SCHRODINGER'S APPROACH TO STATISTICAL MECHANICS

PARTIAL NOTES AS OF FEB. 25, 2019

1. Random samples of a system

These notes are about the first two chapters of Schrodinger's book, "Statistical Thermodynamics."

Schrodinger describes two alternate approaches to defining what a system is. The first views a system as a collection of some large number smaller objects, each of which can lie in one of a certain number of states. A physical example would be a large collection of gas molecules, each of which can have one of a certain number of energies. The second approach, due to Gibbs, thinks of making a large number of copies of the system we actually care about, with each of the copies able to be in one of a certain number of states.

I will also deal with a third possible interpretation. In this, a system may consist of many individual elements which can each be in a certain number of different states, and we take a large number of samples from this system. The third approach is useful for connecting these ideas to a sequence of coin flips or rolls of a die.

The mathematics stays the same in all of the above approaches. I will focus in these notes on what assumptions one is actually making in order for the math to apply.

Let's start by saying that a system S is simply a finite sample space. Suppose each element of the S lies in one of ℓ possible states. This is the same as saying that there is a random variable $X : S \to \{1, \ldots, \ell\}$ such that $X(s)$ for $s \in S$ describes the state of s. We will suppose that if $X(s) = j$ then s has energy E_j. Thus if $f : \{1, \ldots, \ell\} \to \{E_1, \ldots, E_\ell\}$ is the function defined by $f(j) = E_j$, then $Y(s) = f(X(s))$ defines a random variable $Y : S \to \mathbb{R}$ giving the energy of elements of S.

Suppose now that N is a large integer and that $S(N)$ is the set of all possible sequences $\tilde{s} = (s_1, \ldots, s_N)$ of N elements of S. One can think of $\tilde{s}$ as a sequence of N samples from S. The average energy of this set of N samples is

$$E(\tilde{s}) = \frac{1}{N} \sum_{i=1}^{N} Y(s_i) = \frac{1}{N} \sum_{j=1}^{\ell} X_j(\tilde{s}) \cdot E_j \tag{1.1}$$

when $\tilde{X}_j(\tilde{s})$ is the number of components s_i of $\tilde{s} = (s_1, \ldots, s_N)$ for which $X(s_i) = j$, i.e. for which s_i is in state j.

The main question Schrodinger addresses is:

Problem 1.1. What are the "most likely" values of $\tilde{X}_j(\tilde{s})/N$ are as $\tilde{s}$ ranges over all vectors $\tilde{s} = (s_1, \ldots, s_N)$ such that $E(\tilde{s})$ is a prescribed number E.

Since we have not specified a probabilty measure on the set of $\tilde{s}$ described in this problem, we can't talk about the probability that $X_j(\tilde{s})/N$ takes on a certain value yet. What Schrodinger does is to make a certain information theory/thermodynamic hypothesis described in in (2.1) below. He then deduces from this results about the most probable values of $X_j(\tilde{s})/N$.

2. The information theory/thermodynamic hypothesis

Each $\tilde{s}$ of the kind in Problem 1.1 gives a vector

$$b(\tilde{s}) = (X(s_1), \ldots, X(s_N))$$

of states whose components are in $A = \{1, \ldots, \ell\}$. Let A^N be the set of all vectors of length N with components in A. We have a subset B of A^N defined by

$$B(E) = \{(b_1, \ldots, b_N) : \frac{1}{N}\sum_{i=1}^{N} f(b_i) = \frac{1}{N}\sum_{i=1}^{N} E_{b_i} = E\}.$$

This subset corresponds to the set of state vectors $b(\tilde{s})$ associated to N-tuples $\tilde{s} = (s_1, \ldots, s_N)$ with energy $E(\tilde{s}) = E$.

Schrodinger's computations make use of the following hypothesis:

Hypothesis 2.1. Every element of $B(E)$ is equally likely to occur.

One way to view this hypothesis is that it says that any sequence of state vectors which produces the required average energy E is equally likely to occur. To give different vectors in $B(E)$ different probabilities of occurring would require knowing something more about states than just the average energy of a large number of samples. This is an information theoretic argument; any other assignment of probabilities to $B(E)$ would imply one has more information than is supplied by the average energy of all the samples. Another way to say one is assigning all elements of $B(E)$ the same probability of coming up is that we are maximizing our uncertainty regarding which element of $B(E)$ occurs. In the absence of further data, Occam's razor says one should maximize uncertainty. A key philosophical issue is whether the absence of further data is really because nature does not produce additional constraints, or whether this is because we simply have not looked hard enough for such constraints. This is a reason that the thermodynamic hypothesis (2.1) will not always be physically reasonable.

The homework was about the following example of this set up. The sample space S equals $\{1, 2, 3, 4, 5, 6\}$, which we will think of as the 6 sides of a fair die. Let $\ell = 6$. The random variable $X : S \to \{1, 2, 3, 4, 5, 6\}$ is just the identity map. We assign state j the energy $E_j = j$, so that $Y : S \to \{1, 2, 3, 4, 5, 6\}$ is also the identity map. If we roll a fair die N times, the possible outcomes are described by the vectors $\tilde{s} = (s_1, \ldots, s_N) \in S(N)$. If the die is fair and the rolls are independent of one another, the probabiliity of a given $\tilde{s}$ appearing will be $1/\#S(N) = 1/6^N$. The subset $B(E)$ is the set of such $b(\tilde{s}) = \tilde{s}$ for which the average of the rolls is E. Since in this example, every sequence of rolls has the same likelihood of occurring, the same will be said for the elements of $B(E)$. This example is special, though, because not only are the elements of $B(E)$ equally likely to occur, we have created a scenario in which every element of $S(N)$ is equally likely to occur when we no longer care about the average energy of a sequence of rolls. Hypothesis (2.1) only concerns those state vectors which give the prescribed average energy E.

