Lesson 3: Maximizing entropy

Notes from Prof. Susskind video lectures publicly available on YouTube

Introduction

As usual we start with a brief review. The entropy of a discrete probability distribution is defined as

$$S = -\sum_i P(i) \ \log P(i) \tag{1}$$

Here we are in the case where the system under study is in a state belonging to a *discrete collection* of possible states. Our knowledge of which state the system is in is imperfect or incomplete. This incompleteness is modelled with a probability framework $[\ \mathcal{E},\ \Omega,\ P\ ]$ which we described at length in chapter 1. The probability that the system be in state i is $P(i)$.

Where the probability comes from, we have discussed a bit in chapter 1. We will discuss it again later in greater detail in the case of a system made out of large number of molecules forming a gas in thermal equilibrium. But for now we just assume that we know the distribution $P(i)$.

Attached to each state i is also an energy $E(i)$.

The assumptions, we should by now be familiar with, are

$$\begin{aligned} \sum_i P(i) \ &= \ 1 \\ \sum_i P(i) \ E(i) \ &= < E > \end{aligned} \tag{2}$$

The second equation is pretty much a definition: the average energy of the system is $< E >$, which we will soon simply denote E.

The way to read the two equations is straightforward. The sum of all the probabilities is equal to one. And if we add all the energies, each weighted by its probability, we obtain what is called an average quantity in statistics or probability theory[1].

From equation (1), it looks like the entropy is negative. But it is not. Because each probability $P(i)$ is less than one, its logarithm is negative. Therefore S is positive.

We saw last time that the average energy E can be viewed as the parameter indexing a family of probability distributions $P(i,\ E)$. For a given E the system has a given distribution of probability to be in the various states making up the set of possible states, which we called Ω.

Figure 1 shows three such distributions corresponding to three values of the parameter E.

[1] In this course we use the two terms statistics and probability theory synonymously. Specialists make the following distinction: probability theory is concerned with computing the probability distributions of various random variables. It is essentially a section of measure theory. Statistics on the other hand is concerned with the following problem: the distribution of a random variable X is known to belong to a family of distributions indexed by a parameter θ. From a series of observations x_1, x_2,... x_n of X what can we say about θ? θ is unknown but *it is not a random variable.* So it doesn't have a probability to be such and such value; it doesn't have a mean, etc. But it has a maximum likelihood. Statistics is also concerned with other problems of the same nature: testing, comparing, deciding, estimating things which are unknown but are not r.v. Estimators, however, *are* themselves r.v. with all sorts of interesting properties. That is the technical distinction between probability and statistics.

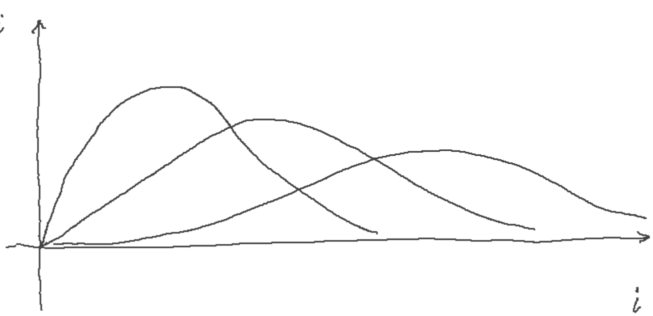

Figure 1: Family of distributions of probability. The states are arranged on the i-axis by increasing $E(i)$.

When the average energy gets larger, the probability distribution shifts to the right. The distribution gets also broader. And as a consequence its entropy gets larger too.

There is a particular probability distribution which is very special. It is the one corresponding to the minimum value of the energy E. It is of course the ground state of the system.

The ground state of a system is described by a probability which is zero everywhere except in the state of lowest energy. In figure 1, it is the left-most state. And the corresponding distribution is a peak of infinite height just at it[2].

[2] Our figure mixes up discrete and continuous distributions. Even though continuous distributions are handled via densities of probabilities and not discrete collections of probabilities, they are somehow intuitively simpler. A density which is peaking at the state of lowest energy is actually a Dirac distribution of integral value one – that is, the "area under its curve" is one, although it is just a peak. See volume 2 in the collection *The Theoretical Minimum* for a more detailed

The entropy of the ground state is exactly zero[3]. Then as the one parameter family of probability distribution shifts to the right, first of all the energy becomes larger, and the entropy increases too.

Remember that the entropy is qualitatively simply a measure of the logarithm of the number of states which have a significant probability underneath the probability distribution. When there are M occupied states with uniform probability, it is exactly true: $S = \log M$, as we saw in equation (23) of chapter 1.

When the blob of occupied states has a more general distribution of probabilities, the definition of S becomes equation (1) above. We checked in chapter 1 that it is consistent with the uniform case. But it is more complicated. It is now the average of the $\log P(i)$'s.

And when we are in the continuous case, equation (1) transforms into its continous analog $S = -\int P \log P$, see equation (34) of chapter 1.

In summary, as the probability distribution gets wider and

description of Dirac distributions. Or, if mathematically inclined, the reader can go to any manual on distributions, which extend the concept of functions by a straightforward passage to the limit.

[3] There is a slight abuse of language here. What is true is that when the system is in the ground state, it is perfectly known, therefore the entropy of the corresponding distribution is nil. Whereas when the parameter E increases the distribution of probability, representing our imperfect knowledge of the state of the system, also becomes broader. The reader should note that so far the family $P(i,\ E)$ is *a given*. We just describe it, without proving much.

wider, in the family of distributions that can describe the system, see figure 1, the average energy E, which is also the indexing parameter in the family, goes up. And at the same time the entropy goes up.

We will always assume that the entropy is a monotonically increasing function of the average energy. Of course one can invent probability distributions for which that is not true. It is possible to make up a family of probability distributions which get narrower and narrower as the energy increases. But, as we will see, the probability distributions that govern thermal equilibrium, at different levels of energy, are not like that.

We will find out that exactly figure 1 is the picture that governs the probability distributions for the states in thermal equilibrium.

Thermal equilibrium is characterized by a temperature T of course, but we can also characterize it by an average energy E.

Suppose we have a box of gas, and we know the average energy in it, and it is in thermal equilibrium. Thermal equilibrium means that whatever was going to happen has finished happening. Now it is simply quiescent and the molecules are all thoroughly mixed up, thoroughly dispersed throughout the system[4]. Not much is happening other than microscopically of course. It is perfectly tranquil, just every molecule moves at a few hundred meters per

[4] Notice that thermal equilibrium does not mean that the system is now in one state i. It is means that it has a *certain distribution* $P(i)$ over Ω which we will study in depth.

second :-) Then, in this state of thermal equilibrium, heat tends not to flow. We will understand that soon enough.

Heat tends not to flow because all parts of the system are in equilibrium. And, as we will see, the probability distributions $P(i,\ E)$ do in fact broaden as the energy E – or parameter E, if you like – goes up. As a consequence the entropy is a monotonically increasing function of energy. That is an important fact.

We are going to go now through some of the laws of thermodynamics. At first again it is a review.

Laws of thermodynamics

The *first law* is just energy conservation.

The *zeroth law* was stated when people were axiomatizing thermodynamics in early XX$^{\text{th}}$ century[5]. In their approach it had to come before almost anything else. But today we think of the zeroth law as a consequence of the first law and the second law in a sense. We will talk about the zeroth law after we talk about the second law.

The *second law* of thermodynamics says that entropy increases. Now what does that mean? It means that if we have a system, described by a given probability distribution, and it is not in equilibrium, and it evolves toward

[5] See for instance the work of Greek mathematician Constantin Carathéodory (1873 - 1950), *Examination of the foundations of thermodynamics*, published in German in 1909.

equilibrium, generally speaking the probability distribution broadens.

We are going to prove that eventually, in a qualitative kind of way. We will prove qualitatively the assertion that if we have a probability distribution and we allow it to evolve, being a little bit careless in our observation of it – that is called *coarse graining* –, we will find that the probability distribution always broadens. It should not be surprising. For a probability distribution to narrow, that would mean we found out more about the system than we originally knew.

