Matthew Schwartz

Statistical Mechanics, Spring 2019

Lecture 6: Entropy

1 Introduction

In this lecture, we discuss many ways to think about entropy. The most important and most famous property of entropy is that it never decreases

$$\Delta S_{\rm tot} \geqslant 0 \tag{1}$$

Here, $\Delta S_{\rm tot}$ means the change in entropy of a system plus the change in entropy of the surroundings. This is the second law of thermodynamics that we met in the previous lecture.

There's a great quote from Sir Arthur Eddington from 1927 summarizing the importance of the second law:

> If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations—then so much the worse for Maxwell's equations. If it is found to be contradicted by observation—well these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation.

Another possibly relevant quote, from the introduction to the statistical mechanics book by David Goodstein:

> Ludwig Boltzmann who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 1933. Now it is our turn to study statistical mechanics.

There are many ways to define entropy. All of them are equivalent, although it can be hard to see. In this lecture we will compare and contrast different definitions, building up intuition for how to think about entropy in different contexts.

The original definition of entropy, due to Clausius, was thermodynamic. As we saw in the last lecture, Clausius noted that entropy is a function of state, we can calculate the entropy difference between two states by connecting them however we like. If we find a reversible path to connect them, then the entropy change is determined simply by the heat absorbed:

$$\Delta S_{\rm system} = \int_{\rm rev.} \frac{dQ_{\rm in}}{T} \qquad (\text{Clausius entropy}) \tag{2}$$

This definition of entropy (change) called **Clausius entropy**. It is a thermodynamic, rather than statistical-mechanic, definition. It says that entropy is generated (or removed) from heating (or cooling) a system. Clausius entropy is important in that it directly connects to physics. As we saw in the last lecture, if the total Clausius entropy change were negative, it would be possible to create a system whose sole function is to turn heat into work.

In statistical mechanics, we can define the entropy as

$$S = k_B \ln \Omega \qquad \text{(Boltzmann entropy)} \tag{3}$$

where Ω is the number of microstates compatible with some macroscopic parameters (E, V, N). This form is usually attributed to Boltzmann, although it was Planck who wrote it down in this form for the first time. We'll call this the **Boltzmann entropy** since it's on Boltzmann's gravestone. Note that there is no arbitrariness in deciding which states to count in Ω or how to weight them – we count all states compatible with the macroscopic parmaters equally, with even weight. That is part of the definition of S. From S, we extract the temperature as $\frac{1}{T} = \frac{\partial S}{\partial E}$ and then integrating over $dE = dQ$ we recover Eq. (2).

Defining S in terms of microstates is useful in that it lets us compute S from a microscopic description of a system. For example, we saw that for a monatomic ideal gas, the Boltzmann entropy is given by

$$S = Nk_B \left[\ln V + \frac{3}{2} \ln\left(\frac{mE}{N}\right) + C \right] \tag{4}$$

for some constant C. This is an example showing that entropy is a state variable: it depends only on the current state of a system, not how it got there (heat and work are not state variables). Clausius formula assumed entropy was a state variable. With the Boltzmann formula, we can check.

Another way to compute entropy came from considering the number of ways N particles could be split into m groups of sizes n_i. This number is $\Omega = \frac{N!}{n_1! \cdots n_m!}$. Expanding for large n_i gives

$$S = -k_B N \sum_i f_i \ln f_i \tag{5}$$

where $f_i = \frac{n_i}{N}$. Since $\Sigma n_i = N$ then $\sum f_i = 1$ and so f_i has the interpretation of a probability: f_i are the probabilities of finding a particle picked at random in the group labeled i.

With the factor of k_B but without the N, the entropy written in terms of probabilities is called the **Gibbs entropy**:

$$S = -k_B \sum_i P_i \ln P_i \qquad \text{(Gibbs entropy)} \tag{6}$$

If all we do is maximize S at fixed N, the prefactor doesn't matter. However, sometimes we care about how S depends on N, in which case we need to get the prefactor right. We'll return to this in Section 3.

All of these ways of thinking about entropy are useful. They are all ways of understanding **entropy as disorder**: the more microstates there are, the less organized are the particles. A solid has lower entropy than a gas because the molecules are more ordered: the constraints on the positions of the atoms in the solid and limitations on their velocities drastically reduce the number of possible configurations. Entropy as disorder is certainly intuitive, and conforms with common usage: a child playing with blocks will most certainly leave them more disordered than how she found them, so we say she has increased the entropy. There are fewer ways for something to be ordered than disordered.

In the second half of the 20th century, it was realized that a more general and useful way of thinking about entropy than entropy as disorder is **entropy as uncertainty**. That is, we associate entropy with our ignorance of the system, or our **lack of information** about what microstate it's in. Entropy as uncertainty makes a lot of unintuitive aspects of the second law of thermodynamics easier to understand. We have already come across this the connection between entropy and information when discussing the principle of molecular chaos – the velocities of particles become correlated when they scatter, but over time the information of their correlations disperses over phase space and is lost. A solid has lower entropy than a gas because we have more information about the location of the atoms.

2 Free expansion

An example that helps elucidate the different definitions of entropy is the free expansion of a gas from a volume V_1 to a volume V_2.

First, consider the Boltzmann entropy, defined as $S = k_B \ln\Omega$ with Ω the number of accessible microstates. Using Eq. (4) which follows from the Boltzmann entropy definition, in going from a volume V_1 to a volume V_2, the gas gains an amount of entropy equal to $\Delta S = N k_B \ln\frac{V_2}{V_1}$. That the Boltzmann entropy increases makes sense because there are more accessible microstates in the larger volume, $\Omega_2 > \Omega_1$.

What about the Gibbs entropy? If P_i is the probability of finding the system in microstate i, then $P_i = \frac{1}{\Omega_1}$ when the gas is at a volume V_1. When it expands to V_2, each microstate of the gas in V_1 corresponds to exactly one microstate of the gas in V_2, so we should have $P_i = \frac{1}{\Omega_1}$ also at volume V_2, and therefore the Gibbs entropy is unchanged! Although there are more possible microstates in V_2, we know that, since the gas came from V_1, that only a small fraction of these could possibly be populated. That is, the state after expansion is in a subset $\mathcal{M}_{\rm sub} \subset \mathcal{M}$ of the full set $\mathcal{M}_2$ of microstates in V_2. The microstates in $\mathcal{M}_{\rm sub}$ are exactly those for which if we reverse time, they would go back to V_1. The size Ω_1 of the set $\mathcal{M}_1$ of microstates in V_1 is the same as the size $\Omega_{\rm sub}$ of $\mathcal{M}_{\rm sub}$.

So in the free expansion of a gas, Boltmann entropy increases but Gibbs entropy does not. How do we reconcile these two concepts?

The origin to this inconsistency is that Boltzmann entropy is defined in terms of the number of states Ω consistent with some macroscopic parameters, V, E, N, etc.. In contrast "the set of states that when time reversed to back to V_1" used for Gibbs entropy depends on more than just these parameters. So we are computing the different entropies using different criteria. If we define the probabilities P_i for the Gibbs entropy the same way as we define Ω for the Boltzmann entropy, that is, as the probability for finding a state with given values of V, E, N, the two definitions will agree. Indeed, the number of new states is $\frac{\Omega_1}{\Omega_2} = \exp(\Delta S) = \left(\frac{V_2}{V_1}\right)^{N k_B}$. Including these new states makes the Gibbs entropy go up by $\Delta S = N k_B \ln\frac{V_2}{V_1}$; removing them makes the Boltzmann entropy go down by the same amount. So if we are consistent with our criterion for computing entropy, the different definitions agree.

Now, you may be wondering why we choose to define entropy using V, E, N and not using the information about where the particles originated from. As an extreme example, we could even say that the gas starts in exactly one phase space point, $(\vec{q}_i, \vec{p}_i)$. So $\Omega = 1$ and the probability being in this state is $P = 1$ with $P = 0$ for other states. We can then evolve the gas forward in time using Newton's laws, which are deterministic and reversible, so that in the future there is still only one state $P = 1$ or $P = 0$. If we do so, $S = 0$ for all time. While we could choose to define entropy this way, it would clearly not be a useful concept. Entropy, and statistical mechanics, more broadly, is useful only if we coarse grain. Entropy is *defined* in terms of the number of states or probabilities compatible with macroscopic parameters. Coarse graining is part of the definition. Remember, in statistical mechanics, we are not in the business of predicting what will happen, but what is overwhelmingly likely to happen. Defining entropy in terms of coarse grained probabilities is the trick to making statistical mechanics a powerful tool for physics.

To justify why we must coarse grain from another perspective, recall the arguments from Lecture 3. Say we have a minimum experimental or theoretical phase space volume $\Delta q \Delta p$ that can be distinguished. Due to molecular chaos, the trajectory of a $\Delta q \Delta p$ phase space region quickly fragments into multiple disconnected regions in phase space that are smaller than $\Delta q \Delta p$ (since phase space volume is preserved under time evolution by Louiville's theorem). Then we coarse grain to increase the phase space volume of each disconnected region to a size $\Delta q \Delta p$. In this way, the phase-space volume of $\mathcal{M}_{\rm sub}$ grows with time. By ergodicity, every point in $\mathcal{M}_2$ will eventually get within $\Delta q \Delta p$ of a point in $\mathcal{M}_{\rm sub}$, so if we wait long enough, $\mathcal{M}_{\rm sub}$, through coarse graining, will agree with $\mathcal{M}_2$.

