right]\left[1-\sum_{m=1}^{\infty}\epsilon^m\langle Q_m\rangle_0\right]. \tag{30–54}$$

In expanding this, note that the double sum can be written as

$$\sum_{k=1}^{\infty}\sum_{m=1}^{\infty} \epsilon^{k+m}\langle S_k C\rangle_0 \langle Q_m\rangle_0 = \sum_{n=2}^{\infty}\epsilon^n \sum_{k=1}^{n-1}\langle S_k C\rangle_0\langle Q_{n-k}\rangle_0 \tag{30–55}$$

and we might as well add the term with $n = 1$, since it vanishes anyway, having no terms at all. So, we have

$$\langle C\rangle = \langle C\rangle_0 + \sum_{n=1}^{\infty}\epsilon^n \left[\langle S_n C\rangle_0 - \sum_{k=1}^{n-1}\langle S_k C\rangle_0\langle Q_{n-k}\rangle_0 - \langle Q_n\rangle_0\langle C\rangle_0\right] \tag{30–56}$$

and, comparing with (30–52), we get a pleasant surprise; a simple final result:

$$\langle C\rangle - \langle C\rangle_0 = \sum_{k=1}^{\infty}\epsilon^n \left[\langle Q_n C\rangle_0 - \langle Q_n\rangle_0\langle C\rangle_0\right]. \tag{30–57}$$

The n'th order contribution to the change $[(\langle C\rangle - \langle C\rangle_0)]$ is just the covariance, in the unperturbed ensemble, of Q_n with C. The first–order term in (30–57) has long been known; to the best of my knowledge, Steve Heims was the first person to see that it can be extended to all orders. In several years of living with this formula, and seeing what it can do for us, I have come to regard it as easily the most important general rule of statistical mechanics; almost every "useful" calculation in the field can be seen as a special case of it. Also, outside of statistical mechanics, almost every nontrivial application of MAXENT will be a special case of (30–57). So, this is the general perturbation expansion that we'll use.

Reciprocity Theorems: Now, the first order correction of course is always the most important one. The first order term has a symmetry property which follows from the cyclic property of the trace, Eq. (30–11). To first order, since $Q_1 = S_1$, we have simply

$$\langle C\rangle = \langle C\rangle_0 = \epsilon\left[\langle S_1 C\rangle_0 - \langle S_1\rangle_0\langle C\rangle_0\right] \tag{30–58}$$

but

$$S_1 \equiv \int_0^1 e^{-xA}Be^{xA}dx$$

so that

$$\langle S_1\rangle_0 = \int_0^1 dx\langle e^{-xA}Be^{xA}\rangle = \frac{\int_0^1 dx \,\mathrm{Tr}\left[e^{(1-x)A}Be^{xA}\right]}{\mathrm{Tr}\left(e^A\right)}. \tag{30–59}$$

Now, as in (30–11), it is true generally that $\mathrm{Tr}(FG) = \mathrm{Tr}(GF)$ even if $FG \neq GF$; and so

$$\langle S_1\rangle_0 = \frac{\int_0^1 dx \,\mathrm{Tr}\left[e^{xA}e^{(1-x)A}B\right]}{\mathrm{Tr}\left(e^A\right)} = \frac{\mathrm{Tr}\left(e^A B\right)}{\mathrm{Tr}\left(e^A\right)} = \langle B\rangle_0, \tag{30–60}$$

so the first–order correction always reduces to

$$\langle C\rangle - \langle C\rangle_0 = \epsilon\left[\int_0^1 dx\langle e^{-xA}Be^{xA}C\rangle_0 - \langle B\rangle_0\langle C\rangle_0\right]. \tag{30–61}$$

At this point, we can verify Eq. (30–30). Make the choices $A = -\lambda_1 F_1 - \cdots - \lambda_m F_m$, $\epsilon B = -\delta\lambda_k F_k$. Then $S(\lambda_1 \ldots \lambda_m) = \mathrm{Tr}\left(e^A\right)$ and from the definition of a derivative,

$$\frac{\partial \log Z}{\partial \lambda_k} = \frac{1}{Z} \lim_{\delta\lambda_k \to 0} \frac{Z\left[\lambda_1 \ldots \lambda_k + \delta\lambda_k \ldots \lambda_m\right] - Z\left[\lambda_1 \ldots \lambda_k \ldots \lambda_m\right]}{\delta\lambda_k}. \qquad (30\text{–}62)$$

In the limit $\delta\lambda_k \to 0$, only the first-order term survives, and so

$$\frac{\partial \log Z}{\partial \lambda_k} = \frac{\mathrm{Tr}\left(e^A S_1\right)}{Z\delta\lambda_k} = \frac{\langle S_1 \rangle_0}{\delta\lambda_k} = -\langle F_k \rangle_0. \qquad (30\text{–}63)$$

This is just (30–30).

Now we note a very important symmetry property; if we interchange B and C in the right-hand side of (30–70), we don't change it. The last term we have worked into a form where it is obvious. We still have to play with the first one a little. Again, let's write this as a ratio of two traces.

$$\int_0^1 dx \langle e^{-xA} B e^{xA} C \rangle_0 = \frac{\int_0^1 dx \mathrm{Tr}\left[e^{(1-x)A} B e^{xA} C\right]}{\mathrm{Tr}\left(e^A\right)} \qquad (30\text{–}64)$$

This time we choose to interchange matrices as follows,

$$\int_0^1 dx \mathrm{Tr} e^{(1-x)A} B e^{xA} C = \int_0^1 dx \mathrm{Tr}\left[e^{xA} C e^{(1-x)A} B\right]. \qquad (30\text{–}65)$$

Now for any $f(x)$, we have

$$\int_0^1 f(x) dx = \int_0^1 f(1-x) dx \qquad (30\text{–}66)$$

consequently we can write (30–70) as

$$\int_0^1 dx \mathrm{Tr}\left[e^{(1-x)A} C e^{xA} B\right], \qquad (30\text{–}67)$$

and writing this back as an expectation

$$\int_0^1 dx \langle e^{-xA} B e^{xA} C \rangle_0 = \int_0^1 dx \langle e^{-xA} C e^{xA} B \rangle_0. \qquad (30\text{–}68)$$

After all this, the only thing that has happened is that we have interchanged B and C.

Now this is a very important symmetry property. If I perturb my density matrix by adding information about B and calculating how that changes my prediction of C, it is the same as if I had perturbed my density matrix by putting in information about C and calculated how that changes the prediction of B. A whole string of reciprocity laws, found originally by physical reasoning in many different contexts, all come out of the single formula (30–68). These include not only the Onsager reciprocity laws in nonequilibrium statistical mechanics, but the Gibbs–Helmholtz equation for the voltage of a reversible electric cell in equilibrium theory; and even the Helmholtz reciprocity theorem in acoustics, and the Lorentz reciprocity law in electromagnetic theory, which are not ordinarily thought of as arising from statistical mechanics at all.

************************* MORE TO COME! *****************************