Typically if we are considering a very complex system, with a large number of degrees of freedom, very difficult to keep track of, exactly the opposite tends to happen. We lose track of things. And because we lose track of things, the probability distribution tends to broaden. Hence the increase of entropy with time[6].

So we have the first law: energy is conserved. We have second law: entropy always increases with time. If it doesn't increase, it stays the same. But it never decreases. Let's assume that the entropy always increases or stays the same.

Now let's talk about the *zeroth law*. First of all the zeroth law asserts that there is such a thing as thermal equilibrium. If we take a box, for instance a box of gas, and we wait long enough, it will come to equilibrium.

[6] We are talking about a system which is not in equilibrium at the beginning of its study. Indeed for a system in equilibrium we saw that its distribution of probability is a monotonic function of its entropy. Therefore if it does not receive energy, its entropy cannot increase. More on this when we talk about adiabatic processes in later lessons.

We take the example of gas because it is easy to think intuitively about how all the molecules positions and velocities should evolve. But it doesn't have to be gas inside the box, it could be liquid, it could be solid, or any mix. In fact it could be any system which is isolated and self-contained. It will come to equilibrium.

What does that mean for it to come to equilibrium? Imagine that the system is made out of separate subsystems. Let's call them A and B. And let's put in a third subsystem C. Assume that they are all connected. By that we mean that they can exchange energy with each other. So we put some little pipes between them that allow energy to flow, figure 2.

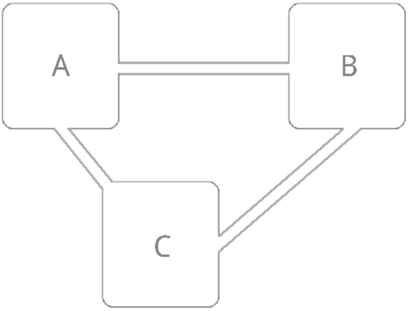

Figure 2: Subsystems exchanging energy through small interfaces.

The zeroth law also says that there is a concept called temperature attached to any equilibrium. Temperature has the characteristic that, if the system is made of several parts separately in equilibrium, and connected together through small interfaces as in figure 2, the energy always flows from hotter parts to colder parts. And if we wait long enough the flows will eventually come to an equilibrium, all the parts of the system having then one unique temperature.

Let's suppose there is only two systems A and B for a moment. Energy will flow from the one with higher temperature, say B, to the one with lower temperature, say A. Until what point? Could T_A become higher than T_B? No, that is not what happens. The flow of energy from B to A will go on until they equilibrate. And equilibrate – we will see – means the temperatures become equal. When the temperatures on both sides have become equal the heat stops flowing.

That is the meaning of temperature, if we like: it tells us which way heat flows.

From this sort of axiomatic description of the laws of thermodynamics, we have the following rule: two systems are in equilibrium with each other if and only if their temperatures are the same. As a consequence, if A is in equilibrium with B, and B is in equilibrium with C, then $T_A = T_B = T_C$. And therefore A is in equilibrium with C.

In other words, to be in equilibrium is an equivalence relation. We have simply replaced a qualitative notion of equilibrium by a quantitative one, through the use of temperature. Then the logical properties of equality apply to equilibrium[7].

[7] Because equality of temperature is an equivalence relation, equilibrium is then one too. Notice that there exist relations which are

We can summarize the zeroth law with these three rules:

1. There is a notion of temperature. We have already introduced it, but we will use it to illuminate this law, see next section.

2. Energy flows from higher temperature to lower temperature.

3. In thermal equilibrium, the temperature of all parts of the system is the same. If the temperature was not the same in all parts of the system, energy would flow until it became the same.

That is the basic zeroth law: the existence of a *temperature* function that tells us which way energy flows; and the existence of an *equilibrium* when the temperature in all parts of the system is the same. And isolated systems evolve toward equilibrium.

Let's see now if we can relate those ideas to the notion of temperature that we introduced in chapter 2, namely the rate of variation of energy with respect to entropy.

Energy, entropy and temperature

If we change the average energy of a system a little bit, the entropy will change. The rate of change is called the temperature.

reflexive and symmetric but not transitive, therefore not equivalence relations. An example is "to share some (unspecified) quality".

$$\frac{dE}{dS} = T \tag{3}$$

Let's see if we can figure out what this definition of temperature has to do with the idea of the flow of heat and the way it goes.

We begin with a system composed of two parts, A and B, and, as usual, a little pipe between them, figure 3.

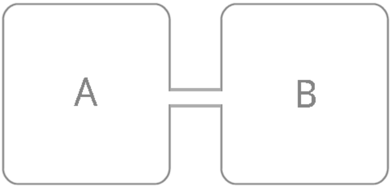

Figure 3: Isolated system made of two subsystems, to study the direction of flow of heat.

Let's write down all the basic equations that we know. For the moment we assume that the temperature of B is higher than the temperature of A. Of course we could have $T_B = T_A$, but let's start with the case where they are not equal.

$$T_A < T_B \tag{4}$$

That implies that the subsystems A and B are not in equilibrium with each other. Or said another way, the global system is not in equilibrium.

The first law tells us that when the system responds and does whatever it does over time, the total energy doesn't change. Since the whole system is not in equilibrium, there will be some energy shift from one subsystem to the other. Whichever way it happens, the following equation holds

$$dE_A + dE_B = 0 \tag{5}$$

For instance, we take our timepiece and we wait one second during which we watch the energy redistribute itself, figure 3. We find out that the energy of A changes; and the energy of B changes. But energy conservation tells us that the sum of the energies doesn't change.

Next, let's look at the change in the entropy of A plus the change in the entropy of B. That is the change of the total entropy in fact. It is greater than zero. In some very special cases it can be equal to zero, but the general statement is that it is greater than zero.

$$dS_A + dS_B > 0 \tag{6}$$

The probability distribution of the combined system shifts in such a way as to become broader. That is what equation (6) says.

Now let's use one more statement: that the change in the energy is equal to the temperature times the change in the entropy. And let's apply this to both A and B.

$$\begin{aligned} dE_A &= T_A \; dS_A \\ dE_B &= T_B \; dS_B \end{aligned} \tag{7}$$

Then equation (5), namely the first law of thermodynamics, rewrites as

$$T_A \; dS_A \;+\; T_B \; dS_B = 0 \tag{8}$$

Let's solve it for dS_B. We get

$$dS_B \;=\; -\frac{T_A}{T_B} \; dS_A \tag{9}$$

Now we take this formula and plug it in equation (6).

This is a general method when doing physics: you first think really hard about what you want to do, and then you blindly just go ahead with the equations and see where they lead.

So plugging equation (9) into equation (6) yields

$$dS_A - \frac{T_A}{T_B} \; dS_A > 0 \tag{10}$$

That is what the second law of thermodynamics says.

Let's multiply equation (10) by T_B. We will assume temperatures are positive, so that doesn't change the direction of the inequality. There are situations in which temperature can be negative, but not for us for the moment. So we get

$$T_B \; dS_A - T_A \; dS_A > 0$$

or

$$(T_B - T_A)\ dS_A > 0 \tag{11}$$

We have already assumed that the temperature of B is greater than the temperature of A. Then what does equation (11) say about dS_A? It says that

$$dS_A > 0 \tag{12}$$

The entropy of the subsystem A increased[8].

Let's multiply equation (12) by T_A. We get $T_A\ dS_A > 0$. But what is $T_A\ dS_A$? It is the change in energy of the subsystem A.

$$dE_A > 0 \tag{13}$$

So what have we found? By a little bit of manipulation and the use of a couple of laws we found that the change in the energy of A is positive. We used the first law of energy conservation, the second law of entropy increase, plus some simple algebraic operations.

It follows of course that the change in the energy of B is negative. In other words heat has flowed – or energy has flowed – from B to A.