It may be helpful to see that if we use Gibbs entropy definition that entropy does in fact increase during diffusion. For simplicity consider a diffusing gas in 1 dimension, with number density $n(x,t)$. By the postulate of equal-a-priori probabilities, the probability of finding a gas molecule at x,t is proportional to the number density $P(x,t) \propto n(x,t)$. Then the Gibbs entropy is $S = -c\int dx\, n \ln n$ for some normalization constant c. Now, the diffusing gas satisfies the diffusion equation $\frac{\partial n}{\partial t} = D\frac{\partial^2 n}{\partial x^2}$. Using this we find

$$\frac{dS}{dt} = -c\int dx\, \frac{\partial}{\partial t}[n\ln n] = -c\int dx\,[1+\ln n]\frac{dn}{dt} = -cD\int dx\,[1+\ln n]\frac{\partial}{\partial x}\frac{\partial}{\partial x}n \tag{7}$$

Next, we integrate the first $\frac{\partial}{\partial x}$ by parts and drop the boundary terms at $x = \pm\infty$ by assuming that the density has finite support. This leads to

$$\frac{dS}{dt} = \underbrace{-cD(1+\ln n)\frac{\partial}{\partial x}n\Big]_{-\infty}^{\infty}}_{=0} + cD\int dx\,\frac{1}{n}\left(\frac{\partial n}{\partial x}\right)^2 > 0 \tag{8}$$

We conclude that during diffusion the entropy strictly grows with time. It stops growing when $\frac{\partial n}{\partial x} = 0$, i.e. when the density is constant, which is the state of maximum entropy.

This calculation illustrates that the problem from the beginning of this section was not that the Gibbs entropy wouldn't increase, but rather that imposing constraints on the probabilities based on inaccessible information about past configurations is inconsistent with the postulate of equal a priori probabilities.

3 Entropy of mixing

Although entropy is a theoretical construction – it cannot be directly measured – it is nevertheless extremely useful. In fact, entropy can be used to do work. One way to do so is though the entropy of mixing.

Say we have a volume V of helium with N molecules and another volume V of xenon also with N molecules (both monatomic gases) at the same temperature. If we let the gases mix, then each expands from V to $2V$. The energy of each is constant so the entropy change of each is (from integrating the discussion of free expansion before or directly from Eq. (4)):

$$\Delta S = Nk_B \ln\frac{2V}{V} = Nk_B \ln 2 \tag{9}$$

So we get a total entropy change of

$$\Delta S = 2Nk_B \ln 2 \tag{10}$$

This increase is called the **entropy of mixing**.

The entropy of mixing is a real thing, and can be used to do work. For example, say we had a vessel with xenon on one side and helium on the other, separated by a semi-permeable membrane that lets helium pass through and not xenon. Say the sides start at the same temperature and pressure. As the helium inevitably diffuses through the membrane, it dilutes the helium side lowering its pressure and adds to the pressure on the xenon side. The net effect is that there is pressure on the membrane. This pressure can be used to do work.

The pressure when there is a mixed solution (like salt water) in which the solute (e.g. salt) cannot penetrate a semi-permeable barrier (e.g. skin) is is called **osmotic pressure**. For example, when you eat a lot of salt, your veins increase salt concentration and pull more water in from your body to compensate, giving you high blood pressure.

In chemistry and biology, **concentration gradients** are very important. A concentration gradient means the concentration of some ion, like Na^+ or K^+ is not constant in space. When there is a concentration gradient, the system is not completely homogeneous, so entropy can be increased by entropy of mixing. Only when the concentrations are constant is there no way to increase the entropy more. Concentration gradients are critical for life. Neurons fire when the concentration of certain chemicals or ions passes a threshold. If systems always work to eliminate concentration gradients, by maximizing entropy, how do the concentration gradients develop? The answer is work! Cells use energy to produce concentration gradients. There are proteins in the cell wall that work to pump sodium and potassium (or other ions) from areas of low concentration to areas of high concentration.

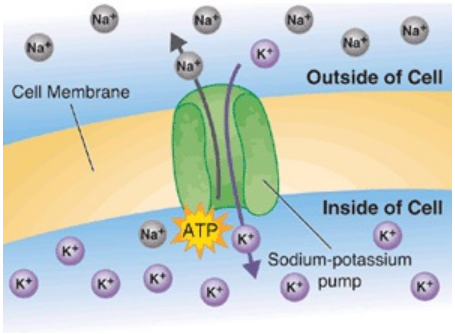


Figure 1. A cellular pump, present in practically every cell of every living thing. It uses energy in the form of ATP to establish concentration gradients

Another familiar effect due the entropy of mixing is how salt is able to melt ice. The salt draws water out of the ice (i.e. melts it) because saltwater has higher entropy than salt and ice separately.

In order to study these physical processes in quantitative detail, we need first to understand entropy better (this lecture), as well as free energy, chemical potentials, and phase transitions, which are topics for the next few lectures. We'll quantify osmotic pressure and temperature changes due to mixing (like in saltwater) in Lectures 8 and 9.

3.1 Gibbs paradox

So entropy of mixing is a real thing, and is very important in physics, chemistry and biology. But it is still a little puzzling. Say instead of two different ideal gases, we just have helium. Again we start with two volumes V, N and E each, and remove a partition between them to let them mix. The calculation is identical to the calculation for helium and xenon and we still find $\Delta S = 2Nk_B \ln 2$.

What happens if we put the partition back in. We start with helium gas with volume $2V$ number $2N$ and energy $2E$. Its initial entropy, by Eq. (4) is

$$S_{\text{init}} = 2Nk_B\left[\ \ln 2V + \frac{3}{2}\ln\left(\frac{mE}{N}\right) + c\right] \tag{11}$$

Now put a partition right down the middle of the gas, splitting it into two equal halves, each with V, N and E. Then the entropy of the sum of the entropies of the two halves:

$$S_{\text{final}} = 2\left\{Nk_B\left[\ \ln V + \frac{3}{2}\ln\left(\frac{mE}{N}\right) + c\right]\right\} = S_{\text{init}} - 2Nk_B \ln 2 \tag{12}$$

Thus the entropy has gone down, by exactly the entropy of mixing. If we mix the gases entropy goes, up, if we split them entropy goes down. That entropy could go down by simply sticking a partition in a gas seems very strange, and apparently violates the second law of thermodynamics. This is called the **Gibbs paradox**.

There are two parts to resolving the Gibbs paradox. First, we will argue that the states we had been counting were not states of identical molecules like helium, but assumed indistinguishability. So we'll have to correct for this. Second, we should understand why the counting that we have been using, which is indeed counting *something*, would have entropy apparently go down. The first question we address now, the second in Section 6.1.

3.2 Distinguishable microstates

To resolve the Gibbs paradox, let's think about why the entropy is changing from the viewpoint of Boltzmann entropy and ergodicity. Again, we start with xenon on the left and helium on the right each in a volume V. At a time t_0 we let them mix. The number of microstates for each gas increases by $\Delta S = Nk_B \ln \frac{2V}{V}$, so the net entropy of mixing is $\Delta S = 2Nk_B \ln 2$ (as we have seen). The entropy increases because there are now new microstates with xenon on the right or helium on the left that weren't there before. By ergodicity, we coarse grain and add these the microstates to Ω. The new microstates we include are not the ones actually populated by the expanding gas, but rather ones exponentially close to those microstates that we can't tell apart. If we imagine our system is truly isolated then the actual microstates populated trace back (in principle) to the separated states at time t_0. The new microstates we add when traced back in time are still mixed at t_0.

Now let's do the same thing for two volumes of helium that are allowed to mix starting at time t_0. As they mix, the entropy change is $\Delta S = Nk_B \ln 2$ just as for the xenon/helium mixture. This entropy increase comes because we are adding to the original microstates new microstates that, when traced back to t_0 have the molecules from the two original volumes still distributed throughout the whole system. But for the helium/helium mixing, these states *do* correspond to states we started with: half helium and the other half helium. So we already had these states and we are including them again and overcounting. This is clearly wrong, so we must undo it.

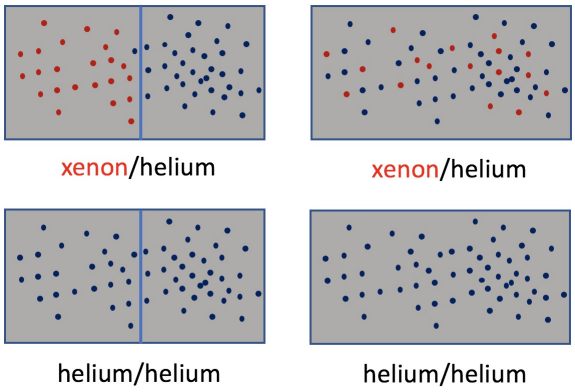


Figure 2. When xenon and helium mix, the new microstates we add don't look like the old ones. When helium mixes with helium, the new microstates are indistinguishable from the old ones.

In our original counting of states, we said that in a length L, with some minimal size Δq for discretizing the box, the number of states was $\frac{L}{\Delta q}$. Each particle could be anywhere, and for $\Delta q \ll L$ the chance of finding two particles in the same place is negligle. This lead to $\Omega = \left(\frac{L}{\Delta q}\right)^N$.

