# Lesson 1: Entropy and conservation of information

*Notes from Prof. Susskind video lectures publicly available on YouTube*

*The section on Elementary Probability Theory, pages 7 to 20, has been developed by the notetaker from the original lecture.*

## Introduction

Statistical mechanics is not really modern physics. It is premodern physics; it is modern physics; and the reader can be assured that it will be post-modern physics. The second law of thermodynamics will probably outlast anything that comes up.

Time and again, the second law of thermodynamics has been sort of a guidepost, or our guiding light, to know what we are talking about and to make sure we are making sense.

Statistical mechanics and thermodynamics[1] may not be as sexy as the Higgs boson, but it is at least as deep. Many would say a lot deeper. Particle physicists shouldn't disown me. It is a lot more general and it covers a lot more ground in explaining the world as we know it. And in fact without statistical mechanics we probably would not know about the Higgs boson.

So what is statistical mechanics about? To answer this question, let's go back a step. The basic laws of physics, Newton's laws, the principles of classical physics, classical mechanics, the topics that were treated in the classical mechanics course, the quantum mechanics course, the electro-

[1]The science of thermodynamics appeared at the beginning of the XIXth century to improve the recently invented steam engine and to explain its functioning on theoretical grounds.

Statistical mechanics – which is also sometimes called statistical thermodynamics – appeared in the second half of the XIXth century when it began to dawn on some physicists, chiefly the Austrian physicist Ludwig Boltzmann (1844 - 1906), that the still controversial atomic hypothesis provided wonderful new and clearer ways to understand thermodynamics, and conversely the successes of statistical mechanics reinforced the atomic hypothesis.

dynamics course and so forth, in the collection *The Theoretical Minimum*, those topics are all about perfect predictability.

Now you can object that in quantum mechanics we cannot predict perfectly. And that is true. But there are some things we can predict perfectly and those things are the predictables of quantum mechanics.

In all these disciplines, we can make our predictions with maximum precision, or maximal predictability, if we know two things:

1. the *starting point* – which is what we call the initial conditions – and
2. the *laws of evolution* of a system.

When working with a *closed system*, that is a system which either comprises everything, or is sufficiently isolated from everything else that external factors don't influence it, if we know the initial conditions exactly or with sufficient precision, and we know the laws of evolution of the system, then we have complete *predictability*. And that is all there is to say.

In many cases, of course, that complete predictability would be totally useless. Having a list of the positions and velocities of every particle in the room would not be very useful to us. The list would be too long, and subject to rather quick changes as a matter of fact.

So we can see that while the basic laws of physics are extremely powerful in their predictability, they can also be

in many cases totally useless for actually analyzing what is really going on.

Statistical mechanics is what we use in those cases. Let's say first of all that the mathematics of statistical mechanics – statistical mechanics as applied to physical systems – is just elementary probability theory. Some people go so far as saying that statistical mechanics simply *is* probability theory.

When is statistical mechanics applicable? It is when we don't know the initial conditions with complete perfection. It may even be the case if we don't know the laws of motion with infinite precision. And it is applicable when the system we are investigating is not a closed system, that is when it is interacting with other things on the outside.

In other words, in all those situations where ideal predictability is impossible, we resort to probabilities.

However because, for instance, the number of molecules in the room is so large and probabilities tend to become very precise predictors when the law of large numbers is applicable, statistical mechanics itself can be highly predictable.

But that is not true for everything. As an illustration, let's say we have a box of gas – it might even be an isolated closed box of gas. It has some energy in it. The particles rattle around. If we know some things about that box of gas, we can predict other things with great precision. If we know the temperature, we can predict the energy in the box of gas. We can predict the pressure. These things are highly predictable. There are other things, as said, we can-

not predict.

We cannot predict the position of every molecule. We cannot predict when there might be a fluctuation. Fluctuations are phenomena which happen, which don't really violate probability theory, but are sort of tails of the probability distributions. They are things which are unlikely but not impossible. Fluctuations happen from time to time. In a sealed room, every so often a density of molecules bigger than the average will appear in some small region. Someplace else molecules may be briefly less dense. Fluctuations like that are hard to predict.

We can calculate the *probability for a fluctuation*. But we can't predict when and where a fluctuation is going to happen.

It is exactly the same sort of phenomenon which we observe when flipping coins. Flipping coins is a good example – it is probably a favorite one – for thinking about probabilities.

If I flip a coin a million times, you can bet that approximately half of them will come up heads and half will come up tails, within some margin of error. But there will also be fluctuations. Every now and then, if we flip the coin enough times, 1000 heads in a row will come up.

Can we predict when a thousand heads will come up? No. But can we predict how often a thousand heads will come up? Yes. Not very often.

So that is what statistical mechanics is for: making sta-

tistical or probabilistic[2] predictions about systems which contain elements either too small, or too numerous, or for any reason too difficult to keep track of. That is when we use statistical mechanics, in other words probability theory.

We are going to go through some of the basic statistical mechanics applications. We will also cover the theory. We will study the laws of thermodynamics, the laws of statistical mechanics, and then see how they apply to gases, liquids, solids. Occasionally we will look at quantum mechanical systems, which were the subject of volume 2 in the collection *The Theoretical Minimum.*

Another striking feature of statistical mechanics that ought to be mentioned is that all great physicists since the second half of XIXth century were masters of statistical mechanics. Why? First of all because it is useful, but second of all because it is truly beautiful.

It is a truly beautiful subject of physics and of mathematics. And it is hard not to get caught up in it, hard not to fall in love with it.

Let's start this course with a brief review of the main concepts of elementary probability theory.

[2]In this course we tend to use the two adjectives interchangeably, although statisticians and probabilists make technical distinctions.

## Elementary probability theory

What is probability? And why does it work?

Everybody has some intuitive feeling about randomness. Yet to many people it is not far from magic. And it is easy to get tricked or make mistakes (see appendix at the end of the chapter).

To say that probablity works means that if the probability of some event is say $1/3$, and we reproduce many times the experiment producing the event, then it will happen roughly one third of the times. But as we know, this doesn't always work nicely. Sometimes the event will happen more often, or less often, than it should. These are exceptions. And by definition of exceptions they are rare.

Any way we want to pinpoint a good definition of probability, it seems to escape like quicksilver. Our explanations end up involving... probability. But let's give it a try.

In nature there are experiments which, when we replicate them, keep producing exactly the same results. They are predictable or deterministic. They are those we mentioned earlier, appearing for instance in classical mechanics when we can apply Newton's laws.

And there are experiments whose outcomes vary. Those are said to be random. The randomness comes either from some fundamental randomness of nature, like certain phenomena in quantum mechanics, or from our incomplete knowledge, as said, of the initial conditions or other things. Yet the results display some experimental stability in the proportions of occurences, when the experiment is repeated many

times, which we will come back to.

This fundamental distinction established, to work with probability theory we need some primitive ideas about randomness, and then we need to construct a framework.

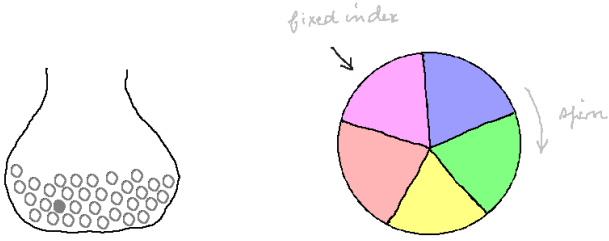


Figure 1: Rare events, and equiprobable events.

There are two primitive ideas, which we present with two illustrative examples, figure 1:

1. *Rare events*: In an urn, suppose there is a *very large* number of white marbles and one black marble. If we plunge the hand in the urn without looking, and pick a marble at random, then picking the black marble would be rare, and we can *"safely"* count on picking a white marble[3].

2. *Equiprobable events*: If the possible outcomes of an experiment display some *symmetry*, then each event

[3]One of the many counterintuitive aspects of probability, which we will have to get used to, is that *any marble*, if we can distinguish each of them, is rare. Yet we do pick one. Seen that way, whatever outcome we obtain is rare. The paradox clears when we distinguish states and events, which are explained in the sequel of the text.

is considered as likely as any other. Thus when we spin the wheel on the right part of figure 1, and wait until it has stopped and the fixed index points to a color, the five colors are said to be equiprobable.

We mentioned the stability of the proportions of occurences, in a random experiment reproduced many times. We intuitively – and correctly – feel that it is linked to the *second idea* listed above. But we shall see in a moment that it is also actually an instance of the *first idea* about rare events.

Let's now turn to the framework to work with a *random experiment* and probabilities. The experiment itself is usually denoted $\mathcal{E}$. Each time we reproduce $\mathcal{E}$, at the end of the experiment the world is in some state $\omega$, which may be different from replication to replication.

We are interested in some aspects of $\omega$: we may want to measure some quantity which has a numerical value, or note a color, or whatever. Sometimes we are interested in only one feature of $\omega$ and sometimes in several, like picking someone at random, in a well-specified procedure, and recording both his height and weight, or level of income and academic degree, or, of a more physical flavor, measuring the position and velocity of a particle.

The main source of mistakes when working with probabilities is the ill-definition of the *experiment* $\mathcal{E}$. For instance, when we simply say "let's pick a person at random", it is ill-defined. The experiment is not sufficiently specified. Do we mean, in the room, or out of a subway entrance in New York, and in that case in which district, at what time of

day, or in the United States, or on Earth?

Another example, of impossibility this time, is when we say "let's pick a point at random uniformly over the entire line". It is actually impossible. There is no such thing as a density of probability with the same value from $-\infty$ to $+\infty$[4]. So we have to be careful.

So, once the experiment $\mathcal{E}$ has been specified, we consider that each time it is performed, the world comes out in some state $\omega$. The set of the $\omega$'s is denoted $\Omega$ and is called the space of states, or set of states. Probabilists call it the universe of possible states attached to performing $\mathcal{E}$.

We are interested in measuring some features, numerical data or non numerical data, about $\omega$. Suppose we are interested in a feature denoted $X$. It depends on $\omega$. In other words it is a function of $\omega$. It is called a *random variable* and the measurement made on a given $\omega$ is denoted

$$X(\omega) \tag{1}$$

It is the result, after having performed the experiment $\mathcal{E}$ once, of the measurement of $X$ on the state $\omega$ that $\mathcal{E}$ produced.

$X(\omega)$ takes its values in the set of possible outcomes of $X$. Let's denote it $A$. Mathematicians write that $X$ is a function from $\Omega$ to $A$ as follows

[4]But a sequence of densities can become more and more flat and therefore with almost the same value – necessarily close to zero – everywhere. Similarly, there is no such thing as a function whose value is zero everywhere except at one point and whose integral is one. But *distributions*, which are limits of functions, can be like that.

$$X : \Omega \to A \tag{2}$$

The set $A$ can be finite, infinite and countable, or infinite and continuous.

For instance if the experiment $\mathcal{E}$ is the throw of a die, $X$ is the number shown on top after a throw, and the set $A$ is

$$A = \{1,\ 2,\ 3,\ 4,\ 5,\ 6\} \tag{3}$$

In this case the abstract set $\Omega$ can just be taken to be the set of results of throwing the die, i.e. $\Omega = A$ itself. Then the $\omega$'s are simply the six possible results, and we don't even have to think about $X$. It is simply the identity.

Although this simplification is often appropriate, it is better – particularly when there are several random variables that we want to measure on the outcome of $\mathcal{E}$ – to clearly distinguish the set of states $\Omega$ that is once and for all attached to $\mathcal{E}$ and the various sets of outcomes of random variables.

Let's stress again that $A$ doesn't have to be numerical. Our die, for instance, could have faces painted with different colors rather than bearing numbers.