So we have proved that, with the definition of temperature as given here – namely the rate of change of E with respect

[8] The same reasoning leads to $dS_B < 0$. But it is *not* because the total entropy is conserved. The total entropy is not conserved, it increases. The entropy of B decreases because of the presence of the cold heat sink A. And A being colder than B, it requires more ΔS for the same ΔE than B lost.

to S –, it indicates a direction of heat flow. Again when we say heat, so far we just mean energy. Energy will flow from B to A until the temperatures become equal.

When the temperatures become equal of course if T_A equals T_B then there is no flow. That is easy to see. And that is when equilibrium is established. So no flow of heat, all temperatures equal, and equilibrium are three equivalent situations.

So that is the zeroth law of thermodynamics.

Questions / answers session

Question: We assumed that the subsystems A and B and C were each separately in equilibrium. What happens if A, for instance, is not in equilibrium?

Answer: Yes, that is true: we assumed that each subsystem was in equilibrium with itself so to speak. But if one is not in equilibrium, you split it into subsystems which are in equilibrium and then apply the same analysis.

Q.: You described the collection of possible states as discrete. But if a state is the exact specification of the position and momentum of each molecule, doesn't it belong to a continuous space Ω?

A.: Yes, I did discretize the problem. But first of all in quantum mechanics we can take those variables to be discrete basically.

On the other hand, we could also use a continuous description of the space of states and the probabilities. As you know, we would then use densities of probability. The formula for the entropy of a distribution of probability would become $-\int P \log P$. But nothing would change essentially in all our reasonings and conclusions.

It is the general problem of the difference between a discrete description and a continuous description of a natural system that at first seems to be best described continuously. Most of the time, the two approaches are equivalent. And we choose the one that is most convenient mathematically.

The important point to remember is that a state always means the same thing. It means "as much as you could possibly know about the system if you were an infinitely powerful observer".

Now, there may be limits that have to do with the fundamental rules of physics, such as quantum mechanics which tells you you can't know as much as you thought you might like to know, such as the position and velocity of every particle. But still a state means as much as can be known about a system, by an infinitely powerful observer.

Then how we decide to represent and manipulate it, is a purely mathematical question. And we often have the choice of a continuous representation or a discrete representation.

Q.: But if we have say a finite number of molecules in the box, their positions and velocities are not discrete variables.

A.: In quantum mechanics they are. In quantum mechanics there is a discrete state counting for a finite box.

What we will have to show is that if we do the correct quantum mechanical statistical mechanics, then in certain limits it becomes equal to classical mechanics[9]. In other words in certain limits you can replace sums by integrals.

That is what it comes down to. And that is not a one liner. That is a subtle business. But it is true.

Let's now turn to the probability distribution $P(i)$'s after equilibrium has been established.

Distribution of probability at equilibrium

We consider our box of gas, or whatever system we are interesed in, after equilibrium has been established. There is some probability distribution $P(i)$ for the system to be in

[9] See volume 2 in the collection *The Theoretical Minimum*, on quantum mechanics, where we already showed that quantum mechanics in certain limits is equivalent to classical mechanics.

its different microstates i's. What is that probability distribution? What is the mathematics of it?

First of all, it is always a useful idea to imagine that our system is not closed, but is in contact with the outside world and can exchange energy with it. It is not only useful, but it usually corresponds to the facts.

In particular, it is a useful idea, if we want to think about the probability distribution of the states of our system, to think of it as being embedded in a huge reservoir, of a much larger number of degrees of freedom, that provides a *heat bath.*

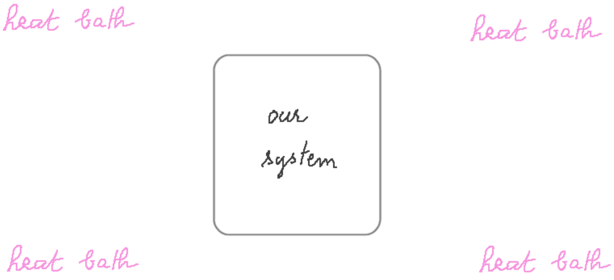

Figure 4: System plunged into a big heat bath.

The heat bath allows energy to flow back and forth between itself and the system. After a while, the system comes to thermal equilibrium with the big surrounding heat bath.

The heat bath is so big that even if a little bit of energy flows into the system the temperature doesn't change much.

This is something we could also prove. But let's just think about the physics: we have a huge system at a certain temperature – the heat bath –, and we have a little system plunged into it – our system. A little bit of heat flows from the big system to the little system. Typically the change in the temperature of the big system will be negligible.

So we can imagine that the big system is at some temperature and we simply wait until the small system comes to equilibrium with the big system and then is at that same temperature.

A particularly convenient choice of heat bath – we will see why –, which often actually corresponds to a fact, is just to imagine that the system in question is one of a very large number of identical systems, which are connected together by little pipes that allow heat to flow back and forth, figure 5.

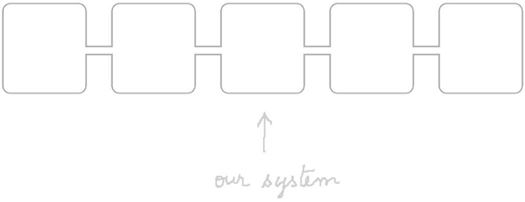

Figure 5: The heat bath is made of a large number of systems identical to the one under study. There are N identical systems.

Figure 5 is only suggestive. It doesn't represent the exact connection between subsystems. The idea is that the system under consideration – our system – is one of many identical systems, all connected together. And all but ours provide the heat bath.

Actually this sometimes corresponds more or less to facts. In particular if we have a very big system and we divide it into numerous small subsystems, the small subsystems may be very much alike and the big system minus one of them then is just the heat bath to that one.

This is a useful trick to pretend that the heat bath is just a repetition of the same system over and over again.

How many such systems? Eventually we will have a large number of them, large enough that we can think of the heat bath as very big. But let's call it capital N for the moment.

Each one of the N systems is in a state. The collection of possible states as usual is $\Omega = \{\ \omega_1,\ \omega_2,\ \omega_3,\ \text{etc.}\ \}$. To fix ideas, and prepare for some combinatorics, let's consider that there are $N = 9$ identical systems, and they each can be in one of three states $\omega_1,\ \omega_2,\ \omega_3$, figure 6.

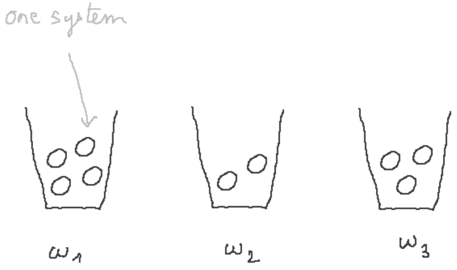

Figure 6: Nine systems distributed among three states.

The three states $\omega_1,\ \omega_2,\ \omega_3$ of Ω are not identical. They have different energy for example. Some of the N systems are in state ω_1. In figure 6, there are four of them. Generally speaking we will denote their number n_1. Similarly there are n_2 systems in state ω_2. In figure 6, $n_2 = 2$. There are n_3 systems in ω_3. In figure 6, $n_3 = 3$. And so forth.

The energy might increase from ω_1 to ω_3. In other words in figure 6 the states might be viewed as forming an axis of increasing energy. We represented three states, but there might be an infinite number of states (not to be confused with N).

For convenience I prefer now to represent the states of increasing energy on a vertical axis, leaving, for later on, the horizontal axis to time[10], figure 7.

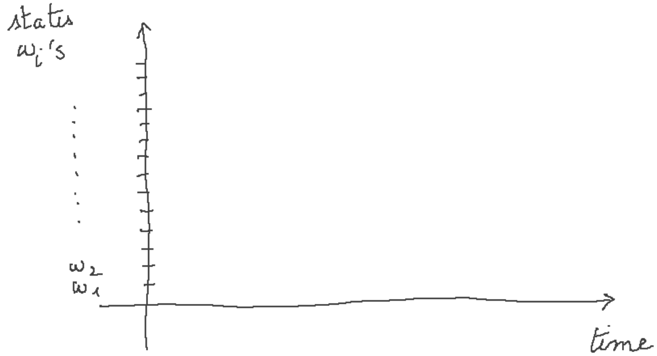

Figure 7: States of increasing energy.