Let's take $L = 2\Delta q$ first, so there are two possible states. If there are N helium molecules, then there are $\Omega = 2^N$ configurations and $S = Nk_B \ln 2$. Now say we partition the system into two halves, with $\frac{N}{2}$ particles in each half. For each half, there is only one possible place for each particle, so the system is fixed, $\Omega = 1$ and $S = 0$. Thus entropy has decreased! What went wrong? The catch is that there were $\Omega_{\text{split}} = \binom{N}{N/2} = \frac{N!}{\frac{N}{2}!\frac{N}{2}!} \sim 2^N$ different ways to split the molecules, and we only considered one of them. If each molecule were different, say we had N different colored balls, then we could tell which of the states we ended up with. In that case, entropy of the system would into go down $\Delta S_{\text{sys}} < 0$, however it would take a lot of work to count all the balls and entropy of the surroundings would increase from doing all this counting. It's easier just to say we do not know which colored ball went where and include the Ω_{split} combinatoric factor. For helium, maybe we can tell them apart, or maybe not, but if we're not planning to try, then we need to account for the Ω_{split} possibilities. We can do this by saying that all the Ω_{split} partitionings of the particles are the same microstate. See section 3.4 for a detailed calculation of these cases.

It's actually easier to account for the combinatoric factor in a large volume than with $L = 2\Delta q$. In a large volume, there are many more positions than there are particles so we can assume no two particles are in the same state. In fact we have to make this assumption classically, since if there is a reasonable chance that two particles are in the same state we need to know if they are fermions or bosons and use the appropriate quantum statistics (Lecture 10). In a large volume with every particle in a different position, we can simply divide by the $N!$ for permuting those positions. This leads to

$$\Omega(q,p) = 2e^{\frac{3}{2}N}\frac{1}{N!}\left(\frac{V}{(\Delta q \Delta p)^3}\right)^N\left(\frac{4\pi m E}{3N}\right)^{\frac{3N}{2}} \tag{13}$$

This is the same as our old formula but has an extra $\frac{1}{N!}$ in front. The extra factor of $N!$ was introduced by Gibbs. After using Stirling's approximation the entropy $S = k_B \ln \Omega$ is

$$\boxed{S = Nk_B\left[\ln\frac{V}{N} + \frac{3}{2}\ln\left(\frac{4\pi m E}{3Nh^2}\right) + \frac{5}{2}\right]} \tag{14}$$

We have used $\Delta q \Delta p = h$ (as will be explained in Lecture 10). This is the **Sackur-Tetrode** equation.

Does the extra $N!$ solve Gibbs paradox? For the gas with $2V, 2N$ and $2E$, Eq. (11) becomes

$$S_{\text{init}} = 2Nk_B\left[\ln\frac{V}{N} + \frac{3}{2}\ln\left(\frac{mE}{N}\right) + c\right] \tag{15}$$

After splitting into two halves, Eq. (12) becomes

$$S_{\text{final}} = 2\left\{Nk_B\left[\ln\frac{V}{N} + \frac{3}{2}\ln\left(\frac{mE}{N}\right) + c\right]\right\} = S_{\text{init}} \tag{16}$$

So the entropy is unchanged by adding, or removing the partition.

What about the xenon/helium mixture? The gases are independent and do not interact, so each one separately acts just like helium alone. Thus inserting a partition in a helium/xenon mixture has a net effect of $\Delta S = 0$ on each separately and therefore $\Delta S = 0$ total as well.

What about the entropy of mixing? Let's start with two separate gases. Using our new formula, the initial entropy is the sum of the two gases' entropies. Each one has volume V, energy E and N. So,

$$S_{\text{init}} = 2\left\{Nk_B\left[\ln\frac{V}{N} + \frac{3}{2}\ln\left(\frac{mE}{N}\right) + c\right]\right\} \tag{17}$$

After letting the gasses mix, each gas goes from $V \to 2V$ but N and E are the same, so we have

$$S_{\text{final}} = 2\left\{Nk_B\left[\ln\frac{2V}{N} + \frac{3}{2}\ln\left(\frac{mE}{N}\right) + c\right]\right\} = S_{\text{init}} + 2Nk_B\ln 2 \tag{18}$$

So now, with the $N!$ factor added, we get a result that makes sense: there is only entropy of mixing if the two gases are different. Inserting a partition never changes the entropy.

Removing a factor of $N!$ from Ω also conveniently changes the general formula for Boltzmann entropy in Eq. (5) to the Gibbs one in Eq. (6):

$$S = -k_B\sum P_i \ln P_i \tag{19}$$

where $P_i = \frac{n_i}{N}$ are the number of particles with the properties of group i.

3.3 Entropy is extensive

Note that with the extra factor $N!$ in Eq. (13) entropy has become an **extensive quantity**. Extensive quantities are those that double when you have twice as much of the same thing. For example, energy E is extensive, as are N and V. The ratio of two extensive, quantities is an **intensive property**: one that characterizes the stuff itself, independent of how much you have. Temperature and pressure and the heat capacity C_V are intensive. We also use **extrinsic** as a synonym for **extensive** and **intrinsic** as a synonym for **intensive**.

To see that entropy is extensive, note from the Sackur-Tetrode formula that doubling V, N and E makes S double. This makes sense from the original definiton – if you have two isolated systems with Ω_1 microstates in one (entropy $S_1 = k_B \ln \Omega_1$) and Ω_2 microstates in the other (entropy $S_2 = k_B \ln \Omega_2$) then the total number of microstates is $\Omega = \Omega_1 \Omega_2$. So

$$S_{12} = S_1 + S_2 \tag{20}$$

This is true if the systems are truly isolated, whether or not we have the extra factor of N in the formula. It the systems are not isolated, Eq. (20) only works if we specify whether the particles are distinguishable – so that we know if we need to add new states (by coarse graining) – or if they are indistinguishable – so that coarse graining would not add anything new. Getting entropy to be extensive both for distinguishable and indistinguishable particles was what motivated Gibbs to add the $N!$ to Ω. The $N!$ is associated with indistinguishable particles.

We can also see the extensive property from the definition of Gibbs entropy in terms of probabilities, in Eq. (6). Say we have two systems A and B with probabilities P_i^A and P_j^B. Then the Gibbs entropy of the combined system is

$$S_{AB} = -k_B \sum_{i,j} P_i^A P_j^B \ln(P_i^A P_j^B) \tag{21}$$

$$= -k_B \left(\sum_i P_i^A \right) \sum_j P_j^B \ln(P_j^B) - \left(\sum_j P_j^B \right) \sum_i P_i^A \ln(P_i^A) \tag{22}$$

$$= S_A + S_B \tag{23}$$

So Gibbs entropy is an extensive quantity. If we had used the formula with the extra factor of N, Eq. (5), we would have found $S_{\rm AB} \neq S_A + S_B$.

To be clear, **indistinguishability** just means we can't think of any way to tell them apart. There is a quantum definition of indistinguishability for identical particles. We're not talking about that. We're just talking about whether we can think up a device to distinguish all of the $N \sim 10^{24}$ helium molecules from each other. Recall that to do work using the entropy of mixing, we need something like a semipermeable membrane that lets one type of thing through and no the other. Perhaps we could tell a He^4 isotope from a He^3 isotope. An entropy definition that can tell them apart would differ from one that says they are distinguishable by 2!, which hardly matters. You are never going to be able to pick out each of the 10^{24} individual He atoms in a gas, so the $N!$ is essential classically or quantum mechanically. In a metal, on the other hand, you can tell where each atom is, so the atoms in a solid should be treated as distinguishable, even if they are identical elements (like in solid gold). We'll talk about metals in Lecture 13.

It's worth adding that extensivity is a convenient property for entropy to have, but it is not guaranteed by any of the definitions of entropy. Indeed, in some systems, such as stars, entropy is not extensive due to long-range forces (gravity). With long-range interactions when you double the amount of stuff, the system can be qualitatively different (a star not much smaller than the sun would not be hot enough to burn hydrogen, see Lecture 15). With that caveat in mind, for the vast majority of systems we consider, where interactions are local (nearest neighbor or contact interactions), entropy will be extensive and we can exploit that property to simplify many formulae and calculations.

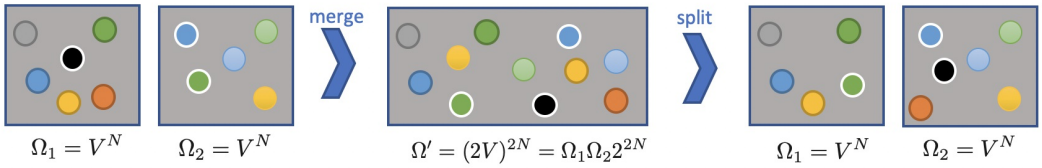
3.4 Mixing example details

Since distinguishability is an important and confusing topic, let's do an example of mixing and unmixing in full detail, with three different assumptions

1. N colored balls + N more colored balls, all distinguishable
2. N molecules helium + N more molecules helium, all indistinguishable
3. N molecules helium + N molecules xenon

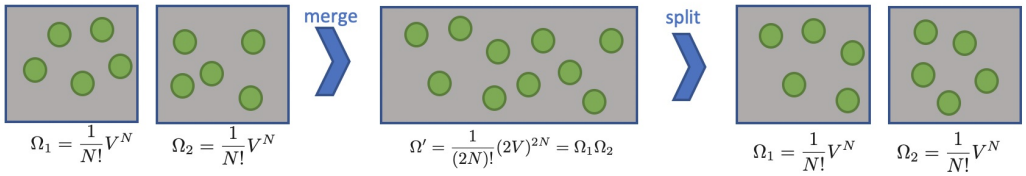
In all cases, we star with a volume V of each, then let them mix, then put a partition in to separate them.