Schrodinger describes another way to view Hypothesis 2.1. Rather than thinking of the components of $\tilde{s} = (s_1, \ldots, s_n)$ as samples from S, he prefers to think of having N copies all at once of the given system. These copies are in contact with one another, but otherwise isolated from the rest of the universe. Each copy can be in one of the ℓ states. A state vector $(b_1, \ldots, b_N) \in A^N$ records the states of the copies, and $B(E)$ is the set of such combinations of states which have average energy E. The idea that all elements of $B(E)$ should be equally likely comes about from the thought that state vectors which have the same energy would be in equlibrium with each other. If two such collections of N systems came into contact, neither would make the other less or more likely to be seen. Admittedly, the implied interaction here between two such systems is a little vague, at least to me.

3. Using the multinomial theorem to guess the most likely distribution of states

We now have a sample space $B(E)$ consisting of all state vectors $\tilde{b} = (b_1, \ldots, b_N)$ giving rise to the average energy

$$E = \frac{1}{N}\sum_{i=1}^{N} f(b_i) = \frac{1}{N}\sum_{i=1}^{N} E_{b_i}.$$

We also have the probability density function on $B(E)$ which gives every element of $B(E)$ probability $1/\#B(E)$. We can define for $j = 1, \ldots, \ell$ a random variable

$$X_j : B(E) \to \{1, \ldots, \ell\}$$

such that $X(\tilde{b})$ is the number of components b_i of $\tilde{b}$ which equal j. Thus

$$X_j(\tilde{b})/N$$

is the proportion of the components of $\tilde{b}$ which equal $B(E)$.

The natural goal now is to find the expected value $E(X_j)/N$ for $j = 1, \ldots, \ell$. This would amount to determining (under hypothesis 2.1) the expected proportion of the time that a sample will come up in state j if we know that the average of all the energies of the samples is E.

Schrodinger refers to $a_j = X_j(\tilde{b})$ as the occuptation number of state j for the state vector $\tilde{b} = (b_1, \ldots, b_N)$. We have to have

$$a_1 + \cdots + a_\ell = N \tag{3.2}$$

since each of the N components of $\tilde{b}$ is in exactly one state and thus is counted exactly once on the left side of (3.2). We have

$$E = \frac{1}{N} \sum_{i=1}^{N} f(b_i) = \frac{1}{N} \sum_{j=1}^{\ell} a_j \cdot E_j \tag{3.3}$$

since in the sum $\sum_{i=1}^{N} f(b_i)$, the set of a_j values of i for which $b_i = j$ contribute a total of $a_j \cdot E_j$ to E because $f(X(b_i)) = f(j) = E_j$.

Suppose now that we ask for the probabiilty of a given set of occupation numbers $\tilde{a} = (a_1, \ldots, a_\ell)$ arising as $(X_1(\tilde{b}), \ldots, X_\ell(\tilde{b}))$ as $\tilde{b}$ ranges over $B(E)$. The odds of any $\tilde{b}$ occurring are $1/\#B(E)$. So to answer this, we need to find the number $c(\tilde{a})$ of ways of writing down a vector $\tilde{b} = (b_1, \ldots, b_N) \in A^N$ the property that exactly a_j of the b_i's are equal to j. This is the same as picking disjoint subsets of $\{1, \ldots, N\}$ of sizes $a_1, \ldots, a_\ell$. The answer is the multinomial coefficient

$$c(\tilde{a}) = \frac{N!}{a_1! \cdots a_\ell!} \tag{3.4}$$

The probability that $\tilde{a}$ will occur is then

$$\frac{c(\tilde{a})}{\#B(E)}$$

4. The Lagrange multiplier argument

What Schrodinger does is to approximate the $\tilde{a} = (a_1, \ldots, a_\ell)$ for which the requirements (3.2) and (3.3) hold and for which the multinomial coefficient in (3.4) is largest.

The idea is to use the approximation $\ln(t!) \cong t \ln(t) - t$ to write

$$\ln(c(\tilde{a})) \cong \ln(N!) - \sum_{j=1}^{\ell} (a_j \ln(a_j) - a_j)$$

To maximize this, we should minimize

$$F(a_1, \ldots, a_\ell) = \sum_{j=1}^{\ell} (a_j \ln(a_j) - a_j)$$

over all vectors $\tilde{a}$ for which the constraints (3.2) and (3.3) hold.

Regard the a_j now as real variables, rather than integer variables. We are now trying to find the minimum of the function $F(a_1, \ldots, a_\ell)$ over all vectors of non-negative numbers a_j satisfying the constraints

$$G(a_1, \ldots, a_\ell) = \sum_{j=1}^{\ell} a_j = 1$$

and

$$H(a_1,\dots,a_\ell)=\frac{1}{N}\sum_{j=1}^{\ell}a_j\cdot E_j=E$$

These two constraints actually define hyperplanes which intersect for generic values of the E_j in a a set L which is a translate of codimension two linear subspace of $\mathbb{R}^\ell$. The gradients

$$grad(G)=(\frac{\partial G}{\partial a_1},\dots,\frac{\partial G}{\partial a_\ell})=(1,\cdots,1)$$

and

$$grad(H)=(\frac{\partial H}{\partial a_1},\dots,\frac{\partial H}{\partial a_\ell})=\frac{1}{N}(E_1,\dots,E_\ell)$$

are perpendicular to the level sets of G and H respectively, and they span the orthogonal complement of the tangent plane to L at each point of L.