Once $\mathcal{E}$, $\Omega$, $X$ and $A$ have been defined, the last fundamental concept to introduce in the framework is a *probability* $P$.

Mathematicians technically talk about a measure of probability $P$ on $\Omega$ that is $\sigma$-additive, etc. And they introduce it

before any random variable. But in this review of elementary probability we don't need to go into this. Furthermore in an elementary approach, it may obscures the link with probabilities as we intuitively know them.

$P$ is defined such that any subset of $\Omega$ – called an *event* – has a probability. In the case of the die, it is particularly simple. Each $\omega$ is itself an interesting event. There are six of them. And if the die is well balanced they are equiprobable. Thus we write

$$P\{X = 5\} = \frac{1}{6} \tag{4}$$

meaning that the probability of getting a 5, when throwing the die, is 1/6. Events are any subsets of $\Omega$, not only the $\omega$'s themselves. We can also write

$$P\{X \leq 2\} = \frac{1}{3} \tag{5}$$

The principle that *symmetry*, or equivalence somehow, between the possible states $\omega$ implies equiprobability – in other words, whatever makes them different doesn't affect their propensity to occur – is often invoked to figure out the distribution of probability $P$ attached to an experiment. Another possible way to figure out $P$ is through a large number of replications of the experiment. We will explain it in a moment.

Finally there may be various probabilistic calculations which we can also make to relate the distribution of a random variable $X$ to those of other random variables we already know. This pertains to the calculus of probability. And we shall

work out many examples in this course.

Our framework is now complete. It consists of a random experiment $\mathcal{E}$, a big set $\Omega$ of possible states of the world after the performance of $\mathcal{E}$, and a probability $P$.

$$\text{framework} = [\ \mathcal{E},\ \Omega,\ P\ ] \tag{6}$$

And we are interested in measuring various random variables $X$, $Y$, $Z$, etc. after having performed $\mathcal{E}$[5].

Let's now simplify a bit the setting and the notations. For the time being, the set of possible states will be

$$\Omega = \{\ \omega_1,\ \omega_2,\ \omega_2, ...\ \omega_n\ \} \tag{7}$$

that is a finite set of outcomes of $\mathcal{E}$. The states are indexed by $i$ running from 1 to $n$. For example, when flipping a coin once, $\omega_1$ = heads, and $\omega_2$ = tails.

Later on, we will extend this to an infinite countable, and then even an infinite countinuous set $\Omega$[6].

[5] In maths manuals, the reader will usually see the framework described as $[\ \Omega,\ \mathcal{A},\ P\ ]$, the experiment $\mathcal{E}$ not being mentioned – which in our teaching experience is regrettable. And the extra $\mathcal{A}$, not to be confused with the target set of any random variable, is the collection of subsets of $\Omega$, but not quite all of them, only the "measurable" ones. We won't be concerned here with those subtleties.

[6] In this last case probabilities will be replaced by *densities of probability*. Instead of considering $P\{X = x\}$, which will usually be equal to 0, we will consider $P\{\ X \in [\ x,\ x + dx\ ]\ \} = p(x)dx$. And, following the custom, we will often still denote it $P(x)dx$, keeping in mind that each random variable has its own density.

Staying with a finite set $\Omega$, the probabilities of the single states $\omega_i$'s will simply be denoted

$$P(i) \tag{8}$$

They must satisfy

$$\begin{aligned} P(i) &\geq 0 \\ \sum_{i}^{n} P(i) &= 1 \end{aligned} \tag{9}$$

Indeed, probabilities are positive numbers. And the total probability, when we add up the probabilities of all possibilities, should be equal to one. When performing $\mathcal{E}$ we certainly should get some result.

Probabilities have all sorts of interesting, beautiful and sometimes surprising properties. The most useful one for us in this course is the *law of large numbers*.

Here is what it says. Suppose that we either make many replicas of the same system, or do the same experiment $\mathcal{E}$ over and over a large number $N$ of times, and we count how many times we get the $i$-th outcome $\omega_i$. That is some count that we denote $N_i$. Then the law says that

$$\lim_{N\to\infty} \frac{N_i}{N} = P(i) \tag{10}$$

This is a statement about probabilities, which can be stated more precisely and rigorously within the framework $[\ \mathcal{E},\ \Omega,\ P\ ]$. But let's approach it at an intuitive level. It says that when we replicate $\mathcal{E}$ a larger and larger number of times,

and measure the *experimental frequency* of occurence of the $i$-th outcome, this experimental frequency gets closer and closer to the actual probability $P(i)$.

For instance, if we toss a coin a thousand times, the frequency of heads will be close to $1/2$. If we throw it 10 000 times, the frequency will be even closer to $1/2$. In each case, it is only a probabilistic statement. There can be – and in fact most of the times there will be – a discrepancy. That discrepancy is itself a random variable. But it will have a distribution more and more concentrated, relatively to its range, around 0.

We said earlier that the convergence of experimental frequencies toward their corresponding theoretical probabilities is actually an instance of the rare event idea, see figure 1 and its comments.

The law of large numbers is neither magic, nor some kind of eerie principle of nature. It stems from the fact that in an urn with one black ball and very many white balls, if we pick one at random, we can assume safely that we will pick a white one. It is not always the case, but it will be extremely rare to pick a black ball. And for all practical purpose it can be neglected.

Let's see why, in the case of tossing coins, the law is a *simple result in numbering*. Consider the experiment $\mathcal{F}$ which consists in tossing the coin 1000 times, i.e. repeating $\mathcal{E}$ 1000 times. The space $\Omega_{\mathcal{F}}$ attached to $\mathcal{F}$ has $2^{1000}$ elements – that is a huge number. Each are equiprobable. When we perform $\mathcal{F}$ once, i.e. when we repeat $\mathcal{E}$ a thousand times, we pick one element in $\Omega_{\mathcal{F}}$.

It turns out – and it is not hard to show, although we won't do it – that most elements in $\Omega_{\mathcal{F}}$ contain about as many heads as tails. View them as the white balls in the urn if you like. And the black balls would be very few. So when we pick one, we pick a white one.

To try to shed even more light on the phenomenon, rather than do some combinatorics, consider the 16 possible results, displayed below, of throwing the coin four times. The reader can check that there is only one result with zero head. Four results with 1 head. Six results with 2 heads. Four results with 3 heads, and one result with 4 heads.

$$
\begin{array}{c}
T,\ T,\ T,\ T \\
T,\ T,\ T,\ H \\
T,\ T,\ H,\ T \\
T,\ T,\ H,\ H \\
T,\ H,\ T,\ T \\
T,\ H,\ T,\ H \\
T,\ H,\ H,\ T \\
T,\ H,\ H,\ H \\
H,\ T,\ T,\ T \\
H,\ T,\ T,\ H \\
H,\ T,\ H,\ T \\
H,\ T,\ H,\ H \\
H,\ H,\ T,\ T \\
H,\ H,\ T,\ H
\end{array}
$$

$$H,\ H,\ H,\ T$$
$$H,\ H,\ H,\ H$$

So there is a kind of concentration around an equal number of heads and tails. The counts of the number of heads actually correspond to the so-called *Pascal triangle*. They are also the coefficients in the development of the polynomial $(a + b)^N$. The concentration about half and half is more marked, of course, when $N$ is larger than 4, and it grows more and more marked as $N$ increases.

That is what the law of large numbers is about. In probability theory, it is stated more rigorously than we have done here. It is proved via an intermediate result called *Chebyshev inequality*[7]. It is not particularly hard, and is rather elegant. But it is outside what we want to do in this review[8].

The law of large numbers, expressed by equation (10), says that the ratio $N_i/N$ converges toward $P(i)$ when $N$ gets very large. In other words,

*when we have repeated an experiment a large number of times, we can use the experimental frequency $N_i/N$ of occurence of the $i$-th outcome as an estimate of $P(i)$.*

[7]Named after Pafnuty Chebyshev (1821 - 1894), Russian mathematician.

[8]Another beautiful and useful result is the *Central Limit Theorem*, which shows in essence that the Pascal triangle looks more and more like a bell-shaped curve called a Gaussian. And it is true in a much more general setting than just flipping a coin many times.

We use this result all the time.

Now let's go back to a random variable we want to measure, which is not the outcome of $\mathcal{E}$ itself. To get closer to physics concepts and notations, let's call the random variable $F$. So let's suppose that there is a quantity, denoted $F(i)$, that is associated with the $i$-th state $\omega_i$. Recall expression (7) defining the set of states.

$F$ can be some meaningful physical quantity. We can also make it up. For example if our system is heads and tails, and nothing but heads and tails, we could assign

$$\begin{aligned} F(\text{heads}) &= +1 \\ F(\text{tails}) &= -1 \end{aligned} \tag{11}$$

If our system has many more states, we may want to assign a larger number of possible values taken by $F$ – not necessarily the same number as the number of elements in $\Omega$ though. $F$ is simply some function of the states. We already mentioned this in expression (2), let's write it again

$$F : \Omega \to A \tag{12}$$

The random variable $F$ acts on the set of states $\Omega$ and takes its value in the set $A$. In the case of the coin, $\Omega = \{H,\ T\}$, and the set $A$ in which $F$ takes its values is $\{+1,\ -1\}$. Thus, we have made up a numerical random variable – or *measurement* – attached to flipping a coin.

As said, $F(i)$ can also be some meaningful physical quantity. It could be the energy of the $i$-th state $\omega_i$. Given the state in which is some system, it has an energy. Its measurement would perhaps be called in that case

$$E(i) \tag{13}$$

Or it could be the momentum of the state. We would have to choose a good notation not interfering with probabilities. Or it could be something else. It could be whatever we happen to like to measure on our system.

Then an important quantity is the *average* of $F(i)$. We will use the quantum mechanical notation for it, even though we are not doing quantum mechanics. It is a nice notation. Physicists tend to use it all over the place. Mathematicians hate it. We just put a pair of brackets around $F$ to mean its average. It is defined as follows

$$< F > = \sum_{i}^{n} F(i)P(i) \tag{14}$$

In words, it is the average of the values $F(i)$ weighted by their respective probabilities $P(i)$.

Notice that the average of $F(i)$ does not have to be any of the possible values that $F$ can take. For example, in the case of the coin, where $F(H) = +1$, and $F(T) = -1$, and we flip it a million times, and the probability is $1/2$ for heads and $1/2$ for tails, the average of $F$ will be 0. It is not one of the possible outcomes of $F$ in the experiment. Yet it is its average. There is no rule why the average of a measure should be one of its possible experimental values.

Thanks to the law of large numbers, we can write equation (14) another way.

$$< F > \approx \sum_{i}^{n} F(i) \frac{N_i}{N} \tag{15}$$

This approximate equality becomes a true equality in the limit when there is a large number of measurement.

That is it for our mathematical preliminary. We need to know what a *random experiment* is, what a *probability* is, what a *random variable* or random measurement is, and what is an *average*, because we will use them over and over.

Now we shall go deeper into the link between probability and symmetry. And we shall introduce time.

## Probability, symmetry and systems evolving with time

Let's start with coin flipping again. For the usual coin, the probability for heads is usually deemed to be 1/2, and the probability for tails is usually also deemed to be 1/2. But why do we think that? Why is it 1/2 and 1/2? What is the logic there?

In this case it is symmetry. Of course no coin is perfectly symmetric. Making a little mark on it, however tiny, to distinguish heads and tails biasses it a little. But apart from that tiny bias, for example a small scratch, the coin is symmetric. Heads and tails are symmetric with respect to each other. Therefore there is no rationale, when we flip a

coin, for it to turn up heads more often than tails.

So it is symmetry quite often – we might even say *always*, in some deeper sense, but at least in many cases – which dictates probabilities.