For the first case we have something like this



So it looks like the entropy goes up by $\Delta S = 2\text{Nln}2$ when the two are mixed and then *down* by $\Delta S = -2N \text{ ln } 2$ when they are split. However, note that there are $\Omega_{\text{split}} = \binom{2N}{N} \approx 2^{2N}$ ways to split them. So in enumerating the final states, we should include this factor, writing $\Omega'' = \Omega_1\Omega_2\Omega_{\text{split}} = V^N V^N 2^{2N}$ so that $\Delta S = 0$ upon the split. If we actually look in the box and enumerate which balls are where, we lose Ω_{split}, but the entropy of the surroundings must go up due to the counting, as we explain in Section 6.1.

We can contrast this to the pure helium case when

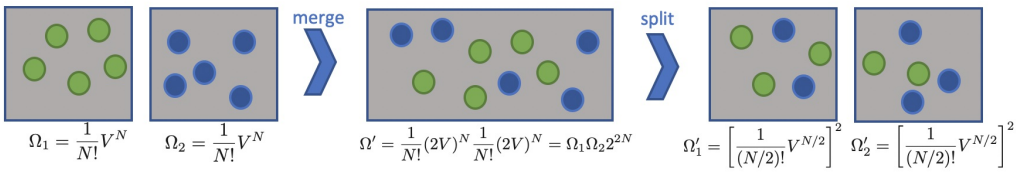


Here, we add the $\frac{1}{N!}$ factors for identical particles. When the two sets are merged, then the number of states is

$$\Omega' = \frac{1}{(2N)!}(2V)^{2N} = \Omega_1\Omega_2\frac{N!N!}{(2N)!}2^{2N} \approx \Omega_1\Omega_2\frac{N^N N^N}{(2N)^{2N}}2^{2N} = \Omega_1\Omega_2 \tag{24}$$

So $\Delta S = 0$ upon merging. Similarly, $\Omega'' = \Omega_1\Omega_2 = \Omega'$, so $\Delta S = 0$ upon splitting.[1]

When we mix helium and xenon we have



In this case, after mixing, each set of N molecules occupy the volume $2V$, so the entropy of mixing is $\Delta S = 2\text{Nln } 2$, just as in the colored balls case. When we split them, since the particles are identical, there is no way to tell apart one splitting from the other. Each half has $\frac{N}{2}$ of each species in a volume $\frac{V}{2}$. So the total number of states in this case is

$$\Omega'' = \left[\frac{1}{\left(\frac{N}{2}\right)!}V^{N/2}\right]^4 = \Omega_1\Omega_2\frac{N!N!}{\left(\frac{N}{2}!\right)^4} \approx \Omega_1\Omega_2\frac{N^N N^N}{\left(\frac{N}{2}\right)^{2N}} = \Omega_1\Omega_2 2^{2N} = \Omega' \tag{25}$$

And therefore $\Delta S = 0$ for the splitting case.

So we see that in the distinguishable case (colored balls) or the helium/xenon mixture case, there is a 2Nln2 entropy of mixing – each of the 2N molecules now has an extra binary choice of where to be, so we get 2Nln 2. In no situation does entropy go down when we split the volume back into two halves.

1. You might be bothered the fact that we had to take the thermodynamic limit $N \to \infty$, which lets us use that the most probable configuration is the only configuration, i.e. that there are always N particles in each side after the splitting. At finite N then indeed $\Omega' > \Omega_1\Omega_2$, so entropy goes up on merging. After splitting, we must allow for m particles on one side and $N - m$ on the other. Then $\Omega'' = \sum_{m=0}^{2N} \frac{1}{(2N-m)!}V^{(2N-m)}\frac{1}{m!}V^m = \frac{(2V)^{2N}}{2N!} = \Omega'$ exactly, so entropy still does not go down upon splitting.

4 Information entropy

Next, we introduce the concept of information entropy, as proposed by Claude Shannon in 1948. We'll start by discussing information entropy in the context of computation, as it was originally introduced, and then connect it back to physics once we understand what it is.

Consider the problem of data compression: we have a certain type of data and want to compress it into as small a file as possible. How good is a compression algorithm? Of course if we have two algorithms, say .jpg and .gif, and some data, say a picture of a cat, then can just compress the data with the algorithms and see which is smaller. But it may turn out that for one picture of a cat the jpeg comes out smaller, and for another, the gif is smaller. Then which is better? Is it possible to make an algorithm better than either? What is the absolute best an algorithm can do?

If you want *lossless compression*, so that the data can always be restored exactly from the compressed file, then it is impossible for any algorithm to compress all data. This follows from the "pigeonhole principle": you can't put m pigeons in n holes if $n < m$ without some hole having more than one pigeon.

So you can't compress every data file. But that's ok. Most data files have some structure. For example, images often have similar colors next to each other. This leads to the idea of run-length-encoding: instead of giving all the colors as separate bytes, encode the information in pairs of byes: the first byte in the pair gives the color and the second byte gives the number of pixels of that color. Run-length-encoding was used in early versions of .gif compression. It will compress almost all pictures. But if you give it white noise, where neighboring pixels are uncorrelated, then the compressed file will be bigger than the original.

Another feature in image data is that it often has smooth features separated by relatively sharp edges. Thus taking a discrete Fourier transform becomes efficient, since high frequency modes are often absent in images. jpeg compression is based on discrete Fourier transforms. Again, white noise images will not compress because their Fourier transforms are not simpler than the original images.

4.1 Text

For data that is text, the information is a sequence of letters. Different letters appear in a typical text with different frequencies. The standard way to write uncompressed text as numbers is with the ASCII code (American Standard Code for Information Interchange). In ASCII every letter is assigned a number from 0 to 127 which takes up 7 bits. For example, the ASCII code for "e" is 101 and the ASCII code for "&" is 38. There is an extended ASCII code as well, with 8 bits, allowing for letters such as "ä" which is 228. Since "e" is much more common than "&" it should be possible to efficiently compress text, allowing for random sequences of symbols not to compress well.

Here is a table of the probability of getting a given letter in some English text:

e	t	a	o	i	n	s	h	r	d	l	u	c
12.7	9.1	8.2	7.5	7.0	6.7	6.3	6.1	6.0	4.3	4.0	2.8	2.8
m	w	f	y	g	p	b	v	k	x	j	q	z
2.4	2.4	2.2	2.0	2.0	1.9	1.5	1.0	0.8	0.2	0.2	0.1	0.1

Figure 3. Probabilities of finding different letters in English text.

For example, if you open a book and pick a letter at random, 12.7% of the time the letter you pick will be "e" and only 0.1% of the time it will be "q". Exploiting these frequencies, a good compression algorithm would find a way to use fewer bits for "e" and more bits for "q".

What is the minimal number of bits you need to encode a given type of data? Shannon's answer was

$$H = \text{minimial } \#\text{bits} = -\sum_i P_i \log_2 P_i \tag{26}$$

This quantity H is called the **Shannon entropy** or **information entropy**. (You may recognize this H as the same as Boltzmann's H with $\log_e$ replaced by $\log_2$.) Shannon proved that you can't ever encode data with fewer than H bits, on average. This is known as the **source coding theorem**.

For example, with the letter sequences above we find

$$H = -[0.127 \log_2 0.127 + 0.091 \log_2 0.091 + \cdots + 0.001 \log_2 0.001] = 4.17 \tag{27}$$

This means the best we can possibly do is to represent each letter with 4.17 bits, on average. Having a non-integer number of bits it is not a problem; it just means that you could encode 100 characters with 417 bits and so on (see coin example below).

Note that 4.17 is better than the 7 bits in ASCII. Of course, we don't need 7 bits to represent 26 letters, but a naive encoding would use 5 bits (so $2^5 = 32$ characters), so since $H = 4.17 < 5$ this says you can do than better than 5 bits. Since we are considering non-integer bits, you might say we should use $\log_2 26 = 4.7$ bits. Indeed, the 4.7 bits is exactly what Shannon entropy would say is the best encoding if the probabilities of finding each letter were equal: $P_i = \frac{1}{26}$. That is

$$H_{\text{equal}} = -\sum_{i=1}^{26} \frac{1}{26} \log_2 \frac{1}{26} = \log_2 26 = 4.7 \tag{28}$$

That reason that we can use only 4.17 bits on average instead of 4.7 is because the probabilities are not equal. A better algorithm *uses this extra information*.

Shannon also noted that in text, letters aren't randomly distributed with probabilities but form words. Using words rather than letters, in his 1948 paper Shannon estimated that $H \approx 2.62$/letter for the entropy of English text. That is, it only takes 2.62 bits/letter to encode words. If a letter is given one byte (8 bits) in extended ASCII, this says that the maximal compression you could get is a compression factor of $\frac{8}{2.62} = 3.05$.

The Hutter prize is a 50,000€ competition to compress a 100MB snapshot of Wikipedia. The current record is 16 MB. For each 1% improvement you get 1,000€. Note that already the compression factor is $\frac{100}{16} = 6.25$ so that each character is represented by $\frac{8 \text{ bits}}{6.25} = 1.28$ bits. This is already much better than Shannon's original estimate. The improvement implies that Shannon's estimate of H is off, probably because he did not use all the information about the regularity of English text (for example, sentence structure); perhaps also Wikipedia articles are not typical text.