In fact in general there will be an infinite number of possi-

[10] In relativity it is standard to represent time vertically, but here we are not doing relativity. And anyway we can use whatever representation we find convenient and like.

ble states ω_i's. But most of them will have so much energy that there is no chance at all that they be occupied. Only a negligible fraction of the N systems will occupy states of very high energy – if anything simply because we may not have that much energy available.

The counts n_i's are called the *occupation numbers.*

A first question is: given the occupation numbers, how many ways are there of redistributing the N systems among the states so as to create that set of occupation numbers?

Obviously that is going to be related to the $P(i)$'s. The more ways there are of distributing these N systems into a given set (n_1, n_2, n_3, ...) of occupation numbers, the more likely that set of occupation numbers is. Or equivalently, the more likely are the probabilities $P(i)$'s defined as $P(i) = n_i/N$.

That is a symmetry argument. If, in an imaginary random experiment distributing the systems into states, all possible redistributions are symmetric with respect to each other and therefore equally probable, then the most probable set of occupation numbers, is that which corresponds to the maximum number of ways of redistributing things.

Figuring out the most probable set of occupation numbers becomes a maximization problem where the unknowns are the n_i's. Without further conditions it would be straightforward to see that the solution is a uniform collection. But there are constraints.

First of all we can see it as a constraint that the following must naturally hold

$$\sum_i n_i = N \tag{14}$$

A second constraint has to do with energy. The total energy of the N systems is a fixed quantity. We are going to denote that fixed quantity N times E, because it is natural to think of it as proportional to the number of boxes, or systems. So let's write that constraint

$$n_1E_1 + n_2E_2 + n_3E_3 + ... = NE$$

or

$$\sum_i n_iE_i = NE \tag{15}$$

where E is the total energy *divided by* the number of boxes. So it is a fixed figure which is the average energy of a box[11].

Now suppose we have found the set $(n_1,\ n_2,\ n_3,\ ...\ )$ which corresponds to the maximum number of ways to distribute the N boxes into the states $\{\omega_1,\ \omega_2,\ \omega_3,\ ...\ \}$, under the constraints (14) and (15). Picking one box at random uniformly over the N boxes, what is the probability that it be in the i-th state? The answer is

$$P(i) = \frac{n_i}{N} \tag{16}$$

Remember that talking about probabilities only has a meaning if there is a clearly defined random experiment $\mathcal{E}$, a set

[11] It is the average, at a given time, over all the boxes. And it is also the average, over time, of a given box.

Ω which is the space of states of the world after having performed $\mathcal{E}$, and a probability distribution, or measure, P over the elements or subsets of Ω.

As we saw in chapter 1, when we are interested in only one random variable X, namely here the state in which a box is, we have a choice of definition Ω and P. We can choose Ω to be the collection of boxes, and P to be, for instance, uniform over them. Or we can choose Ω to be the set of possible values of X, here the states that the boxes can fall into, see figure 6, and modify P accordingly. That is what we did with equation (16).

Let's now use equation (16) to rewrite the two constraints (14) and (15). Simplifying our notation for $P(i)$ into P_i, they become

$$\begin{aligned} \sum_i P_i &= 1 \\ \sum_i P_i \, E_i &= E \end{aligned} \tag{17}$$

So these are now the new expressions for our two constraints. But if we prefer to think about them in terms of the occupation numbers, we can write them back as equations (14) and (15).

Before we worry about the total number of systems being kept constant, and the total energy being kept constant, let's just ask a more primitive question: for a given set of occupation numbers n_1 through n_p, or however infinitely many of them there are, how many ways are there of redistributing our N systems to correspond to that collection of occupation numbers?

That is a problem in combinatorics. The reader can find the solution in any good book on the subject. I will just give the answer, and then we will check it for a couple of cases.

The answer it not the same depending on whether the N boxes are distinguishable or not[12]. When they are distinguishable, the number of arrangements, denoted A, of the N boxes into $(n_1,\ n_2,\ n_3, ...)$ states, as in figure 6, is

$$A = \frac{N!}{n_1!\ n_2!\ n_3!\ ...} \tag{18}$$

For any positive integer m, the symbol $m!$ – read *m factorial* – is by definition the product of all the integers from 1 to m. Also by convention $0! = 1$. Then, thanks to this convention, whether the number of possible states $\{\omega_1,\ \omega_2,\ \omega_3,\ ...\ \}$ is finite or countably infinite doesn't make any difference, because when $n_i = 0$, in formula (18) it corresponds to dividing by 1.

So far, equation (18) gives the number of arrangements for a fixed set of occupation numbers, with no regard to whether the constraints are satisfies. We will come back to the question of the constraints being satisfied. But let's study the right-hand side of equation (18) on its own. It is an interesting expression. We are going to want to approximate it when N and the n_i's become large.

[12] For instance, if the boxes are $b_1,\ b_2,\ b_3$, and the states are ω_1, ω_2, if the boxes are distinguishable, then $(b_1,\ b_2)$ and (b_3) is not the same as $(b_1,\ b_3)$ and (b_2).

Before that, let's check it for a couple of cases. First of all, suppose $n_1 = N$ and all the other occupation numbers are 0. Then formula (18) gives $A = N!/n_1! = 1$. That seems reasonable. Indeed, how many ways are there of putting all the N boxes into one given state? One. So formula (18) is fine in this case.

Consider now the case where $N = 2$, that is there are two systems to distribute. And suppose there are two states. The case $n_1 = 2$ and $n_2 = 0$, we have already done. So let's look at the case $n_1 = 1$ and $n_2 = 1$. Calling the boxes, or systems, *objects*, and the states *cups*, as suggested in figure 6, how many ways are there to put two objects into two cups? If the objects are distinguishable, there are two ways. Formula (18) gives $A = 2!/(1!\ 1!) = 2$. So it is still ok.

Another case: $N = 3$, and $n_1 = n_2 = n_3 = 1$. There are six ways. And formula (18) gives that correct number.

One last case: $N = 4$, and $n_1 = 2$, and $n_2 = n_3 = 1$. There are $\binom{4}{2}$ = 4 x 3 / 2 = 6 ways to fill the first cup, multiplied by two remaining ways to fill the other two. This gives 12 ways. Turning to formula (18), we have $A = 4!/(2!\ 1!\ 1!) = 12$. Again it is hunky dory.

So the general formula for A is formula (18). It is not too hard to prove. It is high school or freshman level. It gives the number A of ways of distributing N objects into a collection of cups, with given occupation numbers n_1, n_2, n_3, etc. Although we haven't specified it with fancy notations, A is a function of the occupation numbers n_i's.

The interesting fact we are going to discover is that, in the circumstances that concern us, of a large number of systems distributed among a collection of states, with some constraints, the function A is highly peaked at a particular set of occupation numbers. Namely, when N gets very big, the occupation numbers cluster very strongly about a particular set of occupation numbers – or better yet a particular set of fractions n_i/N.

The distribution will get tighter and tighter as N gets up, so that the probabilities n_i/N will be well-defined.

In order to do that we ought to have some approximation method.

Approximation using Stirling formula

We want to approximate

$$A = \frac{N!}{n_1!\ n_2!\ n_3!\ ...}$$

The first thing we need to approximate is factorials. When we say approximate, we mean approximate under the circumstances that the n_i's are all very big.

We allow the number N of subsystems to get bigger and bigger in such a way that the occupation numbers also simultaneously get bigger and bigger. So all the n_i's are going to be assumed to be very big. And let's approximate the number A.