Probabilities are usually taken to be equal for configurations which are related to each other by some symmetry. Symmetry means if you act with a symmetry, if you reflect everything, if you turned everything over, that the system behaves the same way.

Another example besides coin tossing would be die throwing. Now the space $\Omega$, instead of having two states, has six states. When we throw the die, and it has finished rolling, the state it is in is the face showing up. To stress that states don't have to be numerical values, let's consider a die with colored faces, figure 2.

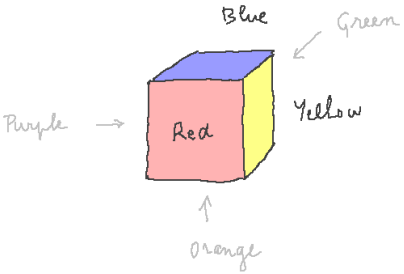


Figure 2: Die with colored faces.

So here we don't keep track of numbers. We keep track of colors. The space of states is

$$\Omega = \{ \text{ red, yellow, blue, orange, green, purple } \} \quad (16)$$

In other words, $\omega_1 = \text{red}$, $\omega_2 = \text{yellow}$, $\omega_3 = \text{blue}$, etc.

What is the probability that after a throw the die turns up, for instance, blue like in figure 2? It is 1/6.

Why? Because there are six possibilities. They are all symmetric with respect to each other. We use the principle of symmetry that tell us that the $P(i)$'s are all equal, therefore

$$P(i) = \frac{1}{6} \qquad \text{for all } i \quad (17)$$

But what if the die is not symmetric? For example, if it is weighted in some unfair way. Or if it is been cut with faces that are not nice parallel squares, thus not making it an honest cube.

Then symmetry won't tell us anything. There are cases where we may be able to use some deeper underlying theory in which symmetry plays a role. But, in the absence of such deeper information, there is no theoretical way to figure out the probabilities.

In that case we resort to experiment. Do the experiment $\mathcal{E}$ a billion times; keep track of the numbers; assume that things have converged. And that way we measure the probabilities and thereafter we can use them.

So either we have some symmetry, or some underlying theory like in quantum mechanics, that tell us what the probabilities are, or we do experiments to get them. Whatever

the case, we end up with probabilities.

Statistical mechanics tends to rely mostly on symmetry as we will see. But if there is no symmetry to guide us in our determination of probabilities then it is experiment.

Now there is another possible answer. This answer is often frequently invoked and it is a correct answer under the circumstances. It can have to do with the evolution of a system, the way a system changes with time. Let's see some examples.

Let's take our six-sided polyhedron, and assume that *it is not* a nice symmetric cube. Furthermore suppose this object has the following funny habit: when it is in one state, it then jumps to another definite state, then to another definite state, etc. It is called the law of motion of the system.

The *law of motion of a system* is that wherever it is – i.e. in whatever state, also called a *configuration* – at one instant, at the next instant it will be in some other state according to a definite rule.

The time spent in each state is some seconds or some microseconds or whatever. And they are all equal. And for simplicity the jumps are instantaneous. So we imagine a discrete sequence. And let's suppose there is a genuine law that tells us how this funny cube moves around.

We have already described such a system evolving over time in different contexts – in chapter 1 of volume 1 of the collection *The Theoretical Minimum*, on classical mechanics, for instance – but it is so important that I feel a need to

emphasize it again.

A law of motion is a rule telling us what the next configuration will be, given where we are. It is a rule of updating configurations. Here is an example. The letters are abbreviations for the colors in expression (16).

$$
\begin{aligned}
\mathrm{R} &\longrightarrow \mathrm{B} \\
\mathrm{B} &\longrightarrow \mathrm{G} \\
\mathrm{G} &\longrightarrow \mathrm{Y} \\
\mathrm{Y} &\longrightarrow \mathrm{O} \\
\mathrm{O} &\longrightarrow \mathrm{P} \\
\mathrm{P} &\longrightarrow \mathrm{R}
\end{aligned}
$$

Given the configuration, that is the state, in which the system is at any time, we know what it will be next. And we know what it will continue to do.

Of course we may actually not know the law. Maybe all we know is that *there is a law* of the type above.

Anyway, let's draw the one above as a diagram, figure 3.

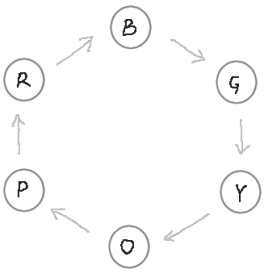


Figure 3: Law of motion.

And we make the assumption that the same amount of time is spent in each state.

We are not assuming that the cube has any symmetry anymore. It may not be symmetric at all. Imagine that the faces are not parallel, they are quadrilaterals but not squares. Some are big, some are small. They may even be not quite flat, etc. But figure 3 is the rule to go from one configuration to another. And each step takes, let's say, a microsecond.

We might have no idea when and in which state the system began. But suppose our job is to catch it at a particular instant and ask what the color is. Even though we don't know the starting point, we can still say that the probability for each state is $1/6$.

It is not related to any symmetry of the die, and as we said the die may not be symmetric at all. But it is a consequence of the symmetry of the time structure. As it passes through the sequence of states, the die spends one sixth of its time red, one sixth of its time blue, one sixth of its time green, etc.

If we don't know where it starts, and we take a flash snapshot, the probability will be $1/6$ for each state[9]. It doesn't really depend on knowing the detailed law. For example the law could have been different. Figure 4 shows a different example

[9]Here is a good example of the importance of specifying precisely what is the random experiment $\mathcal{E}$ that, at least in principle, we can reproduce.

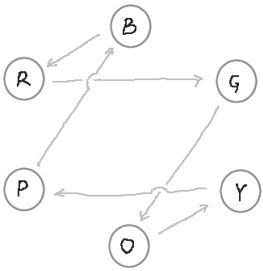


Figure 4: Another law of motion.

This shares with the previous law that there is a closed cycle of events in which we pass through each color once before we cycle around.

We may not know which law of nature is for the system, figure 3 or figure 4 or even another law which makes a complete cycle through all the colors, but we can say again that the probability will be 1/6 for each one of them.

So this prediction of 1/6 doesn't depend on knowing the starting point, and doesn't depend on knowing the law of physics. It is just important to know that there was a particular kind of law.

Are there possible laws for the system which will not give us 1/6? Yes. Let's write another law, figure 5.

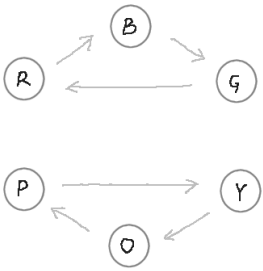


Figure 5: Law of motion with two cycles.

This rule says that if we start with red, we go to blue. If we are in blue, we go to green. And if we get to green, we go back to red.

Or if we start with purple, we go to yellow. Yellow goes to orange. And orange goes back to purple.

Notice in this case, when we are on one of the two cycles, we stay on it forever.

Suppose we knew we were on the upper cycle. It doesn't matter knowing in which state we started, but if we knew that we started on the upper cycle somewhere, then we would know that there is a 1/3 probability to be in the red state when we flash the system, 1/3 probability to be in blue, 1/3 probability to be in green. And of course there is 0 probability to be in yellow, orange or purple, since we are not on their cycle.

The probabilities 1/3, 1/3 and 1/3 to be in red, blue or green are also called the *conditional probabilitie*s, given that

we know we are on the first cycle.

On the other hand, we could have started on the second cycle. We could have started with purple. Or we might not know exactly where we started, just that we started on the lower cycle. Then the probabilities are: purple $1/3$, yellow $1/3$ and orange $1/3$. And again, of course, 0 for the other colors.

Now what about a more general case? It might be that we know with some probability, that we started on the upper cycle and with some other probability on the lower cycle.

In fact, let's give the cycles labels. The upper cycle we call $+1$. And the lower triangle we call $-1$. It is just attaching to them a number, a numerical value. If we are on the upper cycle, something or other is called $+1$, and below it is $-1$.

Now we have to append something that we get from some place else. It doesn't follow from symmetry. And it doesn't follow from cycling through the system. It is the probability[10] that we are on cycle $+1$, or its complement to one that we are on cycle $-1$.

Where might that come from? Well, flipping somebody else's coin might decide which of these two cycles we shall

[10] Again, talking about a probability means talking about a *random experiment* – at least an implicit one – that can be repeated. This touches on the debate between Frequentists and Bayesians in statistics. The latter are willing to consider probabilities of events in experiments that can be performed only once, whereas the former require that there exist a *reproducible experiment*. We refer the reader to the statistics literature on the subject.

be on. And the coin doesn't have to be balanced. So we have two more probabilities to consider:

$$\begin{array}{l} P\{\text{ cycle} = +1 \text{ }\} \\ P\{\text{ cycle} = -1 \text{ }\} \end{array} \tag{18}$$

And they must add up to 1. These two probabilities are not probabilities for individual colors. They are probabilities for individual cycles.

Now what is the probability for blue? From his or her elementary probability course, the reader should remember that it is the product of the probability that we are on cycle +1 times the conditional probability that we get blue, conditioned on knowing that we are on cycle +1.

$$P\{\text{blue}\} = P\{+1\}\ P\{\text{ blue} \mid +1\} \tag{19}$$

where $P\{\text{blue}\}$ is the overall probability to get blue, $P\{+1\}$ is an abbreviation for $P\{\text{ cycle} = +1\text{ }\}$, and $P\{\text{ blue} \mid +1\}$ is the standard notation for the conditional probability to get blue given that we already know that we are on the first cycle.

Since the three conditional probabilities, once we are on cycle +1, are 1/3, formula (19) becomes

$$P\{\text{blue}\} = P\{+1\}\ \frac{1}{3} \tag{20}$$

And we have similar formulas for the other colors. Of course, in the case where $P\{+1\} = 1/2$, we get back to $P\{\text{blue}\} = 1/6$.

So in the case where there are several different possible cycles, we would need to supply another probability that we must get from somewhere else.

The example we just went through, with several cycles, is what we call having a *conservation law.*

In our example, the conservation law would be just the conservation of the number assigned to the cycle. For red, blue and green we have assigned the value $+1$. That $+1$ could be the energy, or it could be something else, whatever. Let's just think of it as the energy to keep things familiar.

The energies of the three configurations in the upper cycle might all be $+1$. And the energies of the three configurations in the lower cycle might be $-1$. The point is that, because the rule of motion keeps us always on the same cycle, the quantity – energy or whatever – is conserved. It doesn't change. That is what a conservation law is.

In summary, a conservation law says that the configuration space divides up into cycles, or trajectories if they don't loop, and that in each of them a certain characteristic of the configuration doesn't change. While the system jumps from configuration to configuration as time passes, it stays on the same cycle or trajectory. As a consequence the quantity, whose value is determined only by the cycle or trajectory, is conserved.

The cycles don't have to have equal size, figure 6.

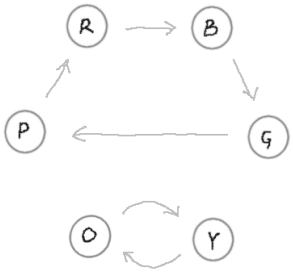


Figure 6: Example of cycles with different size.

We could have a conservation law in this space of configuration, if there is a quantity with one value for the upper cycle, and another value for the lower cycle. The number of states don't have to be the same in each cycle. But still there can be a conservation law.

And again somebody would have to supply for us some idea of the relative probabilities of the two cycles. Where that comes from is part of the study of statistical mechanics.

Another part of the study has to do with the following question: if I know I'm on one of these tracks – another name for cycles or trajectories – how much time do I spend with each particular configuration?