In order for the function F to have a local minimum at some point $\tilde{a}=(a_1,\dots,a_\ell)$ on L, the gradient of F at $\tilde{a}$ should be perpendicular to every tangent vector to L at a. If this were not true, one could move from $\tilde{a}$ in the direction of a tangent vector to L and remain in L while decreasing the value of F. (We are just reproducing here the usual Lagrange mutiplier argument). The conclusion is that there have to be constants ν and μ such that

$$\begin{aligned} grad(F) &= (\frac{\partial F}{\partial a_1},\dots,\frac{\partial F}{\partial a_\ell}) \\ &= (\ln(a_1),\dots,\ln(a_\ell)) \\ &= -\mu\cdot(E_1,\dots,E_\ell)-\nu\cdot(1,\dots,1) \end{aligned} \tag{4.5}$$

Therefore for all $1\le j,k\le\ell$ we have

$$\ln(a_j)-\ln(a_k)=-\mu E_j-\nu-(-\mu E_i-\nu)=-\mu(E_j-E_i)$$

so

$$a_j/a_k=e^{-\mu E_j}/e^{-\mu E_i}$$

We can view this as saying that

$$a_je^{\mu E_j}=c=a_ke^{\mu E_k}$$

is independent of j, so that

$$a_j=ce^{-\mu E_j}\quad\text{for}\quad j=1,\dots,\ell. \tag{4.6}$$

One can solve for c using the fact that

$$N=a_1+\dots a_\ell=c(\sum_{j=1}^{\ell}e^{-\mu E_j}).$$

So (4.6) becomes

$$\frac{a_j}{N}=ce^{-\mu E_j}=\frac{e^{-\mu E_j}}{\sum_{j=1}^{\ell}e^{-\mu E_j}}=\frac{-1}{\mu}\frac{\partial}{\partial E_j}\sum_{j=1}^{\ell}e^{-\mu E_j} \tag{4.7}$$

Here a_j/N is the proportion of the N systems which will be in state j at thermal equilibrium. Notice that all these proportions are determined by one parameter μ. Schrodinger explains why it is reasonable to define the average temperature T of the systems by

$$\mu=\frac{1}{kT} \tag{4.8}$$

when k is the so-called Boltzmann constant.

In terms of multinomial coefficients, the significance of μ is that there is in fact a one parameter family of occupation number vectors $\tilde{a}=(a_1,\dots,a_\ell)$ the multinomial coefficient $c(\tilde{a})$ is (essentially) as large as possible. As μ increases, (4.7) shows that those j for which E_j is large and positive occur with lower frequency a_j/N among all the states in the thermodynamic equilibrium associated to μ.

Since $\mu = \frac{1}{kT}$, this means that as T approaches 0 from above, the proportion of states represented by high energy states decreases. This makes heuristic sense when T is identified with the average temperature of the systems.

Here is the connection of these calculations to Shannon entropy. We have been concerned with finding for which $\tilde{a} = (a_1, \ldots, a_\ell)$ satisying (3.2) and (3.3) the multinomial coefficient $c(\tilde{a})$ in 3.4 is as large as possible. Using the Stirling approximation to factorials we said this is (approximately) the same as finding the maximum value of

$$F(a_1, \ldots, a_\ell) = \sum_{j=1}^{\ell} (a_j \ln(a_j) - a_j) = \sum_{j=1}^{\ell} a_j \ln(a_j) - \sum_{j=1}^{\ell} a_j = \sum_{j=1}^{\ell} a_j \ln(a_j) - N$$

over all such $\tilde{a}$, where we used (3.2) in the last equality. This is the same as trying to maximize

$$\sum_{j=1}^{\ell} a_j \ln(a_j)$$

since N is fixed (but large). The fraction of the N systems which will lie in state j is $p_j = a_j/N$, and

$$\begin{aligned} \sum_{j=1}^{\ell} p_j \ln(p_j) &= \sum_{j=1}^{\ell} \frac{a_j}{N} \ln(\frac{a_j}{N}) \\ &= \frac{1}{N} \sum_{j=1}^{\ell} a_j \cdot (\ln(a_j) - \ln(N)) \\ &= \frac{1}{N} \sum_{j=1}^{\ell} a_j \cdot \ln(a_j) - \frac{\ln N}{N} \end{aligned} \tag{4.9}$$

again using $a_1 + \ldots + a_\ell = 1$. So we are in fact picking $\tilde{a} = (a_1, \ldots, a_\ell)$ satisfying (3.2) and (3.3) so as to maximize the Shannon entropy

$$H_{Shannon}(p_1, \ldots, p_\ell) = \sum_{j=1}^{\ell} p_j \ln(p_j)$$

of the probabilties $p_j = a_j/N$ of the various states occurring.

5. Reformulation using the partition function

Given the energies $E_1, \ldots, E_\ell$ of the states $1, \ldots, \ell$, define the associated partition function of a real variable μ by

$$Z = \sum_{j=1}^{\ell} e^{-\mu E_j} \tag{5.10}$$

We can rewrite (4.7) as

$$\frac{a_j}{N} = \frac{e^{-\mu E_j}}{Z} = \frac{-1}{\mu} \frac{\partial}{\partial E_j} \ln Z \tag{5.11}$$

Now

$$E = \frac{1}{N} \sum_{j=1}^{\ell} a_j \cdot E_j = \sum_{i=1}^{\ell} \frac{e^{-\mu E_j}}{Z} \cdot E_j = -\frac{\partial}{\partial \mu} \ln Z \tag{5.12}$$

The remarkable thing about the paritition function is that its partial derivatives thus determine the average energy E and the probability distribution of states $(a_1/N, \ldots, a_\ell/N)$ at the thermal equilibrium associated to a given temperature $T = \frac{1}{k\mu}$.

6. Heat and work

So far we have not spoken about heat and work in thermodynamic systems. A useful example to clarify the definition is a container in which there are a large number of molecules of different kinds at the same temperature which is placed in a so-called heat bath. The system we will deal with is the set of molecules in the container. We assume that the container is in contact with the heat bath whose temperature could be different than that of the container.

Definition 6.1. (Zemansky, §4) Heat is the energy transferred between a system and its surrounding by virtue of a temperature difference only. The heat transferred is positive if it represents energy entering the system from the surroundings.

For example, if we put a heat insulator between the container and the surrounding heat bath, then one says the system represented by the container is adiabatic. In this case, there can be no heat transfer between the container and its surroundings.

Definition 6.2. (Zemansky, §3.1) If a system exerts a force on its surroundings and a displacement takes place, the amounts of work done by the system is the product of the force with the component of the displacement parallel to the force. Conversely, work can be done on the system by its surroundings in a similar way and is counted as negative work done by the system.

In the physical example above of the container of gas molecules, one could imagine some work being done on some of the molecules by the surroundings, e.g. by microwaves entering the container from the outside. At the same time, there could be a heat transfer between the container and its surroundings, as long as the boundary of the container is not insulating.