We will use the well-known Stirling approximation[13] for factorials.

$$\lim_{N \to +\infty} \frac{N!}{\sqrt{2\pi N}\ N^N\ e^{-N}} \ = \ 1 \tag{19}$$

In other words, as N gets large, $N!$ can be approximated by

$$N! \ \approx \ \sqrt{2\pi N}\ \frac{N^N}{e^N} \tag{20}$$

In order to prove this we are going to work with logarithms. $N!$ is the product of the integers from 1 to N. So we have

$$\log N! \ = \ \sum_{k=1}^{N} \log k \tag{21}$$

Now we are going to use the curve of the continuous function $\log x$ to approximate the right-hand side, figure 8.

The proof works with the logarithm calculated with respect to any base, 2, 10, e, or whatever, it doesn't matter, and we don't need to specify it. But we can think of Napierian logarithm if we like, that is the natural logarithm to the base e.

[13] Named after James Stirling (1692 - 1770), Scottish mathematician.

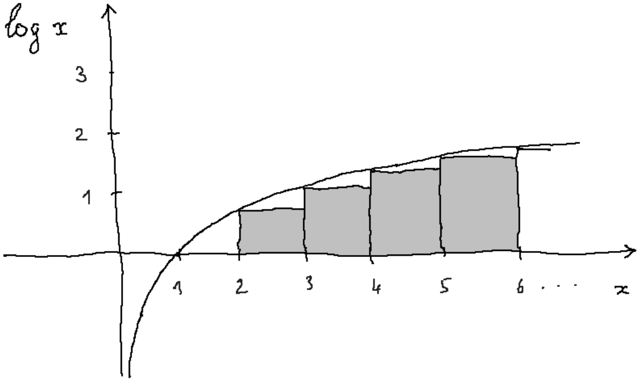

Figure 8: Curve $\log_e x$ and sum $\sum_{k=1}^{N} \log k$ (in grey).

We see that $\sum_{k=1}^{N} \log k$ is slightly smaller that $\int_1^{N+1} \log x \; dx$. But also, had we drawn the grey rectangles shifted one unit to the left, we would see that it is slightly bigger than $\int_1^{N} \log x \; dx$.

Thus we have

$$0 < \left[\sum_{k=1}^{N} \log k \; - \int_1^N \log x \; dx \right] < \int_N^{N+1} \log x \; dx \quad (22)$$

A primitive of $\log x$ is $x \log x - x$. Therefore

$$\begin{aligned} \int_1^N \log x \; dx &= x \log x - x \,\Big|_1^N \\ &= N \log N - N - 1 \log 1 + 1 \\ &= N \log N - N + 1 \end{aligned}$$

And the upper bound on the right of expression (22) is smaller than $\log(N+1)$. So the whole expression can be rewritten

$$0 \;<\; \left[\sum_{k=1}^{N} \log k - \;N\log N + N - 1\right] \;<\; \log(N+1) \tag{23}$$

If we now take the exponentials of the three terms, noting that $e^{N\log N}$ is the same as N^N, we get

$$1 < \frac{N!\; e^N}{N^N\; e} \;<\; N+1 \tag{24}$$

In other words,

$$N! \;\approx\; \frac{N^N}{e^N}\; C(N) \tag{25}$$

where $C(N)$ is a multiplicative factor which grows no faster than N. With further calculations, which are not necessary for our purposes, $C(N)$ can be refined into $\sqrt{2\pi N}$, giving the well-known approximation of equation (20). And there exist even better approximations if need be.

In the subsequent calculations, we will simply use

$$N! \;\approx\; \frac{N^N}{e^N} \tag{26}$$

where the $\approx$ sign here doesn't mean that the ratio of the left-hand side to the right-hand side goes to 1, as it usually means, but that it is bounded by N itself – and in fact the multiplicative factor that we omit is $\sqrt{2\pi N}$.

Now we can turn to approximating our monstrous combinatorial coefficient

$$A = \frac{N!}{n_1!\ n_2!\ n_3!\ ...}$$

Remember that it is the number of ways of rearranging N systems into into a collection of states $\{\omega_1,\ \omega_2,\ \omega_3, ...\}$, under the constraint that there be n_1 systems in the first state, n_2 systems into the second state, n_3 into the third, and so forth.

What we are going to do eventually is to maximize it. We shall look for the set of n_i's which makes this number of ways of rearranging as big as possible. And we will find a very sharp function. But we are not there yet. First we need to deal with the coefficient A.

We want to maximize A with respect to the n_i's, but subject to the two constraints that we wrote before, namely

$$\begin{aligned} \sum_i n_i &= N \\ \sum_i n_i E_i &= NE \end{aligned} \tag{27}$$

What is the set of occupation numbers which makes A as big as possible, under these two constraints? That is the mathematical problem we want to solve.

Once we have found the solution, then we effectively know all of the probabilities $P_i = n_i/N$.

How do we maximize a quantity with respect to a variable? We differentiate it. Here our quantity is a big product. That is unwieldy to work with. But we can equivalently work on maximizing its logarithm, because logarithm is a monotonic function. Since what we are interested in are the maximizing n_i's, it doesn't change the solution.

A second hurdle is that the n_i are not continuous variables but integer variables, so it looks like we cannot use calculus and are stuck. But we shall replace the n_i's by the P_i's. And since N is big, we can treat the P_i's as continuous variables.

But first of all, before going to logarithms, and to continuous variables, let's use Stirling approximation which we derived above in its simple form of equation (26). A can be rewritten

$$A \approx \frac{N^N e^{-N}}{\prod_i n_i^{n_i} e^{-n_i}} \tag{28}$$

where $\prod$ is the standard symbol for a product of terms.

Notice that in the denominator we can factorize separately all the e^{-n_i}. Taking into account that all the n_i's add up to N, this give a factor e^{-N} downstairs. It cancels with the one upstairs. So equation (28) simplifies into

$$A \approx \frac{N^N}{n_1^{n_1}\ n_2^{n_2}\ n_3^{n_3}\ ...} \tag{29}$$

Now we take the logarithms.

$$\log A \;\approx\; N \log N - \sum_i \; n_i \log n_i \tag{30}$$

Now we can start to smell something interesting. Remember that the n_i's, when divided by N, can be interpreted as probabilities. So equation (30) begins to look awfully interesting. The second term on the right-hand side is not quite $-\sum_i P_i \log P_i$. But it is somehow related to it, because the n_i's are related to the P_i's.

And what is $-\sum_i P_i \log P_i$? It is the entropy of the probability distribution P_i.

So let's complete the intermediate steps, and see that what we are doing, when we are looking for the maximum number of ways of rearranging objects into cups as in figure 6, is maximizing some entropy.

We are finding the probabilities which maximize the entropy, subject to the constraints that first of all they add up to one, and secondly that the total energy is some fixed value.

Let's plug into equation (30) the n_i's expressed as probabilities multiplied by N. We get

$$\begin{aligned}\log A \;&\approx\; N \log N - \sum_i \; NP_i \;\log NP_i \\ &\approx\; N \log N - \sum_i \; NP_i \;(\log N + \log P_i)\end{aligned}$$

There is a nice cancellation again. In the summation term, the sum of the $NP_i \log N$'s gives $N \log N$. So the whole thing can be rewritten

$$\log A \ \approx \ -\sum_i \ NP_i \ \log P_i$$

or

$$\log A \ \approx \ - \ N \ \sum_i P_i \ \log P_i \tag{31}$$

So we arrive at N times the entropy of any one of the systems in figure 5. Indeed, remember that the P_i's now are the probabilities for a given system to be in the state ω_i, or said more simply state i.

Thus, apart from the factor N which appears in equation (31) because there are N subsystems altogether, what we want to maximize is the entropy. Or said another way, if we want to find out the occupation numbers which maximize the number of ways that we can rearrange the N systems, keeping the occupation numbers fixed, it simply corresponds to maximizing the entropy.

Questions / answers session (2)

Q.: If we look at the formula $N!/n_1! \ n_2! \ n_3! \ ...n_p!$, that would be maximized when all the n_i's are equal. What is the meaning of that fact?

