**PROFESSOR:** So we started by talking about thermodynamics. And then switched off to talking about probability. And you may well ask, what's the connection between these? And we will eventually try to build that connection through statistical physics. And maybe this lecture today will sort of provide you with why these elements of probability are important and essential to making this bridge.

So last time, I started with talking about the Central Limit Theorem which pertains to adding lots of variables together to form a sum. And the control parameter that we will use is this number of terms in the sum.

So in principle, there's a joint PDF that determines how these variables are distributed. And using that, we can calculate various characteristics of this sum. If I were to raise the sum to some power m, I could do that by doing a sum over i running from let's say i1 running from 1 to N, i2 running from-- im running from 1 to N, so basically speaking this sum. And then I have x of i1, x of i2, x of im. So basically I multiplied m copies of the original sum together.

And if I were to calculate some moment of this, basically the moment of a sum is the sum of the moments. I could do this. Now the last thing that we did last time was to look at some characteristic function for the sum related to the characteristic function of this joint probability distribution, and conclude that actually exactly the same relation holds if I were to put index c for a cumulant.

And that is basically, say the mean is the sum of the means, the variance is sum of all possible variances and covariances. And this holds to all orders. OK? Fine. So where do we go from here? We are going to gradually simplify the problem in order to get some final result that we want. But that result eventually is a little bit more

general than the simplification.

The first simplification that we do is to look at independent variables. And what happened when we had the independent variables was that the probability distribution could be written as the product of probability distributions pertaining to different ones. I would have a p1 acting on x1, a p2 acting on x2, a pn acting on the xn.

Now, when we did that, we saw that actually one of the conditions that would then follow from this if we were to Fourier transform and then try to expand in powers of k, is we would never get in the expansion of the log terms that were coupling different k's. Essentially all of the joint cumulants involving things other than one variable by itself would vanish. So essentially in that limit, the only terms in this that would survive we're the ones in which all of the indices were the same.

So basically in that case, I would write this as a sum i running from one to N, xi to the power of N. So basically for independent variables, let's say, the variance is the sum of the variances, the third cumulant is the sum of the third cumulants, et cetera.

One more simplification. Again not necessary for the final thing that we want to have in mind. But let's just assume that all of these are identically distributed. By that I mean that this is basically the same probability that I would use for each one of them. So this I could write as a product over i one to N, the same p for each xi.

Just to make sure you sum notation that you may see every now and then, variables that are independent and identically distributed are sometimes called IID's. And if I focus my attention to these IID's, then all of these things are clearly the same thing. And the answer would be simply N times the cumulant that I would have for one of them.

This-- actually some version of this, we already saw for the binomial distribution in which the same coin, let's say, was thrown N independent times. And all of the cumulants for the sum of the number of heads, let's say, were related to the cumulants in one trial that you would get. OK? So fine. Nothing so far here.

However let's imagine now that I construct a variable that I will call y, which is the variable x, this sum that I have. From it I subtract N times the mean, and then I divide by square root of N. I can certainly choose to do so. Then what we observe here is that the average of y by this construction is 0. Because essentially, I make sure that the average of x is subtracted.

No problem. Average of y squared-- not average of y squared, but the variance. Surely it's easy to show the variance doesn't really depend on the subtraction. It is the same thing as the variance of x. So it is going to be essentially x squared c divided by square of this. So I will have N. And x squared, big x squared cumulant, according to this rule, is N times small x squared cumulant. And I get something like this. Still nothing interesting.

But now let's look at the m-th cumulant. So let's look at y m c for m that is greater than 2. And then what do I get? I will get to N times x m c divided by N to the m over 2. The N to the power of m over 2 just came from raising this to the power of m, since I'm looking at y to the m. And x to the m c, according to this, is N times x1.

Now we see that this is something that is proportional to the N to the power of 1 minus m over 2. And since I chose m to be greater than 2, in the limit that N becomes much, much larger than 1, this goes to 0. So if I look at the limit where the number of terms in the sum is much larger than 1, what I conclude is that the probability distribution for this variable that I have constructed has 0 mean, a finite variance, and all the other higher order cumulants are asymptotically vanishing.