4.2 Algorithms

The source coding theorem doesn't tell you how to maximally compress the data, just what the maximal compression rate is. Let's think a little about what an optimal compression algorithm might look like. This should give us a better feel for Shannon entropy.

Frist, we'll take some limits. If all the probabilities are equal with $P_i = \frac{1}{N}$ then

$$H = -\sum_i \frac{1}{N} \log_2 \frac{1}{N} = \log_2 N \tag{29}$$

This is just the number of bits to encode N letters. I.e. if $N = 32$, it takes 5 bits, if $N = 64$ it takes 6 bits and so on.

If one of the probabilities is zero, then the effect on H is $\lim_{P \to 0} P \ln P = 0$. So adding something else to the data that never occurs does not affect the entropy. Conversely, if the data is completely uniform, so $P = 1$ for some character, then $H = -1 \log_2 1 = 0$. So it takes zero bits to encode a completely determined sequence.

Say we have a fair coin with 50% heads and 50% tails probabilities. Then

$$H = -\left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) = 1 \tag{30}$$

So it takes one bit to encode each flip. This is the best we can do. If the coin has 2 heads and no tails, then $H = 0$: we know it will always be heads.

Now let's look at a more complicated example. Say the coin is weighted so it is 90% likely to get heads and 10% likely to get tails. Then we find that

$$H = -(0.9 \log_2 0.9 + 0.1 \log_2 0.1) = 0.468 \tag{31}$$

So it is inefficient to encode each coin flip with just 0 or 1. What we really want is a code that uses less than a bit for heads and more than a bit for tails. How could we do this? One way is to say 0 means two heads in a row, 10 means heads and 11 means tails:

0	HH	00	HHHH
10	H	11	T

(32)

This lets us represent a sequence of two flips with one bit if it's HH, and with 4 bits otherwise. Thus for example, the sequence HHHHHHHHTHHH, with 12 digits, becomes 000011100, with 9 digits. For all possible 2 flip sequences, we find:

Sequence	HH	HT	TH	TT
Probability	81%	9%	9%	1%
Code	0	1011	1110	1111

(33)

The expected number of bits needed to encode a 2 flip sequence is the number of bits in the code (1 or 4) times the probabilities, namely $\#\text{bits} = 1\times 0.81 + 4\times 0.09 + 4\times 0.09 + 4\times 0.01 = 1.57$. So instead of 2 bits, we are using 1.57 on average, corresponding to an entropy per bit of $\frac{1.57}{2} = 0.785$. This is not as good as 0.468, but it is better than 1. In other words, our algorithm compresses the data, but not optimally. Can you think of a better compression algorithm?

4.3 Uniqueness

You might like to know that Shannon's formula for information entropy is not as arbitrary as it might seem. This formula is the unique function of the probabilities satisfying three criteria

1. It does not change if something with $P_i = 0$ is added.
2. It is maximized when P_i are all the same.
3. It is additive on uncorrelated probabilities.

This last criteria needs a little explanation. First, let's check it. Say we have two sets of probabilities P_i and Q_j for different things. For example, P_i could be the probability of having a certain color hair and Q_j the probability for wearing a certain size shoe. If these are uncorrelated, then the probability of measuring i and j is P_iQ_j. So the total entropy is

$$H_{\text{PQ}} = -\sum_{i,j} P_iQ_j\log_2(P_iQ_j) = -\sum_i \sum_j Q_jP_i\log_2(P_i) - \sum_j \sum_i P_iQ_j\log_2(Q_j) \tag{34}$$

Doing the sum over j in the first term or i in the second term gives 1 since Q_j and P_i are normalized probabilities. Thus

$$H_{\text{PQ}} = -\sum_i P_i\log_2(P_i) - \sum_j Q_j\log_2(Q_j) = H_P + H_Q \tag{35}$$

In other words, entropy is extensive. This is the same criterion Gibbs insisted on. So Gibbs entropy is also unique according to these criteria.

By the way, there are other measures of information entropy other than Shannon entropy, such as the collision entropy, Renyi entropy and Hartley entropy. These measures do not satisfy the conditions 1-3 above. Instead, they satisfy some other conditions. Consequently they have different applications and interpretations. When we discuss information entropy, we will mean only the Shannon entropy.

5 Information entropy to thermodynamic entropy

One way to connect the information picture to thermodynamics is to say that **entropy measures uncertainty**. For example, suppose you have gas in a box. In reality, all of the molecules have some velocities and positions (classically). If you knew all of these, there would be only one microstate compatible with it and the entropy would be zero. But the entropy of the gas is not zero. It is nonzero because we *don't know* the positions and velocities of the molecules, even though we could in principle. So entropy is not a property of the gas itself but of our knowledge of the gas.

In information theory, the Shannon entropy is 0 if the coins are always heads. That is because we know exactly what will come next – a head – so our uncertainty is zero. If the coin is fair and half the time gives heads and half the time tails, then $H = 1$: we are maximally ignorant. We know nothing about what happens next. If the coin is unfair, 90% chance of heads, then we have a pretty good sense of what will happen next, but are still a little uncertain. If we know the data ahead of time (or the sequence of flips), we can write a simple code to compress it: 1 = the data. So there is no ignorance in that case, as with knowing the position of the gas molecules.

With the uncertainty idea in mind, we can make more direct connections between information theory and thermodynamics.

5.1 Gibbs = Shannon entropy

The easiest way to connect information entropy to thermodynamic entropy is simply by interpreting microstates, and their associated probabilities, as the data. In general, suppose that the probability of finding a system in a given microstate is P_i. Then we can compute a thermodynamic entropy by multiplying the information entropy by a constant (recalling the relation $\frac{\ln x}{\ln 2} = \log_2 x$)

$$k_B(\ln 2)H = -k_B \ln 2 \sum_i P_i \log_2 P_i = -k_B \sum_i P_i \ln P_i = S \tag{36}$$

This is the Gibbs entropy from Eq. (6). Note that from the information theory point of view, the bits are necessarily indistinguishable (if a bit had a label, it would take more than one bit!), so it makes sense that the information entropy leads to Gibbs entropy.

What values of P_i will maximize S? Given no other information of constraints on P_i, the postulate of equal a priori probabilities (or the principle of maximum entropy) gives that the probability of a microstate i is $P_i = \frac{1}{\Omega}$ with Ω the total number of microstates. In terms of information theory, if the P_i are equal, it means that the data is totally random: there is equal probability of finding any symbol. Thus there should be no way to compress the data at all. The data can be compressed only if there is some more information in the probabilities. So the minimal information leads to the maximum entropy. With $P_i = \frac{1}{\Omega}$ the entropy is

$$S = -\sum_{j=1}^{\Omega} k_B \left[\frac{1}{\Omega} \ln \frac{1}{\Omega} \right] = k_B \ln \Omega \tag{37}$$

which is of course the original Boltzmann entropy formula in Eq. (3).

For another connection, consider the free expansion of a gas from a volume V to a volume $2V$. The change in entropy is $\Delta S = k_B N \ln 2$ or equivalently, $\Delta H = \frac{1}{k_B \ln 2} \Delta S = N$. So the number of bits we need to specify the system has gone up by N. But that's exactly what we should have expected: each bit says which of the 2V volumes each particle is in, so we need N more bits to specify the system.

5.2 Irreversibility of information storage

We made an abstract connection between Gibbs entropy and information entropy. Actually, the connection is not just formal, they are actually the same thing. To see this, we need a little bit of the physics of computation.

A key observation was made by Landauer in 1961. Landauer was interested in making powerful computers that used as little energy as possible. How little energy could they use? He considered a model of a bit with a (classical) double well. A ball on the left was 0 and a ball on the right was 1.

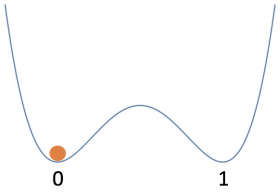


Figure 4. A model of information storage where a ball on the left is 0 and a ball on the right is 1.

The first question is whether we can change the bit from 0 to 1 without using any energy. It seems that the answer is yes. For example, we could hook a line to the ball and tie it to a counterweight ball rolling down a hill with the opposite potential energy so that no work is done:

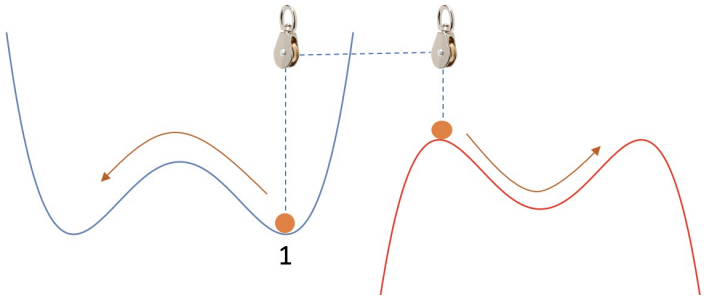


Figure 5. Hooking a pulley to the bit, we can move it from 1 to 0 without doing any work.

We just have to give the ball an infinitesimal nudge and it will roll across, then another infinitesimal nudge will stop it. So no energy is needed to flip this bit. This action is adiabatic and reversible, and can be also used to set 0 to 1 by running in reverse.