A.: It means that, if there was no constraint related to energy, the distribution of probability over all the possible states would be uniform. All the states would be equally probable.

But that mathematical result is of little interest, because it doesn't correspond to the physical situation we are examining.

We must add the constraint that $\sum_i n_i = N$. Secondly and most importantly we must add the constraint that the total energy of the big system formed by the N identical little systems in figure 5 is a fixed quantity. We denoted it NE, in order that E be the mean energy of any little system.

Then it is no longer true that the collection of occupation numbers, or equivalently, the probability distribution is uniform.

Q.: In using Stirling formula we assumed that N gets large. Do we have to justify that the n_i's get large too?

A.: As N gets large, the n_i's will get large in the same proportion. If you double the total number of systems, every average occupation number will also double.

I suppose we can justify it honestly as follows. We chop the big system off into a number of small systems, and say: look I know that I don't have more energy than a certain amount. So we can ignore states above that total energy anyway, see figure 7.

Then as N increases, the little n_i's can't but all increase, at least to maximize A under the two constraints.

Q.: Is that what you meant when you mentioned coarse graining?

A.: No. Coarse graining is an entirely different thing.

Coarse graining, so to speak, is what happens when you look at the world through somebody else's glasses.

Q.: We arrived at equation (30) after having taken the logarithm of something we wanted to maximize. But now in equation (31) we have a minus sign coming from something that was in the denominator. Shouldn't we want to minimize it?

A.: The minus sign does come from terms that were downstairs. But don't forget that the P_i's are less than 1, so their logarithms are negative. That is why the formula for the entropy has a minus sign in front of it. We do want to maximize the right-hand side of equation (31).

We want to maximize the entropy, with respect to the P_i's, but subject to constraints. It is certainly getting somewhat complicated. But if we want to understand where the probability distribution, over the possible states which any of the little systems can occupy, comes from, we have to do it.

If you don't want to see only the slick general public scientific magazine version of it, but you want to see how it really works, you have to get your hands dirty.

We see that we cannot use a simple symmetry argument on the states $\{\omega_1,\ \omega_2,\ \omega_3, ...\}$, which would lead us to some uniform distribution. The states are different. They have different levels of energy. As time passes, in any one box in figure 5, the agitation of the molecules lead to some micro-exchanges of energy with the neighboring boxes. The box goes through a collection of states, with energy E_i's. But they have an average energy E, which is simply the total energy of the big system divided by N.

This transforms into a constraint when looking for the probability distribution with maximum entropy. So our job then is to maximize $-\sum_i P_i \log P_i$ subject to constraints. Maximize with respect to what? Maximize with respect to the choice of probabilities P_i's.

If we go back, we are maximizing it with respect to the possible occupation numbers. But that translates into maximizing the entropy. Remember that entropy has a value which depends on the probability distribution.

Now we come to the conclusion that the most likely collection of occupation numbers corresponds to probabilities which maximize the entropy.

$$-\sum_i P_i \log P_i \tag{32}$$

But we want to maximize the expression subject to two constraints, namely

$$\begin{aligned}\sum_i P_i &= 1 \\ \sum_i P_i \; E_i &= E\end{aligned} \tag{33}$$

The average energy E of any subsystem is a fixed number, and so are the E_i's. They are givens of the problem. And of course the total probability must be equal to one. And we look then for the P_i's which maximize the entropy. That is the rule, and therefore the problem.

It is the second constraint in equations (33) which breaks the symmetry between the different states. Some microstates have a big energy E_i and some have small energy E_i.

So we have reduced now our objective to a mathematics problem: find the collection of probabilities P_i's which maximize expression (32) subject to the constraints (33).

To solve it we need one more mathematical tool: the method of Lagrange multipliers.

Lagrange multipliers come up over and over in statistical mechanics, and more generally in any problem where we need to maximize a function subject to constraints on the variables.

The final section of this chapter is devoted to the general method of Lagrange multipliers. Then in the next chapter we will use it in our problem of maximizing the entropy subject to the constraints we described.

Many thermodynamic quantities are Lagrange multipliers. We will discover, in particular, that *temperature is nothing but a Lagrange multiplier.*

Maximization under constraints, method of Lagrange multipliers

Consider a function F of several independent variables, x_1, x_2, ... , x_p. And F is nicely behaved, meaning that it is smooth – mathematicians would say continuously differentiable – with respect to the independent variables. Suppose we want to maximize F, that is to find the set of x_i's at which it is maximum.

The x_i's will eventually be the P_i's of the preceding problem. But for the moment let's be very general.

If there are no further conditions in the problem, we know how to solve it. We write the set of p equations

$$\begin{aligned} \frac{\partial F}{\partial x_1} &= 0 \\ \frac{\partial F}{\partial x_2} &= 0 \\ &\dots \\ \frac{\partial F}{\partial x_p} &= 0 \end{aligned} \tag{34}$$

We solve them simultaneously, and this gives us the solution for the p unknowns we were looking for[14].

But we want to add something to the problem. We want to add some constraints on what the x_i's can be. If we think geometrically about the space in which the x_i's are, there will be some constraints on *where* they can be.

Let's draw a picture of the space of the x_i's and of F, figure 9. For convenience, we are in two dimensions, $x_1 = x$ and $x_2 = y$. F would be along a third dimension which, instead of being represented in perspective, is suggested by its contour lines.

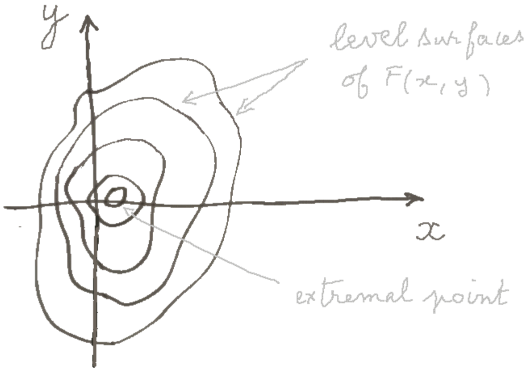

Figure 9: Contour lines of the function $F(x, y)$.

Everybody is familiar with contour map representations. They are used for instance in ordnance survey maps to represent the hills and valleys of a landscape. A line in figure

[14] More precisely, if F takes a maximum value someplace, the point $(x_1,\ x_2, ...\ x_p)$ at which it does satisfies equations (34).

9 corresponds to all the places where F has a given value. When the contour lines become smaller and smaller the final point is an extremal point, either a peak or a trough. When they are close to each other that corresponds to a steep slope in the terrain, etc.

Contour lines are generalized to more than two dimensions. They are then called level surfaces of the function F. If there were three independent variables x_1, x_2 and x_3, the level surfaces would indeed be ordinary surfaces.

In two dimensions, if somebody shows you the contour map of a function and you want to find the extremal points, positive or negative, just look for the places that are surrounded by somehow concentric closed loops zeroing to points.

Now let's add one constraint. In the preceding problem of maximizing entropy, we had two constraints. But for the general explanations of the method of Lagrange multipliers, let's start with one. Then we will add more.

We are now going to look for the place where $F(x_i)$ is maximum – x_i here stands for all the x_1, x_2, ... x_p – but given the constraint that some other function $G(x_i)$ must be equal to zero.

What if the constraint is that G must be equal to 3? Then subtract off the 3 from the original constraint function G, and call the new function G again. Whatever the constraint is, you can always say that it is that some function must be equal to zero.

For instance, the constraint $G(x, y) = 0$ might correspond to some curve in the contour map of F, figure 10.

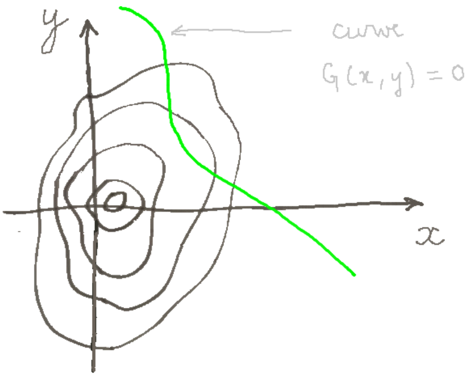

Figure 10: Contour map of $F(x,y)$ and constraint $G(x,y) = 0$.