That is what determines probabilities in statistical mechanics: some probabilities, coming from somewhere, which tell us the probabilities for different conserved quantities, and then going from configuration to configuration through the configuration space.

**Questions / answers session**

Question: So, we assume that once we are on a given cycle, the probabilities of all the configurations on that cycle are the same?

Answer: Right. That comes from the fact that the time spent in each state is the same.

We are in classical physics. Everything is deterministic. The randomness that we introduce is only a consequence of our imperfect knowledge of the initial conditions of the system, or of the exact time at which we observe it.

Because we don't have precise enough instruments, or because we are lazy, we don't know exactly in which state the system was at time zero, or we don't know exactly the time at which we take the measurement of the color.

Let's focus on figure 3 where there is only one overall cycle. Suppose we take a measure at time $t$. Suppose also we know that the system started in state blue, and it stays one microsecond in each state, and $t = 2$ *minutes* 17 *seconds* 754 *milliseconds* and 123 *microseconds*. Then it is easy to calculate that the system should be in orange. But if the precision of our timing is only one millisecond, taking a flash[11] of the system amounts to a random experiment where the probability of each of the six states is $1/6$.

That's the circumstance that we are talking about.

[11]The flash is assumed to be instantaneous. There is no blur between several colors.

Q.: If we take two pictures, presumably we will be able to check that we staid in the same cycle?

A.: Oh yes. If there are several distinct cycles, like for instance in figure 5, and we take two pictures one after the other, the two colors we shall obtain will be on the same cycle.

In fact, once we have made a first picture, we know in which cycle we are, and there are no more probabilities for cycles to heed. Now the probabilities are uniform over the configurations of the cycle we are on.

In other words, once you determine the value of some conserved quantity, then you know it. And then you can reset the probabilities for the various possible configurations of the system.

Let's talk about honest energy for a minute. Suppose we have a closed system. To represent the close system we just draw a box

Figure 7: Closed system.

Now closed means that it is not in interaction with anything else and therefore can be thought of as a whole universe

unto itself. It has an energy. The energy is some function of the state of the system, whatever determines the state of the system.

Now let's suppose we have another closed system which is built out of two, identical or not identical, versions of the same thing.

If they are both closed systems there will be two conserved quantities, the energy of the left system and the energy of the right system. They will both be separately conserved. Why? Because they don't talk to each other. They don't interact with each other. The two energies are conserved. And the probability frameworks for each system will just be independent experiments and analyses.

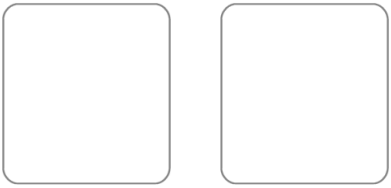

Figure 8: Two closed systems.

But now supposing they are connected by a little tube which allows energy to flow back and forth, figure 9. Then there is only one conserved quantity, the total energy, which is sort of split between the two systems.

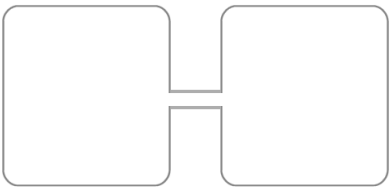

Figure 9: Two systems connected.

We can ask: given a total amount of energy, what is the probability that the energy of one subsystem is one thing and the energy of the other subsystem is the other?

If the two boxes are equal, we would expect that on the average they have equal energy. But we can still ask what is the probability that energy in the left box has such and such value, given some overall piece of information.

So, if we consider only the system on the left in figure 9, that is a circumstance where the probability for which cycle we are on may be determined by thinking about the system as part of a bigger system. And we are going to do that. That is important.

So in general, you need some other ingredients besides just cycling around through the system, represented in diagram 5 or diagram 6, to tell you the relative probabilities of conserved quantities.

This closes the questions / answers session. So we are off and flying in statistical mechanics. Now let's turn to some fundamental laws, and in particular to some forbidden.

## Bad laws of motion and the $-1^{\text{st}}$ law of physics

By bad laws, we don't mean in the sense of DOMA[12] or any of those kind of laws, but in the sense that the principles of physics don't allow them.

If you have read our previous courses, particularly volume 1 of the collection *The Theoretical Minimum* on classical mechanics, you know that a bad law of motion is one that violates the conservation of information – the most primitive and basic principle of physics.

Conservation of information is not a standard conservation law like those we just studied, saying that there are certain quantities which are conserved when a system evolves over time through a cycle or any other kind of trajectory.

Conservation of information is the principle which says that we can keep track of our trajectory both going forward and going backward. We described it in detail in the classical mechanics book, but let's review it briefly.

Figure 10 below is an example of a bad law. There are six states, or configurations, for the system, but the rule says that wherever we are, the next state is red. Even if we are at red, we go to red.

We readily notice that it is not reversible. We can follow a trajectory when the time passes forward. It is very simple: we always end up at red and then stay there. But if we look at the trajectory in reverse, with the time passing

[12] Defense Of Marriage Act.

backwards, it doesn't correspond to a well defined law.

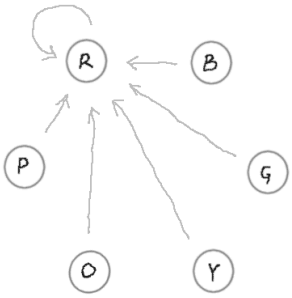


Figure 10: Example of bad law. It keeps track of where we go, but not of where we come from.

When we are at red, where would the law say we should go? It wouldn't say, because there are several arrows pointing at red in figure 10, and therefore when we reverse time several arrows would leave red for different states. That would not be a law at all, because in physics a law of evolution must be deterministic[13].

In other words, the law in figure 10 is bad because it loses track of where we started and therefore is not reversible.

We could make it more complicated, and which still loses track of the past, if we wanted to. But the story would be the same. A law must be deterministic, and it must be re-

[13]Remember that even in quantum mechanics laws of motion are deterministic. When a system is in a state $\psi(t)$, Schrödinger's equation tells us in which state $\psi(t+dt)$ it is going to be next. Measures of observables don't give deterministic results. But that is another story. In quantum mechanics states and measures are different things, see volume 2 in the collection *The Theoretical Minimum* for details.

versible. This second characteristic is equivalent to saying that we don't lose track of the past.

We can express the principle simply:

*A good law of motion is one such that for each state, one and only one arrow arrives at it, and one and only one arrow leaves it.*

The laws in figures 3, 4, 5 and 6, were all good laws according to this principle. In any of those examples, if we start from any state $S_1$ and go, say, 56 steps forward and end up in state $S_2$, then starting from $S_2$ and going backward 56 steps we will end up in state $S_1$. They don't lose information.

The law in figure 10 loses information. That is exactly the kind of thing that classical physics doesn't allow. Quantum physics also doesn't allow laws of motions that are not deterministic in the future, or not deterministic in the past, i.e. not reversible.

There is no name for this law because it is so primitive that everybody always forgets about it. But I call it the $-1^{\text{st}}$ *law of physics*. I wish that it would catch on, that people would start using it, because it is the most basic law of physics, that information is never lost[14].

[14]It was the crux of a debate between Stephen Hawking and Leonard Susskind on black holes. S. Hawking maintained that when a black hole absorbed an object information was lost. L. Sussking maintained that it was a violation of the $-1^{\text{st}}$ law and therefore was incorrect; the information must be preserved somewhere. See L. Susskind's book *The Black Hole War*, Back Bay Books, 2009.

The $-1^{\text{st}}$ law says that distinctions or differences between states propagate with time. Nature never loses track of the difference between two states. In principle, if we have the capacity to follow the system, from wherever we are we will always know exactly where we go, as well as where we come from. Because we don't have the appropriate instruments, or we are too lazy, we may lose track, but nature doesn't have this problem and keeps distinctions. And in principle we could too because nature – by definition – doesn't hide anything from us.

To a physicist or a natural philosopher it doesn't make sense to say that some things *are*, or *have been* in the past, but we cannot one way or another be aware of them. Thus we see that the $-1^{\text{st}}$ law is not a stringent requirement. It is at the foundation of what knowledge means, as opposed to phantasmagoric beliefs.

In continuum classical mechanics, there is a version of this law. It is called Liouville's theorem[15]. We studied it in classical mechanics, see volume 1. We will take a second look at it in this lesson, after we have introduced entropy.

Let's see an apparent counterexample to Liouville's theorem. Friction seems to be such a case. If we consider an object sliding with friction, it eventually comes to rest. It is sort of equivalent to always ending up at red in figure 10. That seems like a violation of the law that tells us that the distinctions have to be preserved.

But of course it is not really true. What is really going on

[15]Named after Joseph Liouville (1809 - 1882), French mathematician.

is that as the object slides, decelerates and stops, it is heating up the surface on which it slid. If we could keep track of every molecule we would find out that the distinctions between trajectories is recorded. In other words, two different starting states of the universe would lead, when taking into account everything, to two different ending states of the universe, even if the object ends up at the same place.

Let's imagine now that there was a fundamental law of physics for a collection of particles. The particles are labeled by an index $n$, they each have a position $x_n$, possibly multi-dimensional, depending on the time $t$. And they satisfy the equation

$$\frac{d^2x_n}{dt^2} = -\gamma \, \frac{dx_n}{dt} \tag{21}$$

The left-hand side is acceleration. And the right-hand side corresponds to a frictional force opposing motion. The factor $\gamma$ is the viscous drag. We could put masses on the left-hand side multiplying the second derivative, but that would not change anything for our purpose, so let's take them all equal to 1.

Again this is friction. Equation (21) has the property that if we start with a moving particle it will very quickly come almost to rest. It will exponentially come to rest, and that will happen fast if there is a significant drag.

Imagine that all the particles, in a gas for example, satisfied this law of physics. It is perfectly deterministic, back and forth. It tells you what happens next, and it is reversible.

But it has the unfortunate consequence that every particle almost comes to rest. That sounds odd. It sounds like no matter at what temperature we start the room, it will quickly come to zero degree. It does not happen that way.

Equation (21) is a perfectly good differential equation. But there is something wrong with it from the point of view of conservation of energy, and from the point of view of thermodynamics.

If a closed system starts with a lot of kinetic energy – we usually call it *temperature* –, it doesn't evolve to zero temperature. That is not what happens.

And it is not only a violation of energy conservation, it also looks like a violation of the second law of thermodynamics. Equation (21) describes a process where things get simpler. We start with a random bunch of particles moving in random directions, and we let it run, and they all come to rest. What we end up with is simpler and requires less information to describe than what we started with. It is very much like, in figure 10, everything going to red.

The second law of thermodynamics generally says things get worse. Things get more complicated not less complicated. So even though it is mathematically sound, deterministic in the future and in the past, equation (21) is not a good law of physics in view of the law of conservation of information.

We saw that one way to express the $-1^{\text{st}}$ law, or law of conservation of information, is to say that every state must have one arrow coming in and one arrow going out. There

is another important way to formulate it.

### Conservation law and probability

Let's consider a system and the collection of states or configurations, indexed by $i$, it can be in. And we assign probabilities, $P(1)$, $P(2)$, $P(3)$, etc. that is, probability, when we look at it, that the system be in state 1 – earlier called $\omega_1$ –, probability that the system be in state 2, probability that it be in state 3,... for *some subset* of the complete set of possible states. All the others, we say that they have probability zero.

For example, we can take our colored die in figure 2 and assign red, yellow and blue, each probability 1/3. And the other three colors have probability zero. Where we got that from? It doesn't matter. We got it from somewhere. Somebody secretly told us in our ear: "It is either red, yellow or blue, but I'm not going to tell you which."[16]

[16]The upper face of the die actually shows one color, that is the die is in one specific state, but for some reason we don't know it. We only know the partial information given by the chap.