So if the bit is 0 it takes no energy to set it to 0 and if the bit is 1 it takes no energy to set it to 0. But what if we don't know what state the bit is in? Here's where it gets interesting. Suppose you had some automated mechanism for "SetToZero" to take the bit from 1 *or* 0 to 0. This is the kind of operation computers need to do all the time. Can we use our pulley gizmo to do it? The answer is no. In fact, the answer is no for any gizmo and any way of representing a bit. The reason as that we are just using Newton's laws, which are time-reversible. So if whatever action we do must be some some kind of 1-to-1 invertible function F acting on the position and momenta of the stuff in the bit. If phase space point $(\vec{q}_i^{\,0}, \vec{p}_i^{\,0})$ represents 0 and point $(\vec{q}_i^{\,1}, \vec{p}_i^{\,1})$ represents 1, then we want $F(\vec{q}_i^{\,0}, \vec{p}_i^{\,0}) = (\vec{q}_i^{\,1}, \vec{p}_i^{\,1})$ and $F(\vec{q}_i^{\,0}, \vec{p}_i^{\,0}) = (\vec{q}_i^{\,0}, \vec{p}_i^{\,0})$. But this is impossible if F is invertible. This argument is very rigorous and holds even in quantum mechanics, since the Schrödinger equation can also be run backwards in time.

Now, computers are very good at SetToZero. How do they do it? If we are allowed to dissipate energy, it's easy. For example, if there is friction in our double well system, then SetToZero could be "swing a mallet on the 1 side with enough energy to knock a ball over the hill." If there is no ball on the 1 side, the this does nothing $0 \to 0$. If there is a ball on the 1 side, it will go over to 0 and then settle down to the minimum due to friction, $1 \to 0$. Note that without friction this wouldn't work, since the ball would come back to 1. In a real computer, the information might be stored in the spin of a magnet on a magnetic tape. Applying a field to flip the bit would release energy if it flips which would then dissipate as heat. No matter how you cut it, we find

- **Landauer's principle**: erasing information requires energy be dissipated as heat.

Erasing information is an essential step in computation. Every time we store information, we erase the information that was previously there. But is the erasing, the throwing out of information, that dissipates heat, not the storing of information. That was Landauer's critical insight.

The key element to showing that SetToZero on an unknown bit is impossible without dissipation was reversibility of the laws of physics. Erasing information cannot be done with a reversible process. Thus thermodynamic entropy increases when information is thrown out.

To be absolutely clear, strictly speaking the information is not really lost. The laws of physics are still reversible, even with friction, so the final state could be run backwards to get the initial state. The final state however requires not just knowing the bit we are interested in, but all the positions and momenta of all the particles carrying off the heat. If we only record the bit, we are averaging over all possible states of the other stuff. It is in that averaging, that purposeful forgetting, where the information is actually lost. Dissipation into thermal energy implies this purposeful forgetting. Coarse graining erases information.

5.3 Energy from information

The connection between information entropy and thermodynamics was pushed further by Charles Bennett in the 1980s. Bennett was very intersted in how much energy computers require, in principle. That is, what are the fundamental limits on computing determined by thermodynamics?

The first relevant observation is that information itself can be used to do work. The setup Bennett considered was a digital tape where the information is stored in the position of gas molecules. We say the tape is made up of little cells with gas molecules either in the bottom of the cell, which we call 0, or in the top of the cell which we call 1. Imagine there is a thin dividing membrane between the top and bottom keeping the molecule from moving from top to bottom.

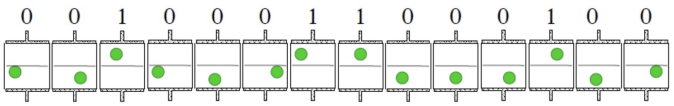


Figure 6. Information is stored in the position of gas molecules.

Let us keep this molecular tape in thermal contact with a heat bath so that the molecules are always at constant temperature. By fixing the temperature, we fix the molecule's momentum. Conversely, if we had allowed the temperature to vary, then the momentum would be variable too, and there would be more degrees of freedom that just the single bit represented by position.

Now, for an individual cell, if we know whether it's 0 or 1, we can use that information to do work. Bennett proposed putting pistons on both the top and bottom of each cell. Say the molecule is on the bottom. If so, we lower the top piston to isolate the molecule on the bottom half (like in Jezzball). This doesn't cost any energy. Then we remove the dividing membrane and let the thermal motion of the molecule will slowly push the piston up, drawing heat from the heat bath and doing useful work:

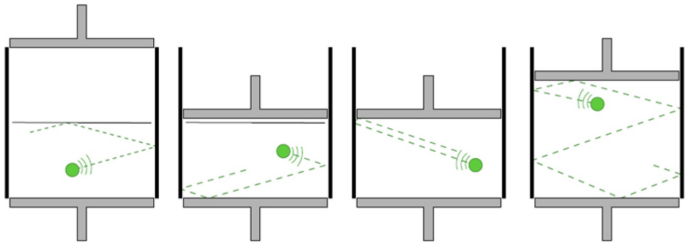


Figure 7. Extracting work from information.

If the molecule were in the top, we would move the bottom piston to the middle, remove the membrane, and let the system do work pushing it back down. Either way, the final state of the system has the gas molecule bouncing around the whole cell so we have lost the information – we don't know if it is 0 or 1 – but we have done work. This way of getting work out of information is known as **Szilard's engine**.

Once the piston is open and the molecule is free to bounce around the whole container, we have lost the information. We can then set the bit to 0 (or 1) by pushing down (up) on the gas with the piston. The work we do during this compression goes off into the thermal bath as heat. This is the SetToZero operation again that we discussed in Section 5.2. Just like there, acting on an unknown bit SetToZero dissipates energy. Once the bit is set, we can then use it to do work. But the work we get out is the same as the work we put in to set the bit. So we cannot do useful work if we do not know the state.

The entropy cost of losing the information, and of not being able to do work, is the entropy increase in doubling the volume available to the molecule

$$\Delta S_{\text{Gibbs}} = k_B \ln 2 \tag{38}$$

or equivalently

$$\Delta H = 1 \tag{39}$$

The information entropy goes up by one bit because we have lost one bit of information – the position of the molecule.

6 Maxwell's demon

We're now ready to tackle the most famous paradox about entropy, invented by Maxwell in 1867. Suppose we have a gas of helium and xenon, all mixed together in a box. Now say a little demon is sitting by a little shutter between the two sides of the box. When he sees a helium molecule come in from the right, he opens a little door and lets it go left. But when it's a xenon molecule coming in from the right, he doesn't open the door.

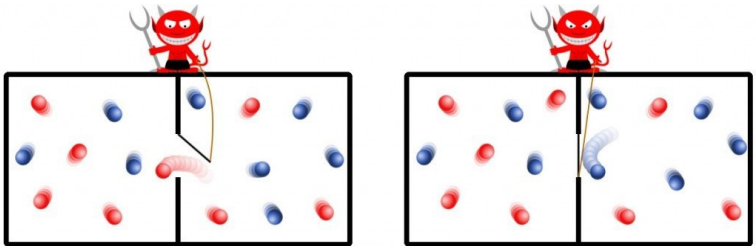


Figure 8. Fig 2: Maxwell's demon lets helium through, but not xenon

After a while, enough helium will be on the left that it will increase the pressure and this can be used to do work. The demon has made the system more ordered and the entropy has gone down.

Your first thought might be that the resolution to this puzzle has to do with identical particles. But this is not true. The paradox holds for pure helium. If the demon lets helium molecules go left and not right, the entropy would go down. Moving the door up and down doesn't take any work (it can be moved with an arbitrarily small push, and moreover the work can be completely recovered by stopping it), yet when the helium gets concentrated on one side it will exert a pressure on the barrier that can be used to do work. So this little demon is converting heat directly into work at constant temperature. In Maxwell's original formulation, the demon would only let the fastest molecules through one way and the slow ones the other way, so that the temperature difference of the two sides would increase. These are all different ways of saying the second law of thermodynamics is violated.

The demon doesn't have to be alive either. A robot could do his job, governed by the laws of physics. You just have to program him to tell xenon from helium or fast from slow. So consciousness doesn't have anything to do with it (although people sometimes like to say it does). For a mechanical example, say you had a little door with a very weak spring on it that only opens one way. If a molecule hits it from the left it will open and let the molecule through, but will not open when hit form the right (do you think this would really work?).

Maxwell's demon has exasperated generations of physicists, for over 100 years. In the 1920s the great physicists Szilard and Brillouin argued that it must take the robot some energy to find out which way the gas is going. The robot must shine light at least one photon of light on the molecule or something equivalent. The energy of this photon will then dissipate as heat increasing the entropy, so the total entropy of the demon/gas system would not go down. While it is true that doing the measurement with light does use energy and increase the entropy, it is possible to make a measurement using an arbitrarily small amount of energy, for example with an arbitrarily low frequency photon.

The correct resolution to Maxwell's demon is that somewhere in the process of letting the molecules pass through from left to right, the robot has to ask: is the molecule on the left? He must store the answer to this question in some variable in his program, somewhere. So he must set a bit, using the "SetToZero" operation. This operation takes work, at least as much work as we get out from moving the molecule to the right side. In terms of information, we gain one bit of information by identifying the particle as left/right, xenon/helium or fast/slow. But we also erase one bit of information by using SetToZero, sending heat into the surrounds. So the net effect is $\Delta S = 0$. Entropy does not go down and there is no contradiction with the second law.