So I know that the probability of y, which is this variable that I have given you up there, is given by the one distribution that we know is completely characterized by its first and second cumulant, which is the Gaussian. So it is exponential of minus y squared, two times its variance divided, appropriately normalized.

Essentially this sum is Gaussian distributed. And this result is true for things that are not IID's so long as this sum i1 to im, one to N, xi1 to xim goes as N goes to infinity, much, much less than 1, as long as it is less than-- less than strictly than N to the m over 2.

So basically, what I want to do is to ensure that when I construct the analog of this, I would have something that when I divide by N to the m over 2, I will asymptotically go to 0. So in the case of IID's, the numerator goes like N, it could be that I have correlations among the variables et cetera, so that there are other terms in the sum because of the correlations as long as the sum total of them asymptotically grows less than N to the m over 2, this statement that the sum is Gaussian distributed it is going to be valid. Yes.

**AUDIENCE:** Question-- how can you compare a value of [INAUDIBLE] with number of variables that you [INAUDIBLE]? Because this is a-- just, if, say, your random value is set [? in advance-- ?]

**PROFESSOR:** So basically, you choose a probability distribution-- at least in this case, it is obvious. In this case, basically what we want to know is that there is a probability distribution for individual variables. And I repeat it many, many times. So it is like the coin. So for the coin I will ensure that I will throw it hundreds of times. Now suppose that for some reason, if I throw the coin once, the next five times it is much more likely to be the same thing that I had before. Kind of some strange coin, or whatever.

Then there is some correlation up to five. So when I'm calculating things up to five, there all kinds of results over here. But as long as that's five is independent of the length of the sequence, if I throw things 1,000 times, still only groups of five that are correlated, then this result still holds. Because I have the additional parameter N to play with. So I want to have a parameter N to play with to go to infinity which is independent of what characterizes the distribution of my variable.

**AUDIENCE:** I was mainly concerned with the fact that you compare the cumulant which has the same dimension as your random variable. So if my random variable is-- I measure length or something. I do it many, many times length is measured in meters, and you try to compare it to a number of measurements. So, shouldn't there be some dimensionful constant on the right?

**PROFESSOR:** So here, this quantity has dimensions of meter to m-th power, this quantity has

dimensions of meter to the m-th power. This quantity is dimensionless. Right? So what I want is the N dependence to be such that when I go to large N, it goes to 0. It is true that this is still multiplying something that has-- so it is.

**AUDIENCE:** It's like less than something of order of N to m/2? OK.

**PROFESSOR:** Oh this is what you-- order. Thank you.

**AUDIENCE:** The last time [INAUDIBLE] cumulant [INAUDIBLE]?

**PROFESSOR:** Yes, thank you. Any other correction, clarification? OK. So again but we will see that essentially in statistical physics, we will have, always, to deal with some analog of this N, like the part number of molecules of gas in this room, et cetera, that enables us to use something like this.

I mean, it is clear that in this case, I chose to subtract the mean and divide by N to the 1/2. But suppose I didn't have the division by N to the 1/2. Then what happens is that I could have divided for example by N. Then my distribution for something that has a well-defined, independent mean would have gone to something like a delta function in the limit of N going to infinity. But I kind of sort of change my scale by dividing by N to the 1/2 rather than N to sort of emphasize that the scale of fluctuations is of the order of square root of N.

This is again something that generically happens. So let's say, we know the energy of the gas in this room to be proportional to volume or whatever. The amount of uncertainty that we have will be of the order of square root of volume.

So it's clear that we are kind of building results that have to do with dependencies on N. So let's sort of look at some other things that happen when we are dealing with large number of degrees of freedom. So already we've spoken about things that intensive, variables such as temperature, pressure, et cetera. And their characteristic is that if we express them in terms of, say, the number of constituents, they are independent of that number.