We could also imagine the die having three faces with three shades of green, and the others with three shades of purple. And, because we did not take a good enough look, we only got the general color of the state, not the precise shade.

And to avoid being in a Bayesian setting where probabilities don't correspond to any reproducible experiment – and therefore, to many of us, would be meaningless –, let's imagine that at least theoretically we could play the game many times and those three colors would turn up randomly with three equal probabilities, because the die is loaded or for any other reason.

Now we follow the system as it evolves in time. The system evolves according to whatever law of physics, as long as it is an allowable law of physics.

After a while, suppose we are capable of following it in detail. We are no longer constrained by our laziness or inaccurate instruments in this case. Then what are the probabilities at a later time?

Well, if we don't know which the laws of physics are, we can't say of course. But we can say one thing: we can say there are three states with probability 1/3, and three states with probability zero.

They may get reshuffled, which ones are probable and which ones are improbable. But after a certain time, there will continue to be three states which have each probability 1/3 and the rest probability zero.

Let's be very precise, because when talking about probabilities it is easy to trip into confusion. If we reproduce the experiment many times here is what will happen. And we suppose we have omnipotent observational powers of colors and time. At first the three colors red, yellow and blue are the possible initial colors. So they will each appear roughly one third of the time. And after a specific time $t$, that is a specific numbers of steps, the three corresponding states after time $t$ will also appear – necessarily – roughly one third of the time. There is no loss of information, even though the initial conditions are random between three states from our point of view.

So in general we could characterize these information con-

serving series as follows. We have a total number of states $N$, and a number of possible states $M$.

$$\begin{aligned} N &= \text{total number of states} \\ M &= \text{number of possible states} \\ M &< N \\ P &= \frac{1}{M} \end{aligned} \tag{22}$$

The number of possible states stays the same over time – even though they may get reshuffled among the $N$ states as time passes. And of course, at a given time, the probability of any of the possible states at that time is $P = 1/M$.

That is a different characterization of the information conserving law. It introduces probability, but it specifies that the randomness over initial conditions is carried along with time, with no further loss. Then the number of states which have non-zero probability will remain constant and their probabilities will remain equal to $1/M$.

As we have described, these possible states may reshuffle as time passes, but the number with non-zero probability will remain fixed[17].

For the information non-conserving laws every configuration goes to red. In figure 10, we may start with a prob-

[17] As already explained, if we could reproduce the *random experiment* generating these probabilities, and we had omnipotent powers to measure time and states, we would see that at time 0 a certain set of $M$ states occur more or less equally frequently, the others having 0 probability. And at another time $t$, it would be another set of states which would be possible, the others having 0 probability. This other set would have the same number $M$ of states as a consequence of the conservation of information, and they would still be equiprobable.

ability distribution that is for instance 1/3 for red, green and purple and 0 for the others. And then a little bit later there is only one state that has a non-zero probability. And that is red.

So we now have another way to describe information conservation. And we can quantify that. Let $M$ be the number of states which are possible, and let's assume they all have equal probability. These states have a name. They are called the *occupied states* – states which have non-zero probability, with equal probability. Again these may shift around as time passes. But $M$ doesn't change.

What is $M$ characterizing? $M$ is characterizing our *ignorance*. The bigger $M$ is, the greater is our ignorance. If $M$ is equal to $N$, that means equal probability for every configuration, or maximal ignorance.

If $M$ is equal to $N/2$, that means we know that the system is in one out of half the states. We are still pretty ignorant, but we are not that ignorant. We are less ignorant.

What is the minimum amount of ignorance we could have? That is when we know precisely, at any time, in which state the system is. In that case $M = 1$. That is when we have no ignorance resulting from one cause or another. You could say that we are then the omnipotent observer that we talked about in explaining the probabilities attached to the occupied states.

So $M$, in relation to $N$, is a measure of our ignorance. And, associated with it, is the concept of entropy.

## Entropy

We come to the concept of entropy. Notice that entropy is coming before anything else. Entropy is coming before temperature. It is even coming before energy. Entropy is more fundamental in a certain sense than any of them.

But it is also different from temperature or energy in the sense that it is not an absolute characteristic of the system – irrespective of what the observer is or does –, but is a characteristic of *the system plus the observer*. This will become clear in a moment.

The entropy is denoted by the letter $S$, and it is defined as the logarithm of $M$:

$$S = \log M \tag{23}$$

It is the logarithm of the number of states that have an appreciable probability, more or less all equal, in the specific circumstances that we described.

When there is no loss of information, the entropy $S$ is conserved. All that happens is that the occupied states reshuffle. But there will always be $M$ of them with probability $1/M$. That is the law of conservation of entropy – if we can follow the system in detail[18].

[18]There is a subtle point here. Think, for a moment, that the initial randomness is not due to our imperfect observations but to some other cause. Randomness in the initial conditions means that if we reproduce the experiment with the system, they will be different. But they belong to a set of $M$ initial states. Then, if we can observe things precisely, this randomness propagates with no loss of information as time goes on. Thus, if there is no loss of information, entropy is conserved.

Now of course in reality we may be again lazy, lose track of the system. We might have, after a point, lost track of the equations, lost track of our timing device, and so forth and so on. Thus we may have started with a lot of knowledge and wind up with very little knowledge.

That is not because the equations cause information to be lost[19], but because we just weren't careful. Perhaps it is impossible to be perfectly careful – to be an omnipotent observer –, perhaps there are too many degrees of freedom to keep track of.

So when that happens, the entropy increases. But it simply increases because our ignorance has gone up, not because anything has really happened in the system. If – the initial randomness being a given – we could follow it, we would find that the entropy is conserved.

That is the concept of entropy in a nutshell. We are going to expand on it a lot. We are going to redefine it with a more careful definition.

What does entropy measure? It measures approximately the number of states that have a non-zero probability. The bigger it is, the less we know.

What is the maximum value of $S$? It is $\log N$. Now of course $N$ could be infinite. We might have an infinite num-

[19]Indeed the laws in physics are deterministic. And the equations used to describe them are deterministic too. But the lack of precision in our observations can generate the equivalent of randomness in initial conditions, or in measurements at time $t$. And the lack of precise information might worsen.

ber of states, and if we do then there is no upper bound to the amount of our ignorance. But when studying a closed system with only $N$ states our ignorance is bounded. So the notion of maximum entropy is also a measure of how many states there are altogether.

We said that entropy is deep and fundamental – and so it is. But there is also an aspect to it which makes it in a certain sense less fundamental. It is not just a property of the system. This is a very important point worth stressing:

*Entropy is a property of the system and our state of knowledge of the system.*

It depends on two things. It depends on characteristics of the system, and it also depends on our state of knowledge of the system. The reader should keep that in mind.

Now let's talk about continuous mechanics.

## Continuous mechanics, phase space, and Liouville's theorem

We want to talk about the mechanics of particles moving around with continuous positions, continuous velocities. How do we describe that? How do we describe the space of states of a mechanical system?

We already studied that in volume 1 of the collection *The Theoretical Minimum* on classical mechanics. So this section will begin as a review for the readers who have studied

our volume on classical mechanics. Then we will show the important relationship between Liouville's theorem and entropy.

We describe real mechanical systems, particles and so forth, as points in a phase space[20]. The phase space consists of positions and momenta. In simple contexts, momentum is mass times velocity. So we can also say, roughly speaking, that the phase space is the space of positions and velocities.

That is sufficient when considering a system consisting of particles moving about, with positions $x_i$ and velocities $\dot{x}_i$, like for instance a gas. When considering more complex systems, the degrees of freedom $x_i$ may be more abstract. They are then usually denoted $q_i$. And the generalized momentum conjugate $p_i$ attached to the degree of freedom $q_i$ is defined as $p_i = \partial L / \partial \dot{q}_i$, where $L$ is the Lagrangian of the system, see volume 1 chater 6.

Figure 11 shows the phase space for a system made of a large number of particles. The $i$-th particle[21] has position $x_i$ and momentum $p_i$.

[20] Note on terminology: sometimes the term configuration space refers only to positions, as we did in volume 1; and sometimes it refers to positions and momenta, in which case it is another name for the phase space. Anyway, *phase space* is the usual and unambiguous name for the space of positions and momenta.

[21] Even though we use the same letter, don't mix up the $i$-th particle and the $i$-th state of the system. The $i$-th particle, we will soon forget about it. But the $i$-th state of the system is a very important concept in statistical mechanics.


Figure 11: Phase space for a system of particles.

The vertical axis is for the momenta $p_i$'s. And as usual this axis is a stand-in for *all* of the momentum degrees of freedom, that is all the $p_i$'s. If there are $10^{23}$ particles there are $10^{23}$ p's. And each of them is actually a vector of three values. But we can't draw more than one of them. Well, we could draw two of them but then we wouldn't have any room for the $q$'s, that is for the $x$'s.

Horizontally we record the positions $q_i$'s of the particles. But here we simply denote them $x_i$'s. And again the horizontal axis is a stand-in for the large number of dimensions corresponding to all the particles. So we just write $x$ to mean all the particles positions.

Let's start with the analog of a probability distribution, which is constant, that is equiprobable, on some subset of the total set of states of the system, and is zero on the remaining states.

We can represent that by drawing a sub-region in the phase space, figure 12.

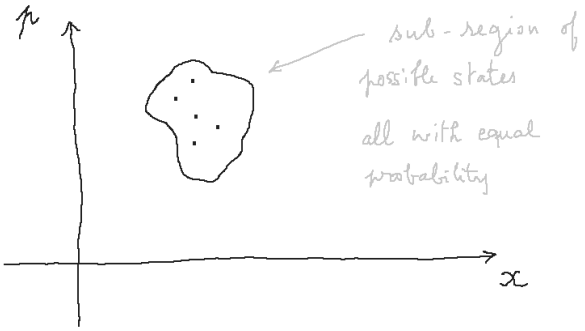


Figure 12: Phase space. Sub-region of possible states. Photograph at time $t$.

In the sub-region shown in figure 12, all the states have equal probability. The time is fixed for the moment. That is, at that time $t$, the system can be in any of the states in the patch, equally probably for each of them. And the probability that the system be outside is zero.

That is the sort of situation where we may have some information about the particles, but we don't know exactly where each of them is and with which velocity. For example we know that all the particles in the room are... in the room. In fact we know quite a bit more than that. We know that they are more or less evenly spread out in the volume of the room. We will formalize it, but let's go one step at a time.

Let's note, to start with, that knowing that the particles are in the room puts some boundaries on where $x$ is in figure

12. Remember that we now call $x$ the entire collection of positions of the particles. $x$ is a unique multi-dimensional variable attached to the state of the system. And for convenience we represent it as a simple number on the horizontal axis.

We may also know that all of the particles have momenta which are within some range. That also confines them between values on the vertical axis.

So a typical bit of knowledge about the particles in the room might be represented, at least approximately, by saying that there is a zero probability that the system be outside the region shown in figure 12, and a uniform probability over all the states within the region.

Now the system evolves over time. As the system evolves $x$ and $p$ change. Over a period $\Delta t$, each point in the sub-region in figure 12 will go to another point elsewhere in the figure, not necessarily in the initial region. Points nearby will go to points nearby. The whole sub-region – also called the *occupied patch* – will flow to another region with a possibly different shape.

The motion of the system with time is almost like a *fluid* flowing in the phase space. We think of the points of the phase space as fluid points moving with time. So the occupied patch flows to another occupied patch, figure 13.

And just like there was equal probability for the system to be at any point in the initial patch in figure 12, after $\Delta t$ there is equal probability for the system to be at any point in the new patch, wherever it is and whatever its new shape

is now.