The first law of thermodynamics has to do with so-called infinitesimal changes in the system during which the system passes at all times through thermal equilibria. The law states that

$$dU = dQ + dW \tag{6.13}$$

where dU represents the change in the internal energy of a system undergoing the infinitesimal change, dW is the work done on the system by its surroundings during this change, and dQ is the heat transferred into the system. We will see that the usual scenario if dW is positive is that some of the work done on the system is not converted into an increase in the internal energy of the system. Thus $0 \leq dU < dW$ which implies dQ is negative and heat is lost by the system to its surroundings. This agrees with our intuition that the energy supplied by work on the system can't be entirely converted to an increase in the energy of the system. Formalizing this will lead to the second law of thermodynamics.

7. Schrodinger's statistical analysis of the first law: First approach

Schrodinger gives a very interesting calculation of the first law using the statistical approach described in the previous sections.

To try to stay as close to physical intuition as possible, let's suppose at first that the system S consists of some large number N or particles, each of which can be in one of ℓ states numbers $1, \ldots, N$. This is the first, naive interpretation of a system described in §1. Eventually I'll return to the other interpretations involving identical copies of a system, with each copy lying in one of ℓ possible states.

Let the energy of the j^{th} state be E_j. We suppose the average energy E of the N particles is known, and that we also know the temperature $T = \frac{1}{k\mu}$. Then we worked out the formula (5.11) for the proportion a_j/N of the N particles which will be in state j assuming the system is in thermal equilibrium and that it obeys the thermodynamic hypothesis in §2.

Schrodinger supposes now that we exert a small external force on the particles in the system. This changes the energy E_j of the those particles in the j^{th} state to $E'_j = E_j + dE_j$, where dE_j is

infinitesimially small. The work one on the system is then

$$dW = \sum_{j=1}^{\ell} a_j \cdot dE_j \tag{7.14}$$

since there are a_j particles in state j. One should imagine this is done by training some tuned microwaves on the molecules, or by some more mechanical means.

The question now is: If the energies are changes to E'_j, what will be the new occupation numbers a'_j associated to the states? These occupation numbers will change in general if the system is in thermal equilibrium, i.e. if adjusts itself so that it continues to obey the thermodynamic hypothesis. The change involved will depend on the new temperature $T' = T + dT$ of the system. One should remember that the system may or may not be able to tranfser heat across it's boundary with the surroundings, depending on whether this boundary is insulating (adiabatic) or not.

Write

$$da_j = a'_j - a_j \tag{7.15}$$

The new average energy of the system is

$$E' = \sum_{j=1}^{N} a'_j E'_j = \sum_{j=1}^{N} (a_j + da_j) \cdot (E_j + dE_j) \tag{7.16}$$

Since we are talking about infinitesimal changes, the product of two infinitesimals is taken to be 0. (One can be more precise by using derivatives and differentials, as in the next section.) Then E' can be written as

$$E' = \sum_{j=1}^{N} a_j E_j + \sum_{j=1}^{N} a_j dE_j + \sum_{j=1}^{N} da_j \cdot E_j \tag{7.17}$$

I will use U for average energy of the system, as in Schrodinger's book, and dU for the change in energy. We get

$$dU = E' - E = dW + dQ \tag{7.18}$$

where

$$\sum_{j=1}^{N} a_j dE_j = dW \quad \text{and} \quad \sum_{j=1}^{N} da_j \cdot E_j = dQ \tag{7.19}$$

Suppose the temperature change dT is 0 and that all of the dE_j are positive. We would expect that dU is in fact less than dW; if we do work to raise the energies of the states, the system will adjust in a way to dissipate some of that work. So that dQ should be negative, meaning that some heat will be transferred to surroundings from the system.

Here is a variation on the container full of molecules example. Suppose the system consists of some large number N of consumers, who fall into ℓ different groups. Suppose group number i has some enthusiasm level E_i for a particular product. The average enthusiasm for the product is then $E = U = \frac{1}{N}\sum_{i=1}^{N} E_{b_i}$ if $b_i \in \{1, \ldots, \ell\}$ is the group in which the i^{th} person falls. The thermodynamic hypothesis requires the equal likelihood of any two ways of assigning the N people to the ℓ possible groups which leads to the same average enthusiasm. This is surely not the case in general, particularly if there is a tendency in the population to have polarized views of the product. But let's suppose the thermodynamic hypothesis holds. Then we can do the experiment Schrodinger describes by doing the work (e.g. via advertising) of increasing the enthusiasm level E_j of members of the j^{th} group to some $E'_j = E_j + dE_j$. The temperature is an interesting invariant which under the thermodynamic hypothesis is related to how quickly the proportion of the population having extreme opinions drops. In general we would not expect that the work done by external forces to increase appreciation of the product will be entirely converted into an increase in the average enthusiasm for it. The difference is lost as "advertising heat."

Before going on to a more detailed mathematical account of Schrodinger's analysis of the first law, it's worth commenting on the above reasoning in the context of the other interpretations of statistical mechanics in §1. The second, Gibbs interpretation involves taking N identical copies of the system we care about. We imagine these are in contact with one another. Such copies are reminiscent of the Everett approach to quantum mechanics, in which the universe is constantly splitting into copies which can follow different paths. It seems more difficult to imagine with this interpretation what it means to do the work of changing the energies of the states in which the N copies can find themselves. One is really creating a new system of states, each of which is tied to one of the previous states. In the multi-universe interpretation, this might be seen as creating a slight alteration in the universe itself. So the first law has to do with what happens when we do work to change the universe. The universe goes along to some degree, but dissipates some of the work as heat.