Along the curve $G = 0$, we are going to look at the various values F takes. And we want to find the point where F is maximum *on the curve* $G = 0$.

The first thing to note is that the point P on the curve $G = 0$, where F is maximum under the constraint, is necessarily a point where the corresponding contour line of F is tangent to the curve $G = 0$, which is itself a contour line of G. Or equivalently, at P the gradient of F and the gradient of G are colinear. It is left as a geometric exercise for the reader to prove.

The technique is to observe that locally every line is straight, every surface is a plane, and all contour lines are evenly spaced. And that is true for F as well as for G, figure 11.

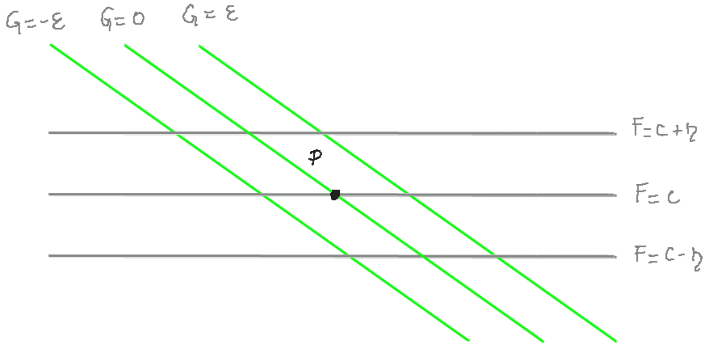

Figure 11: Local analysis, around P, of what would happen if the contour lines of G were not parallel to the contour lines of F. Show that in that case P would not be a maximum of F under the constraint $G = 0$.

The fact that at P the contour line of G must be parallel the contour line of F will be used in a moment.

So we want to find a method for solving the constrained maximization.

At the point P, solution to the problem, let's draw a line L perpendicular to the contour line of F, and therefore also to the curve $G = 0$, figure 12.

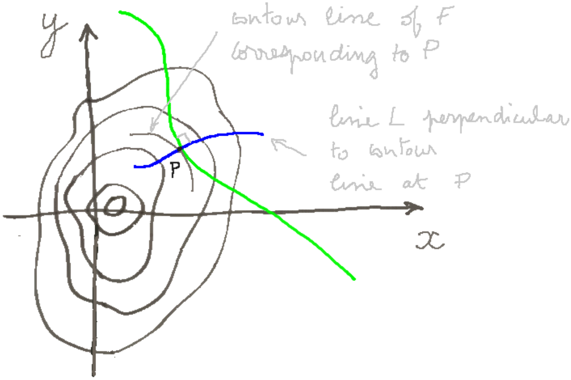

Figure 12: Point P solution to the maximization problem under constraint $G = 0$, and line L perpendicular to the contour line of function F at P.

The point P is the solution to the maximization problem under the constraint of being on the curve $G = 0$. But it is not necessarily the global maximum of F. It could be, but it would be a fluke. If we move along the line L away from P, we will cross other curves corresponding to G equal to different values, $G = -1$, $G = 0$ (at P), $G = +1$, etc.

Let's draw the value of F as function of G as we move along the line L, figure 13.

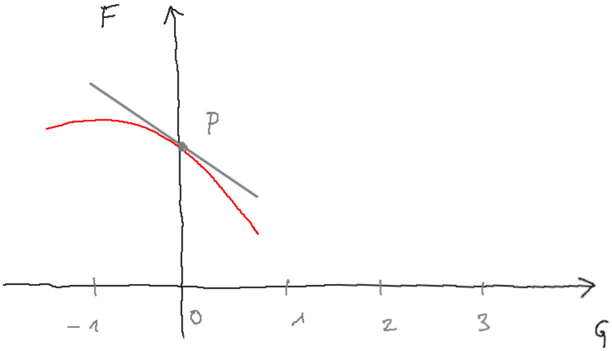

Figure 13: Value of F as a function of G *when we move on line* L perpendicular to the contour line of F and of G at P.

When $G = 0$ we are at the point P solution to the maximization of F under constraint, but the slope of F with respect to G at that point doesn't have to be zero. Why should it be? We were not looking for the place where F was globally stationary. We were looking for the point where F was stationary subject to the constraint of moving along the line $G = 0$.

What can we do in order that on the curve of F as a function of G, along line L, the point P be a stationary point, in other words with flat slope?

A solution is to add to F something proportional to G. Thus let's define the function F' as

$$F' = F + \lambda G \tag{35}$$

F' is a function of all the x_i's, as F and G. And it is also function of some number λ.

The number λ can be chosen to make F' along the line L flat at P – with respect to G or to ordinary distance, it doesn't matter. And in the perpendicular direction, that is along the contour line of F, which is locally the contour line of G as well, F' is also flat at P, since $G = 0$ and F is maximum there on that curve. So P must be a global stationary point of F'.

λ is a new unknown the value of which we will determine later. But it makes the function F' have a global stationary point, without constraint, at the same point P we were looking for. So now it is going to be easy to find P.

This is due to Lagrange[15]. It is called the method of Lagrange multipliers. And the coefficient λ is called the Lagrange multiplier. It is a devilishly clever trick.

Before we see some examples, let's clarify one point. We have added a new unknown λ to the problem. Don't we have now more unknows than equations? No we don't, because we also have the new equation $G = 0$.

That is the whole point. We have replaced a problem with p equations, p unknows, and one constraint, by a problem with $p + 1$ equations, $p + 1$ unknows and no constraint.

The proof that the method of Lagrange multipliers solves the problem of maximizing F under the constraint $G = 0$

[15] Joseph-Louis Lagrange (1736 - 1813), Italian mathematician. Lagrange was born in Turin, where he spent the first thirty years of its life. Then he spent twenty-one years in Berlin. Finally he went to Paris where he lived till the end of his life.

may seem intricate, but the use of it is very easy.

Let's turn to a couple examples of how to maximize or minimize a function under constraint. Of course maximize or minimize just depends on the sign. It is not an important point. It is the same technique.

Let's take

$$F(x,y) = \frac{x^2+y^2}{2} \tag{36}$$

I always like to put a factor of 2 downstairs, because we are going to differentiate F.

We want to minimize F. If there is no further constraint, the answer is easy. It is 0, and that minimum value is taken at $x = 0$ and $y = 0$.

But let's say we want to minimize F given $x + y = 1$.

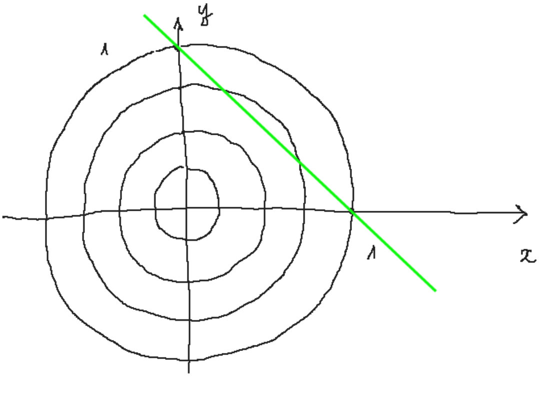

Figure 14: Contours of F, and constraint $G = 0$.

What is G?

$$G(x, y) = x + y - 1 \tag{37}$$

We are looking for the place, along the straight line $x+y = 1$, where $(x^2+y^2)/2$ is minimum, figure 14. Everybody can see by inspection where it is. It is at the point, on that line, which is the closest to the origin. It is of course $x = 1/2$ and $y = 1/2$. But let's find it using the method of Lagrange multipliers.