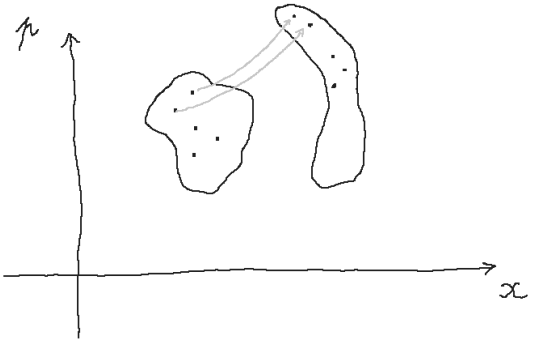


Figure 13: Flowing of the occupied patch in the phase space.

In truth we are in a continuous situation, and probabilities should be replaced by densities of probabilities multiplied by small volumes, but the reasoning is the same.

There is a theorem that goes with this flowing. It is called Liouville's theorem. It says that the volume of the occupied patch is conserved, that is, remains constant over time.

In figures 12 and 13 the patches are surfaces and their volumes are actually areas. But remember that the figures are simplified representations. The $x$ axis can be high-dimensional, and so is the $p$ axis with the same dimension. So the volume of the patch is actually a doubly high-dimensional integral.

Liouville's theorem is the subject of chapter 9 in volume 1 of the collection *The Theoretical Minimum* on classical mechanics. What it tells us is that whatever the initial patch evolves into, it is into something of the same volume, in

other words figuratively speaking the same number of states.

It is the immediate continuous analog of the discrete situation where, if we start with $M$ states and we follow the system according to the equations of motion, we will occupy the same number of states afterwards, as we started with. They will be different states, but the motion will preserve the number of them and the probabilities will remain equal.

**Exercise 1: Explain what is the random experiment, implicitely referred to, when we say that the probabilities of the states in the occupied patch remain the same.**

*Hint: Start with the discrete case corresponding to formulas (22), and then go to the continuous case of figure 13.*

**Exercise 2: Explain why and how these probabilities express our ignorance.**

**Exercise 3: Finally explain how Liouville's theorem expresses the conservation of entropy in the case where we can follow the evolution of the patch.**

Let's stress that Liouville's theorem not only says that the volume of the occupied region will stay the same as it flows, but also, a little bit better, that if we start with a uniform

probability distribution, it will remain uniform.

So there is a very close analogy between the discrete case and the continuous case. And Liouville's theorem is what prevents a system from evolving according to an equation like equation (21) where everything comes to rest. That can't happen.

Let's see more precisely why it can't happen. In figure 13, imagine that no matter where we started, we ended up with $p = 0$. That would mean every point of the occupied patch got mapped onto the $x$ axis. Its area would go from some positive value to zero. But Liouville's theorem prevents that.

What Liouville's result says in fact is if the blob squeezes in one direction, it must expand in another direction, like a drop of water trapped beneath a plastic film coating a surface, which we can move around but not remove.

Let's think of Liouville's theorem in the context of an eraser sliding on a table top, slowing down and coming to rest. So we consider that *we don't know exactly* the state of the system, and therefore it is in a patch like in figures 12 and 13. If, in the phase space, the components of the eraser get shrunk because it stops, it means the components of some other elements of the system [ eraser + table ] in the phase space must expand. It is the $p$'s and $x$'s of the molecules that are in the table.

For the case of the eraser sliding on the table, there is really a very high dimensional phase space. And as the eraser may come to rest, or almost rest, so that the phase space

squeezes one way, it spreads out in another direction having to do with the other hidden microscopic degrees of freedom.

Our possible partial ignorance about the exact position and velocitiy of the eraser cannot vanish as it stops. At the least, it is transferred as an increased partial ignorance about the exact positions and velocities of the molecules of the table top. So if we can – as an omnipotent observer – keep track of the system the entropy is conserved.

As the reader may be aware of, we will see that it can also go up. That will be the subject of the second law of thermodynamics.

We now know how the $-1^{\text{st}}$ law of physics, about conservation of information, is expressed in the context of statistical mechanics and imperfect knowledge of all the degrees of freedom describing a system.

To follow the order of the laws of thermodynamics, let's talk briefly about the zeroth law. Then we will talk about the first law, and then about the famous second law of thermodynamics.

### Zeroth law of thermodynamics

The zeroth or 0-th law of thermodynamics has to do thermal equilibrium. We haven't explained what a thermal equilibrium is, but we can already give an intuitive idea, and give a preliminary statement of the zeroth law.

A gas in a vase is in thermal equilibrium if all its molecules moves around in such a way that the global distribution of their positions and velocities somehow doesn't change over time. And the concept can be applied to any system.

Similarly, we say that two systems $A$ and $B$ are in thermal equilibrium with each other if, when put in contact so that they can exchange energy, the global distributions of the positions and velocities of the molecules in each of them don't change.

The zeroth law of thermodynamics says that if we have several systems, and system $A$ is in thermal equilibrium with $B$, and $B$ is in thermal equilibrium with $C$, then $A$ is in thermal equilibrium with $C$.

We will come back to that. Let's just keep it in the back of our mind for the time being, because we haven't described what thermal equilibrium precisely is. But having gone through the $-1^{\text{st}}$ and the zeroth laws, we can now jump to the first law.

### First law of thermodynamics

The first law is simply energy conservation, nothing more. It is really simple to write down. But its simplicity belies its power.

It is the statement that, first of all, there is a conserved quantity. And the fact that we call that conserved quantity the energy will play for the moment not such a big role.

But let's just say there is energy conservation.

How is that expressed? As we can guess, whatever the energy $E$ is, the equation is written

$$\frac{dE}{dt} = 0 \tag{24}$$

Now this is the law of energy conservation for a closed system.

If a system consists of several parts in interaction with each other, then of course any one of the parts can have a changing energy. But the sum total of all of the parts will conserve energy.

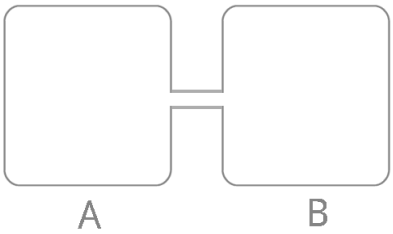


Figure 14: Two systems forming altogether one closed system.

For instance, in the system represented in figure 14 we have

$$\frac{dE_A}{dt} = -\frac{dE_B}{dt} \tag{25}$$

We could have written that the sum of the two derivatives is equal to zero, but equation (25) emphazises that what energy we lose on one side we gain on the other.

That is the first law of thermodynamics. That is all it says. The total energy of a closed system is conserved.

Now in the context of figure 14 there is a slightly hidden assumption. We have assumed that, if the system is composed of two parts, its total energy is the sum of the energies of each parts. That is really not generally true.

If you have two systems and they interact with each other, there maybe for example forces between the two parts. So there might be a potential energy that is a function of both of the coordinates.

For example, let's look at a closed system made of two particles orbiting around each other, or, say, a simplified Solar System made of the Sun and the Earth. The energy consists of the kinetic energy of one body plus the kinetic energy of the other body plus a term which doesn't belong to either body. It belongs to both of them in a sense. It is the potential energy of interaction between the Sun and the Earth.

In that context we really can't say that the energy is the sum of the energy of one thing plus the energy of the other thing. *Energy conservation is still true*, but we can't divide the system into two parts as in figure 14.

On the other hand, there are many contexts where the interaction energies between subsystems is negligible compared to the energy that the subsystems themselves have.

If we were, for instance, to divide the table in front of us up into blocks. How much energy is in each block? Well, the amount of energy that is in each block is more or less

proportional to the volume of each block.

How much energy of interaction is there between the blocks? The energies of interactions are surface effects. They interact with each other because their surfaces touch. And typically the surface area is small by comparison with volume.

We will come back to that. But let's keep in mind that in many contexts, the energy of interaction between two systems is negligible compared to the energy of either of them.

When that happens we can say to a good approximation that the energy of a system made of two parts can just be represented as the sum of the energies of each part, plus a teeny little thing which has to do with their interactions and which we omit. In those circumstances, the first law of thermodynamics is often expressed with equation (25).

In summary, equation (24) is always true, while equation (25) has that little caveat that we are talking about systems where energy is strictly additive.

Now let's return to entropy.

## More on entropy

We are not finished with entropy. We have introduced it. We have also talked about energy. Notice that we haven't talked about temperature yet. Temperature comes in behind entropy.

Indeed temperature is a highly derived quantity. By that we mean that, despite the fact that it is the characteristic of a system we can readily feel with our body, so it is very intuitive and appears fundamental, it is actually a concept derived mathematically from the motion and kinetic energy of molecules, less primitive and less fundamental than either energy or entropy. It will be the subject of chapter 2. For the moment we shall go deeper into entropy.

We defined entropy. But we did it only for certain special probability distributions. Let's represent it schematically, figure 15.

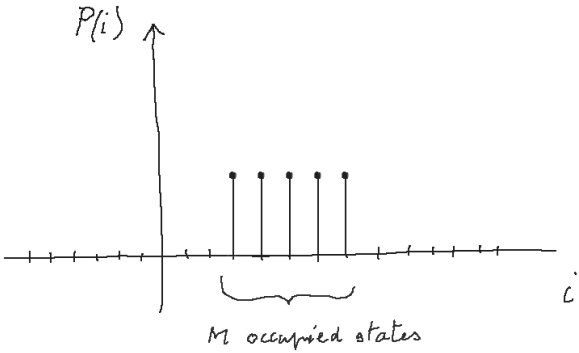


Figure 15: Equiprobable distribution over the set of occupied states.

On the horizontal axis are laid out the different states labelled by their index $i$. And vertically we plot their probability. The probability distributions we considered are those such that inside a subset of occupied states all the probabilities have the same positive value, and outside the probabilities are zero. If there are $M$ occupied states, the probability of each of them is $1/M$.

Then we defined the entropy as

$$S = \log M \tag{26}$$

That is, in the case where the distribution of probability is uniform over all the occupied states, the entropy is simply the logarithm of their number.

Generally speaking, however, we don't have probabilities evenly distributed like in figure 15. We have probability distributions which are more complicated. In fact they can be anything as long as they are positive and all add up to 1. Let's draw another one, and for simplicity, even though we are in a discrete case, let's draw it as a continuous curve.

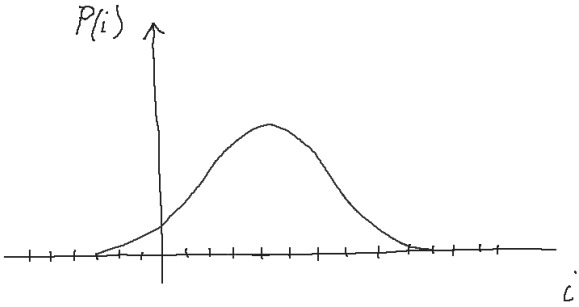


Figure 16: General distribution of probabilities over the states.

So the question is: How do we define entropy in a more general context where the probability distribution looks something like in figure 16 or has an even more odd shape?

In this lesson, we won't construct logically the formula. I will just give the definition. We will check that, in the

equiprobable case, it corresponds to equation (26). Then our objective will be to get familiar with it and begin to see why it is a good definition.

It is representing something about the probability distribution. It is in some sense the average number of states which are importantly contained inside the probability distribution in figure 16.

The narrower the probability distribution, the smaller the entropy will be. The broader the probability distribution, the bigger the entropy will be. In other words, a spread out distribution will correspond to a high entropy, while a narrow distribution over a small number of states will correspond to a small entropy.

The formula for the entropy $S$ attached to a probability distribution of $P(i)$'s is

$$S = -\sum_i \; P(i) \; \log P(i) \tag{27}$$

Remembering formula (14) for the average of a function $F(i)$, we see that formula (22) can actually be read as the average of the function $\log P(i)$. It is worth emphasizing:

*The entropy of a distribution $P$ is the average of* $\log P$.