8. Statistical analysis of the first law: Second approach

Using (5.10) let us write

$$F = \ln(Z) = \ln(\sum_{j=1}^{\ell} e^{-\mu E_j}) \tag{8.20}$$

This F is a function of μ together with $\{E_1, \ldots, E_j\}$. To be consistent with Schrodinger's book, we will use U for the average energy $E = \frac{1}{N}\sum_{j=1}^{N} a_j E_j$. The total differential of F is:

$$dF = \frac{\partial F}{\partial \mu} d\mu + \sum_{j=1}^{\ell} \frac{\partial F}{\partial E_j} dE_j \tag{8.21}$$

This total differential is an example of a differential form. One should think of a differential form as assigning a number to pairs which consist of a point $p = (\mu, E_1, \ldots, E_\ell)$ in $\mathbb{R}^{1+\ell}$ and vector v in the tangent space at this point in $\mathbb{R}^{1+\ell}$. This function should be linear in v when $(\mu, E_1, \ldots, E_\ell)$ is fixed. One can integrate such differentials over paths in $\mathbb{R}^{1+\ell}$ by breaking the paths into ever shorter segments. A differential form is exact if it has the form dG for some function G on $\mathbb{R}^{1+\ell}$, where the value of dG on a tangent vector v at some point p is the directional derivative of G at p in the direction of v.

From (5.12) and (5.11) and (8.21) we get

$$dF = -U d\mu - \frac{\mu}{N} \sum_{j=1}^{\ell} a_j dE_j \tag{8.22}$$

where remember we have $U = E$ relative to the earlier formulas. Therefore

$$d(F + U\mu) = \mu(dU - \frac{1}{N} \sum_{j=1}^{\ell} a_j dE_j) \tag{8.23}$$

The differential form dE_j is the one which produces the value δ_j when evaluated on the tangent vector $v = (0, 0, \ldots, \delta_j, \ldots, 0)$ at each point $p = (\mu, E_1, \ldots, E_\ell)$. Consider what happens when we integrate $d(F + U\mu)$ along a path in $\mathbb{R}^{1+\ell}$ space from the point with coordinates $(\mu, E_1, \ldots, E_\ell)$ to a nearby point with coordinates $(\mu', E_1', \ldots, E_\ell')$. When the displacement vector

$$(\delta\mu, \delta E_1, \ldots, \delta E_\ell) = (\mu', E_1', \ldots, E_\ell') - (\mu, E_1, \ldots, E_\ell) \tag{8.24}$$

is small, this integral is approximated to first order by evaluating the element of the tangent space at $(\mu, E_1, \ldots, E_\ell)$ given by (8.23) at the displacement vector in (8.24). The term

$$\sum_{j=1}^{\ell} a_j \cdot \delta E_j$$

which results is the work done by increasing the energy levels by δE_j. The value of dU on the displacement is a first order approximation to the change in energy of the system. In terms of the previous section, its value on the displacement is to first order given by

$$\sum_{j=1}^{\ell}(a'_j E'_j) - a_j E_j$$

when a'_j are the new occupation levels associated to $(\mu', E'_1, \ldots, E'_\ell)$. The significance of (8.23) is that it gives an expression for the differential of heat, dQ as

$$dQ = (dU - \frac{1}{N}\sum_{j=1}^{\ell} a_j dE_j) = \frac{1}{\mu} d(F + U\mu) \tag{8.25}$$

Notice that this differential is not in general the differential of a function! This is a very significant fact: it means that the integral of heat over paths is not just a function of the endpoints. However,

$$dS = k \cdot \mu \cdot dQ = \frac{dQ}{T} = k \cdot d(F + U\mu) \tag{8.26}$$

is the differential of a function, i.e. an exact differential, where k is as before the Boltzmann constant. This means that integrals of dS over paths will depend only on the endpoints of the path.

We can therefore define a new function S, the entropy of the system, by the formula

$$-\Psi = U - ST = -kT\ln(Z) \tag{8.27}$$

where $-\Psi$ is called the free energy of the system. If we hold T constant, the differential of free energy is just

$$dU - TdS = dU - dQ = dW.$$

This describes the work the system can do, or in other words, the part of the energy that is available for doing work.

It might be interesting to pursue the interpretation of entropy, energy and free energy in a psychological context. Suppose, for example, that one can reside in some number of energy levels. A given individual might have a characteristic average energy. If their likelihood of being in various energy levels satisfies the thermodynamic hypothesis, we could then assign them a temperature, which would be related via the statistics we have worked out to the likelihood of finding them in various energy states at a given time. The entropy of an individual then has to do with their free energy via (8.27). The free energy describes how much of their energy, at a given temperature, they can apply to do work on their environment, e.g. by writing or grading homework. Increasing either the temperature or the entropy of the individual decreases their ability to convert energy into work.

9. Some formulas for the entropy S.

The first formula connects the entropy S to the free energy $-\Psi$ and temperature T:

$$S = -\frac{\partial(-\Psi)}{\partial T} \quad \text{when} \quad T = \frac{1}{k\mu} \tag{9.28}$$

To check this, use

$$\begin{aligned}
-\frac{\partial(-\Psi)}{\partial T} &= \frac{\partial\mu}{T}\cdot\frac{\partial\Psi}{\partial\mu} \\
&= -\frac{1}{kT^2}\frac{\partial(\frac{1}{\mu}\ln(Z))}{\partial\mu} \\
&= -\frac{1}{kT^2}\left(-\mu^{-2}\ln(Z)+\frac{1}{\mu}\cdot\frac{\partial(\ln(Z))}{\partial\mu}\right) \\
&= k\ln(Z)+\frac{-1}{kT^2}\cdot\frac{1}{\mu}\cdot(-U) \qquad (9.29)\\
&= k\ln(Z)+\frac{U}{T} \\
&= \frac{\Psi+U}{T} \\
&= S \qquad (9.30)
\end{aligned}$$

Here (9.29) is from (5.12) because $U = E$, and (9.30) is from (8.27).

It's an interesting question to have a heuristic interpretation of (9.28). We have said that the free energy $-\Psi$ is the quantity of total energy available for doing work. So (9.28) says that if the entropy S is very large and positive, the free energy is decreasing rapidly with temperature when the other variables going into the free energy are held constant. Thus a person with large entropy will become less efficient more quickly as the their temperature increases. A person with entropy $S = 0$ would not see the amount of energy they can apply to do work change with temperature. In this sense they are more unflappable than a high entropy person.