You might instead suppose that we knew the bit on our tape was 0 to begin with. Then recording the position of the molecule with $0 \to 0$ or $0 \to 1$ does not require heat dissipation. In this case, it seems that Maxwell's demon does violate the second law. Note, however, that as we write the (random) locations of the molecules to our tape, our tape randomizes. So we are just moving the disorder from our gas into the disorder of the tape. In other words, $\Delta S_{\text{gas}} = -\Delta S_{\text{tape}}$ so the net entropy increase is still zero. Moreover. if we only have a finite sized tape then after a finite number of measurements we must start irreversibly erasing information. As soon as we do this, work is done, heat is dissipated, and the entropy of the surroundings increases.

This thought experiment also clarifies why information entropy really is entropy. If we had a finite size tape of 0's, then Maxwell's demon could indeed make a finite amount of heat flow from a hot bath to cold bath. As Bennett puts it, a tape of 0's has "fuel value" that can be used to do work. So we must include the specification of this tape as part of the definition of the system. If we do so, then the entropy never decreases at any step, it just moves from the disorder of the molecules in the box to the disorder of the information on the tape. Thus the entropy for which $\Delta S \geqslant 0$ is strictly true should include both thermodynamic and information theoretic entropy.

In this way Maxwell's demon was resolved by Bennett in 1982. after 115 years of confusion. As Feynman says in his Lectures on Computation (p. 150)

> This realization that it is the erasure of information, and not measurement, that is the source of entropy generation in the computational process, was a major breakthrough in the study of reversible computation.

6.1 Distinguishable particles

Having understood Maxwell's demon, we are now prepared to return to the Gibbs paradox for distinguishable particles. If we have a gas of entirely differently colored balls and partition it, as discussed in Section 3.4, the entropy does not go down because there are $\binom{2N}{N}$ ways of picking which half of the balls end up on the right and which half on the left. An objection to this is that once we partition the set, we can just look and see which half went where. If only one of the $\binom{2N}{N}$ choices is made, then the entropy would go down by $\Delta S_{\text{sys}} = -Nk_B \ln 2$, an apparent contradiction with the second law. Now that we understand information entropy, we know to ask *how* we knew which balls were on which side? To find out, we have to look at the color of each ball. Equivalently, we have to measure for each ball which side it's on. Each such measurement must SetToZero some bit in whatever we're using to do the measurement. The entropy consumed by this measurement is exactly the entropy lost by the system. This confirms that a system of distinguishable particles is perfectly consistent and does not lead to violations of the second law of thermodynamics.

7 Quantum mechanical entropy (optional)

This section requires some advanced appreciation of quantum mechanics. It's not a required part of the course, but some students might find this discussion interesting.

In quantum mechanics, distinguishability takes a more fundamental role, as does measurement. Thus, naturally, there are additional ways to quantify entropy in quantum mechanics. These all involve the density matrix ρ. Recall that in quantum mechanics, the states of the system are linear combinations of elements $|\psi\rangle$ of a Hilbert space. You may know exactly what state a system is in, in which case we say that the system is in a pure state $|\psi\rangle$. Alternatively, you may only know the probabilities P_i that the system is in the state $|\psi_i\rangle$. In such situations, we say that the system is in a mixed state (technically, a mixed ensemble of states). The density matrix is defined as

$$\rho = \sum P_j |\psi_j\rangle\langle\psi_j| \tag{40}$$

The **von Neumann entropy** is defined as

$$S = -k_B \text{Tr}[\rho \ln \rho] \tag{41}$$

Because S is defined from a trace, it is basis independent. Of course, we can always work in the basis for which ρ is diagonal, $\rho = \sum P_i |\psi_i\rangle\langle\psi_i|$, then $\rho \ln \rho$ is diagonal too and

$$S = -k_B \sum_j \langle\psi_j|\rho \ln \rho|\psi_j\rangle = -k_B \sum_j P_j \ln P_j \tag{42}$$

In agreement with the Gibbs entropy. Thus, in a pure state, where $P_j = 1$ for some j and $P_j = 0$ for everything else, $S = 0$. That is, in a pure state we have no ignorance. The von Neumann entropy therefore gives a basis-independent way of determining how pure a state is.

For example, say we start with a pure state $|\psi\rangle = |\rightarrow\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle + |\downarrow\rangle)$. This has $P_1 = 1$ and so $S = 0$. Now say we measure the spin along the z axis, but don't record the result. Then the system is either in the state $|\psi_1\rangle = |\uparrow\rangle$ with probability $P_1 = \frac{1}{2}$ or the state $|\psi\rangle = |\downarrow\rangle$ with $P_2 = \frac{1}{2}$. The density matrix is therefore

$$\rho = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \tag{43}$$

and the entropy is $S = k_B \ln 2$. The entropy has gone up since the measurement has collapsed the wavefunction from a pure state to a mixed state. We no longer know what the state is exactly, so our ignorance has gone up.

The von Neumann entropy also gives a useful way to quantify correlations. In quantum mechanics correlations among different particles are encoded through entanglement. For example, if there are two electrons, possible states have their spins aligned $|\psi\rangle = |\uparrow\uparrow\rangle$, anti-aligned $|\psi\rangle = |\uparrow\downarrow\rangle$, or entangled, $|\psi\rangle = |\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle$. To quantify entanglement in general, let us suppose our Hilbert space has two subspaces A and B, so $H_{AB} = H_A \otimes H_B$. Then we can compute a reduced density matrix for a subspace A by tracing over B, and for B by tracing over A

$$\rho_A = \mathrm{Tr}_B[\rho], \quad \rho_B = \mathrm{Tr}_A[\rho] \tag{44}$$

The von Neumann entropies of the reduced density matrices

$$S_A = -k_B \mathrm{Tr}[\rho_A \ln \rho_A], \quad S_B = -k_B \mathrm{Tr}[\rho_B \ln \rho_B] \tag{45}$$

are called the **entanglement entropies** of the subspaces.

For example, consider the system in pure state

$$\psi = \frac{1}{2}(|\uparrow\rangle_A + |\downarrow\rangle_A) \otimes (|\uparrow\rangle_B + |\downarrow\rangle_B) = \frac{1}{2}\Big[\ |\uparrow\uparrow\rangle + |\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle + |\downarrow\downarrow\rangle\Big] \tag{46}$$

Because the state is pure, the density matrix $\rho = |\psi\rangle\langle\psi|$ has zero von Neumann entropy. The density matrix for A is

$$\rho_A = \mathrm{Tr}_B(\rho) = \langle\uparrow|_B \rho |\uparrow\rangle_B + \langle\downarrow|_B \rho |\downarrow\rangle_B \tag{47}$$

$$= (|\uparrow\rangle_A + |\downarrow\rangle_A)(\langle\uparrow|_A + \langle\downarrow|_A) \tag{48}$$

$$= |\uparrow\rangle\langle\uparrow| + |\uparrow\rangle\langle\downarrow| + |\downarrow\rangle\langle\uparrow| + |\downarrow\rangle\langle\downarrow| \tag{49}$$

This is the density matrix for a pure state, $|\psi\rangle = |\uparrow\rangle_A + |\downarrow\rangle_A$, so $S_A = 0$. Similarly, $S_B = 0$. There is no entanglement entropy.

Now consider the system in an entangled pure state

$$\psi = \frac{1}{\sqrt{2}}\Big[\ |\uparrow\rangle_A \otimes |\downarrow\rangle_B + |\downarrow\rangle_A \otimes |\uparrow\rangle_B\Big] = \frac{1}{\sqrt{2}}\Big[\ |\uparrow\downarrow\rangle + |\downarrow\uparrow\rangle\ \Big] \tag{50}$$

Then, $S = 0$ since $\rho = |\psi\rangle\langle\psi|$ is still based on a pure state. Now the reduced density matrix is

$$\rho_A = \mathrm{Tr}_B(\rho) = \frac{1}{2}\Big[\ |\uparrow\rangle_A\langle\uparrow|_A + |\downarrow\rangle_A\langle\downarrow|_A\Big] = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \tag{51}$$

This is is now a mixed state $P_1 = \frac{1}{2}$ and $P_2 = \frac{1}{2}$. Thus $S_A = k_B \ln 2$. So the entangled state has entanglement entropy. Tracing over B amounts to throwing out any chance of measuring B. By doing so, we cannot exploit the entanglement anymore, so the information is lost and entropy goes up.

We can think of the whole universe as being described by a single wavefunction evolving in time. It's a pure state with entropy zero. Everything is entangled with everything else. As it becomes practically impossible to exploit that entanglement, exactly like it was impossible to exploit the correlations among scattered molecules classically, we coarse grain. Coarse graining in quantum mechanics means tracing over unmeasurable components. This increases the entropy and moreover turns a pure state into a mixed state. In this way, classical probabilities emerge from a completely deterministic quantum system.

In summary, von Neumann entropy lets us understand both the information loss by measurement and by losing entanglement. Entanglement is the quantum analog of correlations in a classical system. Discarding this information is the reason quantum systems become non-deterministic and entropy increases. We don't have to discard the information though. In fact, figuring out how to exploit the information stored in entanglement is critical to the function of quantum computers.

8 Black hole entropy (optional)

This section will be hard to follow if you don't know any general relativity. It's not a required part of the course, but some students might find this discussion interesting.