Let's see what this yields in the special case where the probability distribution is $1/M$, over $M$ states.

For the unoccupied states the term $P(i) \log P(i)$ looks like it is going to cause a problem. It is zero times logarithm of zero. Of course logarithm of zero is not defined. The

limit when $P$ tends to 0 of $\log P$ is minus infinity. But the logarithm grows, in the negative numbers, to infinity much slower than $P$ goes to zero. So $P \log P$ goes to 0. It is left as a little exercise for the reader to show that. So in general the contribution from states with very small probability will be very small, and by continuity the contribution of unoccupied states will be zero.

Now what about the states which have positive probability in the equiprobable case? The probability of each one is $1/M$. So formula (27) becomes

$$S = -\sum_{i=1}^{M} \frac{1}{M} \log\left(\frac{1}{M}\right)$$

Then, noting that $\log(1/M) = -\log M$, summing the $M$ terms which are all the same, thus getting rid of the factor $1/M$, and cancelling the minus signs, we get

$$S = \log M$$

which is formula (26) again, as we want.

Notice that the minus sign in the general definition (27) for the entropy is there simply because $P(i)$ is always less than one, therefore all the $\log P(i)$ are negative. If we want the entropy to be a positive number, we have to introduce a minus sign in its definition.

Our new definition now makes sense even when we have a more complicated probability distribution than just uniform over the occupied states. We will discover, as we use

it, that this average of $\log P$ is a good and effective definition.

Notice that entropy is associated with a probabilitiy distribution. *It is not something like energy which is a property of the system itself.* It is not a thing like momentum. Entropy is associated with our imperfect knowledge of which state the system we are considering is in.

This idea requires some getting used to. When we consider a system whose state is perfectly known, there is no concept of entropy. Or rather the entropy is zero. The reader should check that it is consistent with formula (27).

When we consider a gas made of a very large number of molecules in a box, and it is in thermal equilibrium – assuming for the moment we intuitively know what that means –, at the same time we feel that we know well in which state it is, described by volume, pressure, temperature, etc. as classical thermodynamics tells us, but we must also admit that it is only a *statistical knowledge* over a large number of possible exact configurations of positions and velocities of the molecules. And that is so even without taking into consideration the further limits imposed, at a much higher level of precision, by quantum mechanics on the simultaneous knowledge of positions and momenta.

So entropy is a special kind of measure. It is a measure of our only statistical knowledge. It is not a pure measure attached to the system itself. It involves and reflects the incomplete knowledge we have of it.

We must develop a familiarity with this probability distri-

bution representing our imperfect knowledge of the system. Once in a while the reader should go back to the fundamental questions: What is the random experiment that is implicitely referred to when talking about the $P(i)$'s? What are the states $i$'s? In what sense do we have only an imperfect or incomplete knowledge of the system? What are the $P(i)$'s? Etc.

The beauty of statistical mechanics, or statistical thermodynamics, over classical thermodynamics, is that, once we have introduced the fundamental probabilistic description of the system, reflecting our imperfect knowledge of which state it is in, it offers new and stunning explanations of what is entropy and temperature, and their relations to energy and pressure. We will discover little by little this beauty as we progress in our study. The present chapter is devoted to the probabilistic framework and to the introduction of entropy. Temperature will be treated in chapter 2, pressure in chapter 5. And we will study many other subjects in statistical mechanics.

Hopefully the reader now understands why entropy is a somewhat more obscure quantity from the point of view of intuitive definition, than the other thermodynamics entities. It is because its definition has to do with both the system and our state of knowledge of the system.

Let's now do some examples.

### Examples of entropy of simple systems

Let's calculate some entropy for a couple of simple systems. Our first system will be with coins, not a single coin but a lot of coins. We consider $n$ coins, figure 17.

Figure 17: Collection of $n$ coins, each showing heads or tails.

Each coin can be heads or tails. Suppose that we have no idea what the state of the system is. We know nothing. The probability distribution, in other words, is the same for all states of this system consisting in $n$ coins. We are considering the case of absolute ignorance on our part.

What is the entropy associated to the system and us[22] in such a situation?

All the probabilities are equal. Under this circumstance, we just get to use logarithm of the number of possible states. How many states are there altogether? We have

$$N = 2^n \tag{28}$$

[22]Soon we will stop recalling that the entropy is related to us and our knowledge of the system. It will be implicit, and we will drop the "and us".

There are two states for the first coin, two states for the second coin, etc. That makes $2^n$ possible states for the collection of $n$ coins.

What is the entropy, given that we know nothing? It is the logarithm of the number of possible states, therefore we have

$$S = n \log 2 \tag{29}$$

Here we see an example of the fact that entropy is kind of additive over the system. It is proportional to the number of degrees of freedom in this case.

And we also discover a *unit of entropy*. The unit of entropy is called a bit. It is the abbreviation of binary digit. That is what a bit is in information theory. It is the basic unit of entropy for a system which has only two states, up or down, heads or tails, open or closed, or whatever.

The entropy is proportional to the number of bits, or in this case the number of coins times the logarithm of 2. So $\log 2$ plays a fundamental role in information theory as the unit of entropy.

It does not mean that in general that entropy is an integer multiple of $\log 2$. We will see it in a second.

Above our state of knowledge was zilch. We knew nothing. The value $n \log 2$ has an important interpretation. It is the *maximum entropy*.

With the same collection of $n$ coins, let's try another state

of knowledge. Suppose, at the other extremity of the range of possible knowledge, that we know the state completely. That is the case where the number of occupied states is $M = 1$. The probability distribution now is shown in figure 18.

The entropy is now $S = 1 \log(1)$, and that is equal to 0. So complete knowledge corresponds to zero entropy. And, generally speaking, the more you know about the system, the smaller the entropy is.

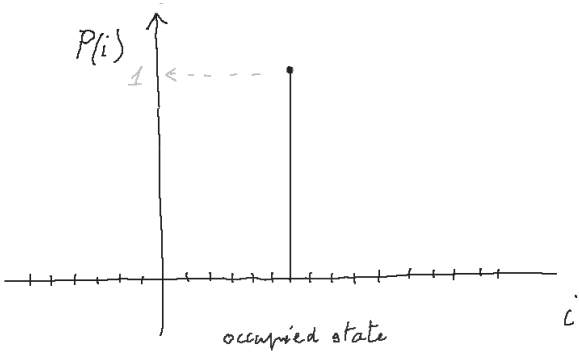


Figure 18: Case of perfect knowledge of the state of the system.

With our collection of $n$ coins, lets consider another interesting case. We know that all of them are heads, except one that is tails. But we don't know which one. And we have no clue about where it could possibly be. So our imperfect knowledge of the system corresponds to an even probability distribution over all the states with one tails and $(n-1)$ heads. How many are there? Well, clearly there are $n$ such states. Tails could be the first coin, or it could be the second coin, etc. In other words, $M = n$. Therefore, according

to formula (27) or its simpler version (26), in this situation the entropy is

$$S = \log n \tag{30}$$

Notice that it doesn't have to be an integer multiple of $\log 2$. In general entropy is not an integer multiple of $\log 2$. Nevertheless $\log 2$ is a good unit of entropy. And it is called the bit.

If the coins of our collection were objects with three possible states[23], formula (28) would become $N = 3^n$. The maximum entropy would be $n \log 3$. And the example where we know that $(n-1)$ objects are in one state, and one of them is in another state, and we don't know which one, would have to be entirely reworked.

**Exercise 4: Suppose we have a collection of $n$ identical objects, each of them having three possible states. Build various cases of imperfect knowledge of the state of the collection, and compute each time the entropy.**

Now computer scientists of course, for a variety of reasons, like to think in terms of 2. First of all the mathematics of it is nice. 2 is the smallest integer which is not 1.

Moreover, the physics that goes on inside your computer is connected with electric switches which are either on or off.

[23] An example is a die where we consider the three states $\omega_1 = \{1 \text{ or } 2\}$, $\omega_2 = \{3 \text{ or } 4\}$ and $\omega_3 = \{5 \text{ or } 6\}$.

So it is easy to have them manipulate information expressed in bits[24].

### Questions / answers session (2)

Q.: Is the logarithm in the definition of the entropy calculated in base 2 or in base $e$?

A.: Basically it doesn't matter because there is just a simple multiplicative factor between the two.

Remember that we have much freedom in the choice of our units. The speed of light can be expressed in meters, or just taken to be 1. The same is true with units of length, mass or time. We have the MKS units system, the Planck units, etc. And the same is true with the bit.

The usual choice depends on who you are. If you are a computer scientist, or an information theorist disciple of Shannon[25] then you like logarithm to the base 2. In this case $\log 2$ is just 1. And the entropy, for instance in the example leading to formula (29), it is just $n$. That is, *the*

[24]Notice that we are talking about conventional computers, not quantum computers which are working with qbits. A fundamental practical discovery in electronics, which made the digital electronic computers possible, was made by W. Eccles and F. W. Jordan, in 1918, when they figured out how to build a circuit capable of staying in one of two states, after having received some electrical input, without the help of any mechanical device.

[25]Claude Shannon (1916 - 2001), American mathematician, electrical engineer, and information theorist.

*maximum entropy of a collection of n coins is* $n$[26].

If you are a physicist, then you usually work in base $e$. But the relationship is just a multiplicative factor. Logarithm to the base $e$ and logarithm to the base 2 are just related by a numerical factor that is always the same[27].

In this book, when we write log we mean logarithm to the base $e$. But very little would change if we used some other base for the logarithms. For us physicists the maximum entropy of the collection of $n$ coins is aproximately 0.7 x $n$. But anyway we just write it $n \log 2$.

### Entropy in phase space

We are back in a continuous space of configurations of the possible states of the system, and the phase space, which combines positions and momenta, is continuous too. Let's see what is the definition of entropy in this case.

We begin by supposing that the probability distribution is just uniform over some blob formed by the occupied states, and zero outside the blob. In other words we are in a simple situation, figure 19.

[26] Notice how this sentence can be puzzling or confusing if we don't know all that we have learned about entropy in this lesson: the probabilistic framework, the possible states of the system, the fact that entropy is a measure characterizing the system *and* our knowledge of it, etc. That is the difference between our courses and articles in scientific reviews for the general public.

[27] Logarithm to the base $e$ of 2 = 0.69314718056...

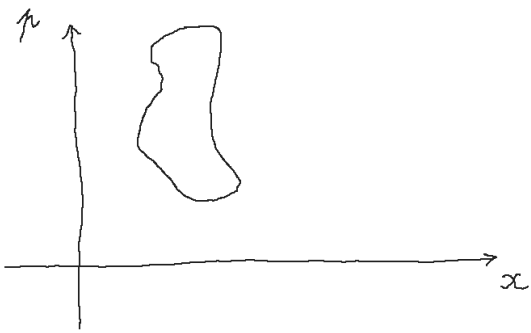


Figure 19: Region, or blob, in the phase space formed by the occupied states.

Then the definition of the entropy should be simple: we take the logarithm of the number of states in the blob – except that we are in the continuous case and the number of occupied states is infinite.

So instead we just say it is the logarithm of the volume of the probability distribution in phase space.

$$S = \log V_{PS} \tag{31}$$

$V_{PS}$ is the volume in the phase space, which is high-dimensional. Whatever the dimensionality of the phase space is, the volume is measured in units of position times units of momentum to the power of the number of coordinates in the system. It is the volume of the region which is occupied and has a nonzero probability distribution.

This is the closest analog that we can think of to $\log M$ where $M$ represented the number of equally probable states

of a discrete system. Here we didn't have to weigh the states with their probability or density of probability since we are in a uniform case.