As opposed to extensive quantities, such as the energy or the volume, et cetera,

that are proportional to this. We can certainly imagine things that would increase [INAUDIBLE] the polynomial, order of N to some power. If I have N molecules of gas, and I ask how many pairs of interactions I have, you would say it's N, N minus 1 over 2, for example. That would be something like this.

But most importantly, when we deal with statistical physics, we will encounter quantities that have exponential dependence. That is, they will be something like e to the N with some something that will appear after.

An example of that is when we were, for example, calculating the phase space of gas particles. A gas particle by itself can be in a volume V. Two of them, jointly, can occupy a volume V squared. Three of them, V cubed, et cetera. Eventually you hit V to the N for N particles. So that's a kind of exponential dependence. So this is e g V to the N that you would have for joined volume of N particles. OK?

So some curious things happen when you have these kinds of variables. And one thing that you may not realize is what happens when you summing exponentials. So let's imagine that I have a sum composed of a number of terms i running from one to script N-- script n is the number of terms in the sum-- that are of these exponential types. So let's actually sometimes I will call this-- never mind. So let's call these e to the N phi--

Let me write it in this fashion. Epsilon i where epsilon i satisfies two conditions. One of them, it is positive. And the other is that it has this kind of exponential dependence. It is order of e to the N phi i where there could be some prefactor or something else in front to give you dimension and stuff like that that you were discussing.

I assume that the number of terms is less than or of the order of some polynomial. OK? Then my claim is that, in some sense, the sum S is the largest term. OK?

So let's sort of put this graphically. What I'm telling you is that we have a whole bunch of terms that are these epsilons i's. They're all positive, so I can sort of indicate them by bars of different lengths that are positive and so forth. So let's say

this is epsilon 1, epsilon 2 all the way to epsilon N. And let's say that this guy is the largest. And my task is to add up the length of all of these things.

So how do I claim that the length is just the largest one. It's in the following sense. You would agree that this sum you say is certainly larger than the largest term, because I have added lots of other things to the largest term, and they are all positive.

I say, fine, what I'm going to do is I'm going to raise the length of everybody else to be the same thing as epsilon max. And then I would say that the sum is certainly less than this artificial sum where I have raised everybody to epsilon max. OK?

So then what I will do is I will take log off this expression, and it will be bounded by log of epsilon max and log of N epsilon max, which is the same thing as log of epsilon max plus log of N. And then I divide by N. And then note that the conditions that I have set up are such that in the limit that N goes to infinity, script N would be P log N over N. And the limit of this as N becomes much less than 1 is 0. Log N over N goes to 0 as N goes to infinity.

So basically this sum is bounded on both sides by the same thing. So what we've established is that essentially log of S over N, its limit as N goes to infinity, is the same thing as a log of epsilon max over N, which is what? If I say my epsilon max's have this exponential dependence, is phi max.

And actually this is again the reason for something that you probably have seen. That using statistical physics let's say a micro-canonical ensemble when you say exactly what the energy is. Or you look at the canonical ensemble where the energy can be all over the place, why do you get the same result? This is why. Any questions on this? Everybody's happy, obviously. Good.

**AUDIENCE:** [INAUDIBLE] a question?

**PROFESSOR:** Yes.

**AUDIENCE:** The N on the end, [INAUDIBLE]?

**PROFESSOR:** There's a script N, which is the number of terms. And there's the Roman N, which is the parameter that is the analog of the number of degrees of freedom. The one that we usually deal in statistical physics would be, say, the number of particles.

**AUDIENCE:** So number of measurements [INAUDIBLE] number of particles.

**PROFESSOR:** Number of measurements?

**AUDIENCE:** So the script N is what?

**PROFESSOR:** The script N could be, for example, I'm summing over all pairs of interactions. So the number of pairs would go like N squared. Now in reality practicality in all cases that you will deal with, this P would be one. So the number of terms that we would be dealing would be of the order of the number of degrees of freedom. So, we will see some examples of that later on.