Finally we connect the entropy defined by (8.27) to Shannon entropy. Recall that he proportion of the N systems in the Schrodinger model which are in state $j \in \{1,\dots,\ell\}$ was shown in (5.11) to be

$$p_j = \frac{a_j}{N} = \frac{e^{-\mu E_j}}{Z} = \frac{-1}{\mu}\frac{\partial}{\partial E_j}\ln(Z). \qquad (9.31)$$

Here

$$\sum_{j=1}^{\ell} p_j = 1 \quad \text{and} \quad \sum_{j=1}^{\ell} p_j E_j = E = U. \qquad (9.32)$$

The Shannon entropy of the set of probabilities $(p_1,\dots,p_N)$ is

$$H(p_1,\dots,p_N) = -\sum_{j=1}^{N} p_j \ln_2(p_j) \qquad (9.33)$$

So we calculate

$$\begin{aligned} H(p_1,\dots,p_N) &= -\sum_{j=1}^{N} p_j \ln_2(\frac{e^{-\mu E_j}}{Z}) \\ &= -\sum_{j=1}^{N} p_j(-\mu E_j - \ln(Z))/\ln(2) \\ &= \left(\mu\cdot\sum_{j=1}^{N} p_j E_j + \sum_{j=1}^{N} p_j \ln(Z)\right)/\ln(2) \\ &= (\mu\cdot U + \ln(Z)))/\ln(2) \qquad (9.34) \\ &= \left(\frac{U+kT\ln(Z)}{kT}\right)/\ln(2) \\ &= \frac{S}{k\ln(2)} \qquad (9.35) \end{aligned}$$

Here (9.34) is from (9.32) and (9.35) is from (8.27). So we conclude

$$S = k\ln(2)H(p_1,\dots,p_N)$$

This explains why the thermodynamic entropy S is the same, up to multiplication by a universal constant, as the information theoretic entropy $H(p_1,\dots,p_N)$ of the probabilities of the various states arising. Since we showed in the multinomial argument that the values of $p_1,\dots,p_N$ arose from the thermodynamic hypothesis, we can now say that these probabilities are the ones which maximize uncertainty for a given total average energy $E = U$.

In psychological terms, if the thermodynamic hypothesis holds, then one will have greater uncertainty about the state of a person with high entropy and a given total energy than one will have about the state of a person with the same energy level but lower entropy. It would be interesting to see if one could make this statement more quantitative.

10. Gibbs measures and the variational principle

There is another way to characterize the proportion

$$p_j = \frac{a_j}{N} = \frac{e^{-\mu E_j}}{Z} = \frac{-1}{\mu}\frac{\partial}{\partial E_j}\ln(Z) \qquad (10.36)$$

of N samples in the j^{th} state at thermal equilibrium. This leads to the concept of Gibbs measures which is central to many applications of thermodynamics.

We suppose as always that we are considering state vectors $b = (b_1,\dots,b_N)$ of length N. Each component b_i of b lies in one of the ℓ states $\{1,\dots,\ell\}$. The energy of the ℓ^{th} state is E_ℓ and the average energy is

$$E = E(b) = \frac{1}{N}\sum_{i=1}^{N} E_{b_i} = \frac{1}{N}\sum_{j=1}^{\ell} a_j(b)E_j$$

where $a(b) = (a_1(b),\dots,a_\ell(b))$ is the vector of state occupation number vectors associated to b. As before, if T is temperature we define $\mu = \frac{1}{kT}$ when k is Boltzmann's constant.

We now consider a function $\phi : \{1,\dots,\ell\} \to \mathbb{R}$ of the states. In our application $\phi(j)$ will turn out to be $e^{-\mu E_j}$, but we needn't make this assumption at first. Define the associated partition function by

$$Z(\phi) = \sum_{j=1}^{\ell}\phi(j). \qquad (10.37)$$

To a probability density function $f : \{1, \ldots, \ell\} \to [0, 1]$ we can define a mean energy

$$-\Phi(f) = -\sum_{j=1}^{\ell} f(j) \ln(\phi(j)) \tag{10.38}$$

The Shannon entropy of f is

$$H(f) = H(f(1), \ldots, f(\ell)) = -\sum_{j=1}^{\ell} f(j) \cdot \ln_2(f(j)) \tag{10.39}$$

Another way to formulate the existence of thermodynamic equilibrium probability distributions are the following variational results:

Theorem 10.1. *There unique probability density function $f : \{1, \ldots, \ell\} \to \mathbb{R}$ which maximizes*

$$\ln(2)H(f) + \Phi(f). \tag{10.40}$$

This f is the probability distribution $f_{therm} : \{1, \ldots, \ell\} \to \ell$ defined by

$$f_{therm}(j) = \frac{\phi(j)}{Z(\phi)} \tag{10.41}$$

Example 10.2. Suppose we fix a temperature $T = \frac{1}{k\mu}$ and that we define $\phi(j) = e^{-\mu E_j}$. The thermodynamic probability distribution is then

$$f_{therm}(j) = \frac{\phi(j)}{Z(\phi)} = \frac{e^{-\mu E_j}}{Z(\phi)} = \frac{a_j}{N}$$

where $(a_1, \ldots, a_N)$ is the approximation we derived earlier for the most likely set of occupation numbers for temperature $T = \frac{1}{k\mu}$. The statement that f_{therm} is the choice of f maximizing (10.43) for the given T and $E_1, \ldots, E_j$ is the same as saying that f_{therm} minimizes the Gibbs free energy

$$F(f) = -kT\Phi(f) - k\ln(2)H(f)T \tag{10.42}$$

Here

$$-kT\Phi(f) = -kT\sum_{j=1}^{\ell} f(j)\ln(\phi(j)) = \sum_{j=1}^{\ell} f(j)E_j = U(f)$$

is the average energy associated to the probability density function f. The function $k\ln(2)H(f)$ is the thermodynamic entropy $S(f)$ associated to this distribution. So the Gibbs free energy is

$$U(f) - S(f)T$$

in accordance with the definition of free energy in the previous section. Note that when $f = f_{therm}$, we derived in (9.34) the fact that

$$k\ln(2)H(f_{therm}) = S$$

when S is defined by (8.27).