The rule is: add to F the quantity λ times G. That defines F'. Remember that F' here doesn't denote a derivative of F, but the new function we want to minimize.

$$\begin{aligned} F'(x, y, \lambda) &= F(x, y) + \lambda\ G(x, y) \\ &= \frac{x^2 + y^2}{2} + \lambda\ (x + y - 1) \end{aligned} \tag{38}$$

Next step, thinking temporarily of λ as a known coefficient, minimize F'. That is the rule.

So that means instead of writing $\partial F/\partial x = 0$ and $\partial F/\partial y = 0$, we write

$$\begin{aligned} \frac{\partial F'}{\partial x} &= 0 \\ \frac{\partial F'}{\partial y} &= 0 \end{aligned} \tag{39}$$

This gives

$$\begin{aligned} x + \lambda &= 0 \\ y + \lambda &= 0 \end{aligned} \tag{40}$$

Hence, $x = -\lambda$ and $y = -\lambda$.

But, remember, we also have the extra equation $G = 0$. By the way, it is nothing more, in the procedure of minimizing $F'(x, y, \lambda)$ without contraint, than the third equation $\partial F'/\partial\lambda = 0$. It yields

$$-\lambda - \lambda - 1 = 0 \tag{41}$$

or $\lambda = -1/2$. Therefore, as expected,

$$\begin{aligned} x &= \frac{1}{2} \\ y &= \frac{1}{2} \end{aligned} \tag{42}$$

That is the trick of Lagrange multipliers.

Exercise 1 : Minimize $x^2 + y^2$ under the constraint $2x + y = 1$.

Exercise 2 : Minimize $x^2 + y^2$ under the constraint $y = (x - 1)^2 + 1$.

What happens if we have more variables and more constraints?

Suppose we want to maximize, or minimize, a function F of three variables x_1, x_1 and x_3, and there are two constraints

$$\begin{aligned} G_1(x_1, x_2, x_3) &= 0 \\ G_2(x_1, x_2, x_3) &= 0 \end{aligned} \tag{43}$$

The way to proceed now is to define F' as F to which we add two terms, one for each constraint.

$$F' = F + \lambda_1 G_1 + \lambda_2 G_2 \tag{44}$$

So as usual F' depends on the three initial variables, x_1, x_2 and x_3, but also on the two new variables λ_1 and λ_2.

Then we minimize F' pretending that we knew what the lambdas were. So for each variable x_i we write

$$\frac{\partial F'}{\partial x_i} \tag{45}$$

But we still have the two unknowns λ_1 and λ_2. Indeed the solution to the set of three equations (45) will be a triplet of variables which depend on λ_1 and λ_2.

The two extra equations are again simply $G_1 = 0$ and $G_2 = 0$, which are nothing more than the partial derivatives of F' with respect to λ_1 and λ_2 set equal to zero.

Plugging the solution to (45) into the two extra equations gives a set of two equations for the two unknowns, λ_1 and λ_2. This enable us to find λ_1 and λ_2, and in turn x_1, x_2 and x_3 without anything unknown anymore.

This seems like a rather complicated things to do. But it is by far the easiest way to minimize something or maximize it when we have constraints.

Exercise 3 : Minimize $x^2 + y^2 - z^2$ under the two constraints $x + y = 1$ and $x + z = 1$.

Why are we interested in the method of Lagrange multipliers in this course? Because in many problems in statistical mechanics we want to maximize the entropy subject to some constraints.

The constraints might be that the total electric charge has some value, or the total energy has some value, or the total of something else has some value. Those are the kinds of constraints that we impose.

In particular, in the problem of figuring out the probability distribution of the states of the N small systems making up the big system of figure 5, we want to maximize the entropy subject to the two constraints that the total energy is fixed, and of course that the total probabilities add to one. Those are two constraints.

So let's now write the mathematics problem that we wrote it before. The problem is

$$\text{maximize } F(P_1,\ P_2,\ ....) \tag{46}$$

under the constraints

$$\begin{aligned} G_1(P_1,\ P_2,\ ....) &= \sum_i P_i - 1 = 0 \\ G_2(P_1,\ P_2,\ ....) &= \sum_i P_i E_i - E = 0 \end{aligned} \tag{47}$$

That will give us the probability distribution as a function of i, that corresponds to the most likely possible values for the occupation numbers. The occupation numbers tend to cluster around these values.

A note on statistics: What is the technical difference between a *probable value* and a *likely value*, or more generally between *probability* and *likelihood*? We saw that we talk about probabilities when there is a well defined random experiment $\mathcal{E}$ producing "states of the world" ω's belonging to a big set Ω, and there exist of measure of probability P on the subsets of Ω.

A random variable X is a function from Ω onto some set of things, numerical or not. X enables us to define subsets of Ω, as for instance $\{X < 2\}$ when the set in which X takes its values is numerical. And we can then talk about, calculate, or measure through a large number of replications of $\mathcal{E}$, the probability $P\{X < 2\}$.

Sometimes the problem we are confronted with is this: $\mathcal{E}$ and Ω are well defined and known, but we don't know P, and we don't have the possibility of reproducing $\mathcal{E}$ a large number of times. However we know that P belongs to a known family of probability distributions indexed by a parameter θ. And we have at our disposal some experimental data about a r.v. X. For instance, $\mathcal{E}$ was reproduced n times, and we obtained the measures x_1, x_2, ... x_n for X. The problem is: what is θ? Of course we don't have enough information to know θ exactly. But the experimental data we have do give us some information.

We can *estimate* θ. The value of the parameter θ which gives the maximum probability to the data we observed is called the maximum likelihood of θ. We don't talk of a "probable value" of θ because θ is not a r.v. But we talk about a "likely value" of θ.

The maximum likelihood of θ is a well known and well respected estimator. Let's denote it $\hat{\theta}$. It *is* a random variable produced by the experiment $\mathcal{F}$ which consists in n replications of $\mathcal{E}$. Indeed F produces x_1, x_2, ... x_n and therefore $\hat{\theta}$. If we reproduce $\mathcal{F}$, we will get a new set of data, and a new experimental value of $\hat{\theta}$.

The maximum likelihood estimator $\hat{\theta}$ is very natural estimator of θ. If we know that the following measures 2, 5, -3, 2 and 1 came from a Gaussian distribution of unknown mean θ and unknown width σ, we won't surmise that θ was 50. We will estimate it to be the mean of the five data values, namely 1.4, which happens to be the maximum likelihood of θ. Maximum likelihood estimators have plenty of nice properties, and a few weird ones[16].

[16] For the interested reader, see the James-Stein estimator, which is uniformly better than the maximum likelihood estimator according to some very natural measure of quality. Here is a reference in a general public scientific magazine:
`http://statweb.stanford.edu/~ckirby/brad/other/Article1977.pdf`

That is why when we speak of the estimated value of some parameters, be they probabilities like P_1, P_2, ... P_n in our problem of entropy, we talk not about their probable values, but about their *likely* values.

Back to maximizing the entropy: It is not easy to solve. But once we have set up the problem, the quantity to maximize, and the constraints, the rest is just pretty straightforward in its principle. It is then just probability theory and statistics.

It is the same set of rules which happens to govern many problems.

Once the problem is set, we are not doing the theory of gases; we are not doing the theory of superconductors; we are doing basic probabilistic theory of maximizing some entropy, the number of ways of rearranging things, subject to some constraints.

It is not special to any particular kind of physical system. It is special to the kinds of physical systems which have so many degrees of freedom that you are forced to deal with them statistically and probabilistically. But once you know that, the rules are always the same.

In the next chapter, we will solve the problem. F will be the entropy. As usual it will be called S. The problem will be to maximize

$$S(P_i) = -\sum_i \; P_i \; \log P_i \tag{48}$$

subject to the two constraints the reader should by now be familiar with.

The variables to shift around to maximize S are the P_i's. In other words, we have a probability distribution and we are going to vary the probability distribution until we find that probability distribution P_i's which maximizes the entropy, given that the sum of all the probabilities is 1, and the average energy is E.

That is statistical mechanics in a nutshell: solving that problem for different systems.