More generally, if the distribution of probability over the blob in figure 19 is not uniform, we will introduce in formula (31) the *density of probability* $P(p, x)$ that the system be at point $(p, x)$ in the phase space.

Remember that, because we are in a continuous case, we cannot deal with discrete probabilities and their ordinary sum. The usual formula

$$\sum_i P(i) = 1 \tag{32}$$

becomes now the integral of a probability density over the phase space

$$\int P(p, x)\ dp\ dx = 1 \tag{33}$$

The density $P(p,\ x)$ is non zero only over the blob of occupied states but that doesn't make any difference in formula (33).

Now, in the continuous case with non uniform density of probability, the formula for the entropy is

$$S = -\int P(p, x)\ \log P(p, x)\ dp\ dx \tag{34}$$

In first approximation – we don't mean first approximation numerically, but first conceptual approximation –, it is

measuring the logarithm of the volume of the probability blob in phase space. That was the meaning of formula (31) which we now extended to formula (34).

We have now defined, in the general continuous case, what is entropy, which as we have seen depends on the probability distribution, or density, over the blob of occupied states.

The next topics are temperature and the Boltzmann distribution. Temperature will be treated in chapter 2, maximizing entropy in chapter 3, and the Boltzmann distribution in chapter 4.

The Boltzmann distribution is the probability distribution for thermal equilibrium. We haven't quite defined thermal equilibrium yet but we will.

Let's finish this substantial first chapter in statistical mechanics with a last questions / answers session.

### Questions / answers session (3)

Q.: In the example with $n$ coins, do you assume that the tosses of each coin are independent?

A.: Well, I haven't made such an assumption. I did not describe the experiment, or sequence of experiments, which produced the series of $n$ elementary results, one for each coin. Of course what you say could be the case, but the point is that the assumption in not necessary as such, and in fact would not always be true. I just described what we

know.

In the first example we know nothing, and the $2^n$ possible series of tosses are considered to be equally probable. Then it is indeed equivalent to your assumption.

In the second case, we know that $(n-1)$ tosses are heads, and one is tails, and we don't know which one. Now your assumption doesn't hold.

We just say what we know, without specifying how we got to our incomplete knowledge. And our incomplete knowledge we translate into a probability distribution over occupied states.

To say, like in the first example, that all states are equally probable is closely related to saying that there are no correlations.

Let's go back briefly to elementary probability theory. Consider two random variables $X$ and $Y$, on the state space $\Omega$. For instance the experiment $\mathcal{E}$ could be the tosses of two coins, coin 1 and coin 2. $X$ would be the result of the first toss, heads or tails. And $Y$ would be the result of the second toss. The space of states here is

$$\Omega = \{\ (H,H),\ (H,T),\ (T,H),\ (T,T)\ \} \qquad (35)$$

The r.v.[28] $X$ is a function from $\Omega$ to the set $\{H,\ T\}$. It consists in picking the first value in the pairs forming $\Omega$. $X$ has a probability distribution, which can be calculated from the fundamental probabilities for the four elements of

[28] Abbreviation for random variable.

$\Omega$. They don't have to be 1/4 each. They could be, for instance, $P(\omega_1) = 1/2$, $P(\omega_2) = 1/4$, $P(\omega_3) = 1/4$, and $P(\omega_4) = 0$.

Then we can calculate the distribution of probability attached to $X$:

$$\begin{aligned} P\{X = H\} &= \frac{3}{4} \\ P\{X = T\} &= \frac{1}{4} \end{aligned} \tag{36}$$

This is denoted $P_X$, the probability distribution of the random variable $X$. We can do the same for $Y$.

We can also define and compute a joint probability distribution for $(X, Y)$. It is denoted $P_{XY}$. For instance $P_{XY}(H, H) = 1/2$, $P_{XY}(H, T) = 1/4$, etc.

$X$ and $Y$ are said to be independent if and only if their joint distribution is the product of each of their distributions. In other words, iff

$$P_{XY} = P_X \; P_Y \tag{37}$$

then, by definition, $X$ and $Y$ are independent.

**Exercise 5: Show that, in the above example, $X$ and $Y$ are not independent.**

Another way to define, and to feel, independence is as follows: $X$ and $Y$ are independent if knowing the value of $X$,

after having performed $\mathcal{E}$, doesn't give any information on $Y$ – technically, if the conditional distribution of $Y$, given a specific value of $X$, is the same as its marginal distribution (another name for $P_Y$).

**Exercise 6: In the above example, what is the conditional distribution of $Y$ if we know that $X = T$?**

Finally correlation is a probabilistic concept close to dependence. However uncorrelation is less stringent than independence. Strictly speaking correlation is defined only for numerical random variables. We say that two random variables are uncorrelated iff the expectation of their product is the product of their expectations[29], that is iff

$$E(XY) = E(X)\ E(Y) \tag{38}$$

Correlation is defined as follows

$$Corr(XY) = \frac{E(XY) - E(X)\ E(Y)}{\sigma_X\ \sigma_Y} \tag{39}$$

where $\sigma_X$ is the standard deviation of $X$, and $\sigma_Y$ is the standard deviation of $Y$.

Why introduce correlation and noncorrelation if dependence and independence is approximately the same concept? Because it is simpler to compute or check.

[29] We use in formulas (38) and (39) the standard notations of probabilists.

It raises a little problem however: the two concepts are equivalent within the family of Gaussian random variables. But they are not equivalent in general. It is possible to construct counterexamples, where $X$ and $Y$ are not correlated but are not independent. These counterexamples usually play no role in statistical mechanics, or even more generally in physics, where, thanks to the central limit theorem, many random variables we deal with are Gaussian, and when they are not, well, we go into the details.

Let's go back to our series of $n$ coins. In the case where at first we know nothing, then knowing that coin 1 is heads, doesn't give any information on the others. They are indeed uncorrelated and independent.

Now let's look at the case where what we know is: $(n-1)$ are heads, and one is tails, but we don't know which one. If we look at the first coin and it is heads, does it give us any information on the second one? Well, not much, but actually a little bit. Instead of having probability $1/n$ of being tails, it is now $1/(n-1)$. So the coin tosses cannot have been independent. And in fact if the first coin is tails, then the second coin is surely heads, etc.

We gave a precise definition for correlation. For practical purposes, with the above little caveat, what is important to remember is this:

*There is correlation between variables if measuring a first variable gives us information on the others.*

Or, said another way, there is correlation when the probability distribution for the other things is modified by mea-

suring the first.

In the complete ignorance case, there is no correlation. In any other kind of initial incomplete knowledge about the state of the system, in general there is some correlation.

Q.: In the system made of two parts $A$ and $B$, in figure 14, once we measured one thing in $A$, which is a conserved quantity in the overall system, that gives us information on what it is in $B$, doesn't it?

A.: Oh yes. That is certainly true for energy.

Concerning entropy we have to be much more careful.

Incidentally, entropy is additive. It is the sum of the entropies of all the individuals parts. If your system is made of many identical parts, the overall entropy is proportional to the number of parts.

Entropy is additive whenever there is no correlation. But this fact is more subtle than it looks, because entropy is not as simple a concept as energy. The concept of entropy is related to our lack, or partial lack, of information. So we will have to continue to develop our understanding of it.

In summary, uncorrelated systems have additive entropies. But that is a theme that we will come back to.

Here is an interesting question, that we submit to the reader:

**Exercise 7: We have a collection of $n$ coins laid out in a row. If we measure one of them and it is up, then the probability for its neighbor on the right is three quarters to be down. And it is also the probability for its neighbor on the left to be down. That is given. And that is all we know.**

**Calculate the entropy of such a distribution.**

There is correlation of course, because when we measure one coin we immediately know something about its neighbors.

We encourage you to make up your own example like that, and compute the entropy. You will learned something from it.

Q.: Was this idea of entropy of a series of coins invented in 1949 with information theory?

A.: Shannon rediscovered, in the context of information theory that he was developping, the idea of entropy. And he defined the entropy of a message made of a series of bits – that is the same as our collection of $n$ coins.

However the formula

$$S = -\int P \log P \tag{40}$$

is due to Boltzmann.

What is the formula on his tomb? (figure 20)

$$S = k \log W \tag{41}$$

But he meant $S = -\int P \log P$ :-)

He did write the formula $S = -\int P \log P$. It is Boltzmann final formula for entropy. And the only difference with Shannon entropy is that Shannon uses $\log 2$.

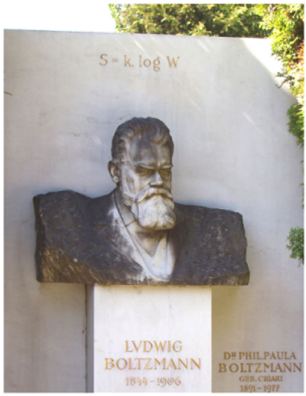


Figure 20: Boltzmann tombstone.

Now of course Shannon discovered this entirely by himself. He didn't know Boltzmann's work. And he worked from an entirely different direction, from information theory rather than from thermodynamics.

But none of it would have surprised Boltzmann. Nor do I think Boltzmann definition would have surprised Shannon.

They are really the same thing.

If you put the minus sign in $S = -\int P \log P$, it is called *entropy*.

If you don't put the minus sign, then you write $I = \int P \log P$ and it is called *information*.

Q.: Is the incompleteness of our knowledge about the state of the system, in statistical mechanics, related to the Heisenberg uncertainty principle in quantum mechanics?

A.: No. These are two separate issues. $S$ doesn't have to do with the quantum mechanical uncertainty which we encounter when measuring an observable, if the state of the system is not one of the eigenstates of the observable.

In fact $S$ has to do with the uncertainty implicit in *mixed states*, if you remember chapter 7 of volume 2 in the collection *The Theoretical Minimum*, on quantum mechanics. That chapter is devoted to entanglement. It explains in particular the so-called density matrices, which do represent an incomplete knowledge like in statistical mechanics. But it has nothing to do with Heisenberg's principle, which is related to observables which cannot be measured simultaneously because their Hermitian operators don't commute.

In other words, $S$ has nothing to do with the randomness of measures implicit in *pure states*.

Q.: Why isn't there the Boltzmann constant $k_B$ in the definition of entropy in formula (40)?

A.: Well, remember that there are conversion factors between units, so that in the appropriate units

$$c = \bar{h} = G = k_B = 1$$

Boltzmann's constant was a conversion factor from temperature to energy. The natural unit for temperature is really energy.

But the energy of a molecule for example is approximately equal to its temperature in certain units. Those units contain a conversion factor $k_B$.

## Appendix: customary mistakes in elementary probabilities

When talking about probabilities, even educated people may say wrong things like: since having two accidents is very rare, and since I already had one, now the chances that I have another one are much smaller.

On a more sophisticated level, the paradox of Monty Hall, for instance, can befuddle even the best minds. Here is how it goes: there are two people, one is the guesser who must provide a best guess, the other is the operator of the game. There are three closed doors, $A$, $B$ and $C$, facing the guesser. Behind one and only one is a prize. Step one: the guesser must make a guess, for instance guess door $A$. Step two: the operator, who knows where the prize is, doesn't open $A$, but selects among $B$ and $C$ a door where the prize

is not, and opens it. The guesser sees this new piece of information. Step three: the guesser is invited to guess again where the prize is.

Question: Should the guesser change his or her guess, and now choose the other non opened door, or it doesn't matter? Answer: it *does* matter. The guesser should change guess, and now choose the other non opened door. The probability of winning will go from 1/3 to 2/3. Before reading the solution, try to solve it by yourself.

One way to see it is to note that, if the guesser follows this strategy, he or she will lose only when the prize was behind the initial guess.