Theorem 10.3. *Suppose one fixes a real number Φ_0 and that $f : \{1, \ldots, \ell\} \to \mathbb{R}$ is a probability density function which maximizes*

$$\ln(2)H(f) + \Phi(f) \tag{10.43}$$

over all f such that $\Phi(f) = -\Phi_0$. Then there is a number $T \in \mathbb{R} \cup \{\infty\}$ such that

$$f(j) = \frac{\phi(j)^{-1/kT}}{Z(\phi^{-1/kT})} \tag{10.44}$$

Remark 10.4. The general pattern behind Theorem 10.1 is that one starts with a phase space X which describes the possible states of a system. In Theorem 10.1, X is just $\{1, \ldots, \ell\}$. One then assigns a function $\phi : X \to \mathbb{R}$ and considers probability measures ν on X. To each such ν we define an entropy $H(\nu)$ which generalizes $k \ln(2) H(f)$ as well as a mean energy function $-\Phi(\nu)$ which generalizes $-k\Phi(f)$. The Gibbs measure ν_{therm} should then be the unique one maximizing

$$H(\nu) + \Phi(\nu). \tag{10.45}$$

This ν_{therm} should describe the thermal equilibrium distributions of states. It minimizes free energy when this is a positive multiple of the negative of (10.45). I'll describe how this works out in the next section the case of Hamiltonian mechanics.

Proof of Theorem 10.1. We consider variables $x_j = f(x_j)$ for $j = 1, \ldots, \ell$ satisfying the constraint

$$x_1 + \ldots + x_\ell = 1 \quad \text{and} \quad x_j \geq 0 \quad \text{for} \quad j = 1, \ldots, \ell. \tag{10.46}$$

By definition

$$F(x_1, \ldots, x_\ell) =_{def} \ln(2) H(f) + \Phi(f) = -\sum_{j=1}^{\ell} x_j (\ln(x_j) - \ln(\phi(j))) \tag{10.47}$$

Then

$$grad(F) = (\frac{\partial F}{\partial x_1}, \ldots, \frac{\partial F}{\partial x_\ell}) = (\ln(\phi(1)) - \ln(x_1) - 1, \ldots, \ln(\phi(\ell)) - \ln(x_\ell) - 1) \tag{10.48}$$

and

$$Jacobian(F) = \left(\frac{\partial^2 F}{\partial x_j \partial x_i} \right) = \text{diag}(\frac{-1}{x_1}, \ldots, \frac{-1}{x_\ell}) \tag{10.49}$$

By the usual Lagrange multiplier argument, the maximum of $F(x_1, \ldots, x_\ell)$ on $x = (x_1, \ldots, x_\ell)$ subject to the constraint (10.46) can only occur where

$$grad(F) = \tau \cdot grad(x_1 + \cdots + x_\ell) = \tau \cdot (1, \ldots, 1)$$

for some constant τ. So (10.48) then implies

$$\ln(\phi(j)) = \ln(x_j) + c$$

for some constant c. Exponentiating, we see $x_j = \phi(j) \cdot d$ for some constant d such that

$$1 = \sum_{j=1}^{\ell} x_j = \sum_{j=1}^{\ell} \phi(j) \cdot d = Z(\phi) \cdot d.$$

Thus $d = 1/Z(\phi)$ and the only x which can lead to a constrained maximum has $x_j = \frac{\phi(j)}{Z(\phi)}$.

To see that a unique constrained maximum does occur at this x, we first observe that F is a strictly concave function of x, in the sense that

$$F(\alpha y + (1 - \alpha) z) > \alpha F(y) + (1 - \alpha) F(z) \tag{10.50}$$

for all y and z where F is defined and all $0 \leq \alpha \leq 1$. To see this, just use the fact that F is the sum of the one variable strictly concave functions $x_j \to x_j(\ln(\phi(j)) - \ln(x_j))$. We now let y be $x = (\phi(1), \ldots, \phi(\ell))/Z(\phi)$. Suppose there is a $z \neq x$ satisying the contraints for which $F(z) \geq F(x)$. The fact that F is concave then implies that F must be increasing over some small interval starting from x in the direction of z. We now show this impossible because of the second order Taylor expansion of F at x. The directional derivative of F at x in the direction of $z - x$ vanishes by our choice of x and the Lagrange multiplier argument, while the second order term is given by (10.49) and is negative definite.

Proof of Theorem 10.3.

With the notation of the proof of Theorem 10.1, the condition that $\Phi(f) = -\Phi_0$ now imposes the additional constraint

$$-\Phi_0 = -\Phi(f) = -\sum_{j=1}^{\ell} x_j \ln(\phi(j)). \tag{10.51}$$

By the usual Lagrange multiplier argument, the $x = (x_1, \ldots, x_\ell)$ where $F(x_1, \ldots, x_\ell)$ can have a maximum over all x satisfying both (10.46) and (10.51) must have

$$\begin{aligned} grad(F) &= (\ln(\phi(1)) - \ln(x_1) - 1, \ldots, \ln(\phi(\ell)) - \ln(x_\ell) - 1) \\ &= \tau \cdot grad(\sum_{j=1}^{\ell} x_j) + \xi \cdot grad(-\Phi_0 + \sum_{j=1}^{\ell} x_j \ln(\phi(j))) \\ &= \tau \cdot (1, \ldots, 1) + \xi \cdot (\ln(\phi(1)), \ldots, \ln(\phi(\ell))) \end{aligned} \tag{10.52}$$

for some $\tau, \xi \in \mathbb{R}$. Hence

$$\ln(x_j) = (1 - \xi) \cdot \ln(\phi(j)) - (1 + \tau) \tag{10.53}$$

for $j = 1, \ldots, \ell$. We can rewrite this as

$$x_j = c \cdot \phi(j)^{-1/kT}$$

with $c = e^{-(1+\tau)}$ and $\frac{-1}{kT} = 1 - \xi$, with the understanding that if $\xi = 1$ then $T = \infty$. This leads to (10.44), and the rest of the argument is the same as the proof of Theorem 10.1.