Using general relativity, you can prove some interesting results about black holes. General relativity is described by Einstein's equations, which are like a non-linear version of Maxwell's equations. Instead of $\partial_\mu F_{\mu\nu} = J_\nu$ we have

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = T_{\mu\nu} \tag{52}$$

The right-hand side of this equation, $T_{\mu\nu}$ is the energy-momentum tensor, which is the source for gravitational radiation like the current J_μ is the source for electromagnetic radiation. The object $R_{\mu\nu}$ is called the Ricci curvature, it is constructed by taking 2 derivatives on the metric $g_{\mu\nu}$ in various combinations: $R_{\mu\nu} = \partial_\mu \partial_\alpha g_{\alpha\beta} + \partial_\mu g_{\nu\alpha} \partial_\alpha g_{\alpha\gamma} + \cdots$. So $g_{\mu\nu}$ plays the role that the vector potential A_μ plays in Maxwell's equations where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$.

If we set $J_\mu = 0$, there is a spherically-symmetric static solution to Maxwell's equations, $A_0 = \frac{e}{4\pi\epsilon_0 r}$ and $\vec{A} = 0$. This is the Coulomb potential. It has one free parameter e. The Coulomb potential is singular at $r = 0$ indicating that there is some charge localized there. In fact, this solution corresponds to a current which is zero everywhere but the origin: $\vec{J} = 0$ and $J_0 = e\delta(\vec{x})$. In Newtonian gravity, the spherically-symmetric static solution is the Newtonian potential $\Phi = -G\frac{M}{r}$. with G Newton's constant.

In general relativity, the spherically-symmetric static solution to Einestin's equations is

$$g_{00} = \left(1 - \frac{r_s}{r}\right) c^2, \quad g_{rr} = \frac{1}{1 - \frac{r_s}{r}}, \quad g_{\theta\theta} = r^2, \quad g_{\phi\phi} = r^2 \sin^2\theta \tag{53}$$

and $g_{ij} = 0$ for $i \neq j$. This solution, called the Schwarzschild solution, describes a black hole. This solution is unique up to a single parameter r_s called the **Schwarzschild radius**. Its uniqueness implies black holes have no "hair", meaning that every black hole is identical to an external observer (up to possible conserved charges like electric charge which can be seen through electric field lines ending at the black hole). Note that the solution is singular not only at $r = 0$ but also at $r = r_s$.

In the non-relativistic limit, general relativity reduces to Newtonian gravity. The precise correspondance is that $g_{00} = 1 + 2\Phi$. Matching on to $\Phi = -G\frac{M}{r}$ lets us relate the parameter r_s in the solution to the black hole mass M:

$$r_s = 2\,\frac{MG}{c^2} \tag{54}$$

The sphere at $r = r_s$ called the event horizon. It turns out that nothing inside the event horizon can ever escape. The size (surface area) of the event horizon is

$$A = 4\pi r_s^2 = 16\pi \frac{M^2 G^2}{c^4} \tag{55}$$

Classically, things only fall in to a black hole, so their energy only goes up, and therefore the area of the event horizon only increases.

Because the potential is singular on the event horizon, unusual things can happen. One such thing is that due to quantum field theory the infinite potential energy can be turned into kinetic energy, with photons produced that radiate inwards and outwards. Stephen Hawking showed that the spectrum of these photons is identical to a hot gas (a blackbody, to be covered in Lecture 12) at temperature

$$T = \frac{\hbar c^3}{8\pi G M k_B} \tag{56}$$

This Hawking temperature is inversely proportional to the mass: very small black holes are very hot, and very large black holes are cold. This unusual behavior is associated with a negative heat capacity. Indeed, the specific heat of a black hole is

$$c_S = \frac{1}{M}\frac{\partial M}{\partial T} = -\frac{1}{T} = -\frac{8\pi G k_B}{\hbar c^3} M < 0 \tag{57}$$

As things fall into a black hole, its mass goes up and its temperature goes down. A solar mass black hole has a temperature $T = 10^{-8} K$. A supermassive black hole, like Sagittarius $A^\star$ in the center of our galaxy is about 1 million solar masses and has $T = 10^{-14} K$.

If nothing falls into a black hole, the black hole will completely evaporate due to Hawking radiation in finite time

$$t_{\text{evap}} = 5120\pi \frac{G^2 M^3}{\hbar c^4} \tag{58}$$

As a black hole evaporates, its mass goes down and its temperature goes up. The bigger the black hole, the longer it takes to evaporate. A solar-mass black hole would take 10^{74} years to evaporate. An atom-mass black hole would evaporate in 10^{-98} seconds.

You probably know that the universe is filled with cosmic microwave background (CMB) radiation at a temperature of $3K$. A black hole radiating at this temperature has mass $M_{3K} = 10^{22}$kg, around the mass of the moon. So black holes less massive than M_{3K} will be hotter than the CMB and therefore radiate more energy than they absorb, eventually evaporating. Black holes more massive than M_{3K} will be colder than the CMB; these will absorb CMB radiation slowly increasing their mass. But as they increase in mass, their temperature drops further. Thus, although it is possible for a black hole to be in equilibrium with the CMB if it has exactly the right mass, this equilibrium is unstable. This is the typical behavior of systems with negative heat capacity.

Black holes also have entropy. Since $\frac{\partial S}{\partial E} = \frac{1}{T}$, taking the energy of a black hole as its rest mass, $E = Mc^2$ the entropy is

$$S = \int \frac{dE}{T} = \frac{8\pi G}{\hbar c} \int M dM = \frac{4\pi G}{\hbar c} M^2 = \frac{c^3}{4\hbar G} A \tag{59}$$

Note that black holes have entropy proportional to their surface area. String theory even provides a way of counting microstates for certain supersymmetric black holes that agrees with this formula.

So black holes have entropy, but no hair, and they evaporate in finite time into pure uncorrelated heat. This means that if some data falls into a black hole, it is lost forever. In this way, black holes destroy information and radiate it out as heat, much like Landauer or Bennett's SetToZero operation. There is one important difference though. When we "SetToZero" a bit, the information is not destroyed, just lost by being embedded irretrievably in correlations in the heat. We know this because the law of physics are reversible. When a black hole destroys information it really destroys it – it cannot be stored in correlations of the outgoing radiation because nothing can get out of a black hole, including information. This is the **black hole information paradox**.

To see this another way, information can fall into a black hole well before the radiation is emitted. Since black holes have no hair, that information cannot be accessed in any way by an external observer. For example, suppose we just through bits of information into a black hole, one by one, at such a rate that the energy input exactly equals the thermal radiation rate. Then the black hole's horizon stays constant so the information must be going out in the radiation. However, this is impossible since once something passes the black hole horizon, it can never affect anything outside the horizon. Thus the information really seems to be lost as it falls into the black hole.

The basic conflict is that the laws of gravity and quantum mechanics are deterministic and reversible – if we know the exact starting state, we should be able to predict the exact final state. The precise statement is that in quantum mechanics and gravity, as well as in string theory, time evolution is unitary. Information cannot be lost in a closed, reversible, unitary theory.

The conflict between unitarity and black hole evaporation can be understood clearly with von Neumann entropy. Say the initial state is a wavefunction describing two electrons moving towards each other at super high energy. This is a pure state. They then collide to form a black hole. The black hole then evaporates and the information leaves as heat. The entropy goes up, so the outgoing state is mixed. Thus black holes mediate the evolution from a pure state into a mixed state. This is in conflict with Schrodinger's equation, or more generally, any theory with unitary evolution (such as string theory). If unitarity can be violated by black holes, then it would contribute through virtual effects in quantum field theory to unitarity violation in every other process, in conflict with observation.

9 Summary

We have seen a lot of different ways of thinking about entropy this lecture. The Gibbs entropy is

$$S = -k_B \sum P_i \ln P_i \tag{60}$$

Here P_i is the probability of finding the system in a microstate i and the sum is over all possible microstates i consistent with some macroscopic parameters (volume pressure etc.). In equilibrium, this definition is equivalent to $S = k_B \ln \Omega$ with Ω the number of microstates, but Eq. (60) can be used in any situation where the probabilities are well-defined, including time-dependent non-equilibrium systems.

This Gibbs entropy is proportional to the (Shannon) information entropy:

$$H = -\sum P_i \log_2 P_i \tag{61}$$

In this equation, P_i is the probability of certain data showing up. H has the interpretation as the minimal number of bits needed to encode the data, on average. Information entropy quantifies our ignorance of the system. The more entropy, the less information we have.

An important result from information theory is Landauer's principle: erasing information dissipates heat. This connects information to thermodynamics. When 1 bit of information is erased, the information entropy goes up by 1 bit. Doing so is necessarily accompanied by the release of heat which increases the thermodynamic entropy by the same amount. While the information is technically somewhere encoded in the heated-up molecules, we accept that we will never recover this information and forget about it. Thus we increase our ignorance of the state of the whole system and the entropy goes up.

The broader lesson from this lecture is the modern view of entropy is not as a measure of disorder but as a measure of ignorance. Indeed, the information-theoretic point of view unifies all the different forms of entropy and is the cleanest way to resolve the various entropy-related paradoxes (Gibbs paradox, Maxwell's demon, etc.). It's not that there are two kinds of entropy that we must add: counting microstates and information, but rather that *all* entropy measures the lack of information. When we count the number of microstates Ω these are the states that give the same macroscopic parameters. Thus, given only the information E, V, N etc, Ω measures the things we don't know, namely which microstate we have, consistent with our information. The Gibbs and Shannon entropy formulas are equivalent and therefore both measure ignorance as well. Thus, if you get one thing out of this lecture it should be that **entropy = ignorance**.