11. Hamiltonian mechanics and Gibbs measures

We can now describe how to view N interacting molecules of identical mass m from the point of view of classical mechanics. Let $q = (q_1, \ldots, q_N)$ and $p = (p_1, \ldots, p_N)$ be the position and momenta of the particles, where each $q_i = (q_{i,x}, q_{i,y}, q_{i,z})$ and $p_i = (p_{i,x}, p_{i,y}, p_{i,z})$ lies in $\mathbb{R}^3$. The corresponding phase space X is a subset of $\mathbb{R}^{2 \cdot 3^N}$. The total energy of the particles is the sum of kinetic and potential energy terms:

$$\psi = \sum_{i=1}^{3N} \frac{1}{m} p_i \cdot p_i + W(q_1, \ldots, q_N) \tag{11.54}$$

The equations of motion are

$$\frac{dq_i}{dt} = (\frac{dq_{i,x}}{dt}, \frac{dq_{i,y}}{dt}, \frac{dq_{i,z}}{dt}) = (\frac{\partial \psi}{\partial p_{i,x}}, \frac{\partial \psi}{\partial p_{i,y}}, \frac{\partial \psi}{\partial p_{i,z}}) \tag{11.55}$$

and

$$\frac{dp_i}{dt} = (\frac{dp_{i,x}}{dt}, \frac{dp_{i,y}}{dt}, \frac{dp_{i,z}}{dt}) = -(\frac{\partial \psi}{\partial q_{i,x}}, \frac{\partial \psi}{\partial q_{i,y}}, \frac{\partial \psi}{\partial q_{i,z}}). \tag{11.56}$$

Here (11.55) is just the definition of momentum, while (11.56) is Newton's law

$$\text{Force} = \text{mass} \times \text{acceleration}.$$

The main goal of statistical mechanics is to determine the so-called equilibrium measure ν on the phase space X. This should have the property that the probability of finding the system in some measurable subset $A \subset X$ is $\nu(A)$.

Let $d\zeta$ be Lebesgue measure on X. This is the canonical measure on the Borel sets which gives cartesian boxes a measure equal to their usual Euclidean volume. We will consider measures of the form $\rho d\zeta$ in which $\rho = \rho(\zeta)$ is a (Lebesgue) integrable function. The associated thermodynamic entropy is

$$S(\rho) = -k \int_X \rho \cdot \ln(\rho) d\zeta. \tag{11.57}$$

The associated mean energy is

$$-\Phi(\rho) = \int_X \psi \cdot \rho d\zeta. \tag{11.58}$$

The variational principle in this case says that the equilibrium distribution ρ_{therm} for a given fixed value the mean energy $-\Phi(\rho)$ should be the choice of ρ having this mean energy which maximizes

$$S(\rho) + k\Phi(\rho) = -k \int_X \rho \cdot \ln(\rho) d\zeta + k \int_X \psi \cdot \rho d\zeta \tag{11.59}$$

Here ρ_{therm} is called the Gibbs measure.

We now use the following calculus of variations argument to determine ρ_{therm}. Consider changing ρ to $\rho + \delta$ for some function δ on X which is unformly small. Since ρ is a probability measure and we want $\rho + \delta$ to be one as well, this forces

$$\int_X \delta \; d\zeta = 0 \tag{11.60}$$

Since we also do not want to change the mean energy $-\Phi(\rho)$, we should have

$$\int_X \delta \; \psi \; d\zeta = 0 \tag{11.61}$$

There is also the constraint that $\rho + \delta$ is non-negative. We will ignore this, since we will be taking δ uniformly small in comparison to ρ.

Consider now how the right hand side of (11.59) changes when we replace ρ by $\rho + \delta$. For all real numbers $r > 0$ one has

$$(r + h) \cdot \ln(r + h) - r \cdot \ln(r) = (\ln(r) + 1) \cdot h + o(h)$$

where $\lim_{h \to 0^+} o(h)/h = 0$. So if δ is uniformly small in comparison to ρ, we should have

$$\int_X (\rho + \delta) \cdot \ln(\rho + \delta) d\zeta - \int_X \rho \cdot \ln(\rho) d\zeta = \int (\ln(\rho) + 1) \cdot \delta d\zeta + \quad o(\delta)$$

where $o(\delta)$ goes to 0 faster than δ. (To make this rigorous, one could use L^2 norms, but I will not do this.) Using this in (11.59), we see that replacing ρ by $\rho + \delta$ will change the right hand side of (11.59) by

$$k \int_X (-\ln(\rho) - 1 + \psi) \cdot \delta d\zeta + o(\delta). \tag{11.62}$$

If ρ produces a local maximum for the right side of (11.59), then (11.62) should be $o(\delta)$ whenever (11.60) and (11.61) hold. This forces there to exist constants α and β such that

$$-\ln(\rho) - 1 + \psi = \alpha\psi + \beta \tag{11.63}$$

This is equivalent to

$$\rho = e^{\beta - 1} \cdot e^{(1-\alpha)\psi}.$$

Since α and β are constants, we end up with the following result of Gibbs.

Theorem 11.1. *There is a constant T (the temperature) so that for all measurable subsets A of X, the measure $\nu_{therm} = \rho_{therm} d\zeta$ satisfies*

$$\nu_{therm}(A) = \frac{\int_A e^{-\frac{\psi}{kT}} d\zeta}{\int_X e^{-\frac{\psi}{kT}} d\zeta} \tag{11.64}$$

We can write

$$\rho = \frac{e^{-\frac{\psi}{kT}} d\zeta}{\int_X e^{-\frac{\psi}{kT}} d\zeta} \tag{11.65}$$

which is the natural generalization of the discrete probability formula of Theorem 10.3.

Note that the partition function is

$$Z = \int_X e^{\frac{-\psi}{kT}} d\zeta$$

and the free energy is

$$F = -kT \ln(Z)$$

just as in the discrete case.