**AUDIENCE:** [INAUDIBLE] script N might be N squared?

**PROFESSOR:** If I'm forced to come up with a situation where script N is N squared, I would say count the number of pairs. Number of pairs if I have N [? sides ?] is N, N minus 1 over 2. So this is something that goes like N squared over 2. Can I come up with a physical situation where I'm summing over the number of terms? Not obviously, but it could be something like that.

The situations in statistical physics that we come up with is typically, let's say, in going from the micro-canonical to the canonical ensemble, you would be summing over energy levels. And typically, let's say, in a system that is bounded the number of energy levels is proportional to the number of particles.

Now there cases that actually, in going from micro-canonical to canonical, like the energy of the gas in this room, the energy axis goes all the way from 0 to infinity. So there is a continuous version of the summation procedure that we have that is then usually applied which is in mathematics is called the saddle point integration.

So basically there, rather than having to deal with a sum, I deal with an integral. The

integration is over some variable, let's say x. Could be energy, whatever. And then I have a quantity that has this exponential character.

And then again, in some specific sense, I can just look at the largest value and replace this with e to the N phi evaluated at x max. I should really write this as a proportionality, but we'll see what that means shortly.

So basically it's the above picture, I have a continuous variable. And this continuous variable, let's say I have to sum a quantity that is e to the N phi. So maybe I will have to not sum, but integrate over a function such as this. And let's say this is the place where the maximums occur.

So the procedure of saddle point is to expand phi around its maximum. And then I can write i as an integral over x, exponential of N, phi evaluated at the maximum. Now if I'm doing a Taylor series, then next term in the Taylor series typically would involve the first derivative. But around the maximum, the first derivative is 0.

Again if it is a maximum, the second derivative phi double prime evaluated at this xm, would be negative. And that's why I indicate it in this fashion. To sort of emphasize that it is a negative thing, x minus xm squared. And then I would have higher order terms, N minus xm cubed, et cetera. Actually what I will do is I will expand all of those things separately. So I have e to the minus N over 6 phi triple prime. N plus N over 6 phi triple prime, evaluated at xm, x minus xm cubed, and then the fourth order term and so forth. So basically there is a series such as this that I would have to look at.

So the first term you can take outside the integral. And the integration against the one of this is simply a Gaussian. So what I would get is square root of 2 pi divided by the variance, which is N phi double prime. So that's the first term I have taken care of.

Now the next term actually the way that I have it, since I'm expanding something that is third order around a potential that is symmetric. That would give me 0. The next order term, which is x minus xm to the fourth power, you already know how to

calculate averages of various powers with the Gaussian using Wick's Theorem. And it would be related to essentially to the square of the variance. The square of the variance would be essentially the square of this quantity out here. So I will get a correction that is order of 1 over N.

So if you have sufficient energy, you can actually numerically calculate what this is and the higher order terms, et cetera. Yes.

**AUDIENCE:** Could you, briefly remind what the second term in the bracket means?

**PROFESSOR:** This? This?

**AUDIENCE:** The whole thing, on the second bracket.

**PROFESSOR:** In the numerator, I would have N phi m, N phi prime. Let's call the deviation y y. But phi prime is 0 around the maximum. So the next order term will be phi double prime y squared over 2. The next order term will be phi triple prime y cubed over 6. e to the minus N phi triple prime y cubed over 6, I can expand as 1 minus N phi triple prime y cubed over 6, which is what this is. And then you can go and do that with all of the other terms.

Yes.

**AUDIENCE:** Isn't it then you can also expand as N the local maximum?

**PROFESSOR:** Excellent. Good. So you are saying, why didn't I expand around this maximum, around this maximum. So let's do that. xm prime xm double prime. So I would have a series around the other maxima. So the next one would be N to the phi of xm prime, root 2 pi N phi double prime at xm prime. And then one plus order of 1 over N And then the next one, and so forth.

Now we are interested in the limit where N goes to infinity. Or N is much, much larger than 1. In the limit where N is much larger than 1, Let's imagine that these two phi's if I were to plot not e to the phi but phi itself. Let's imagine that these two phi's are different by I don't know, 0.1, 10 to the minus 4. It doesn't matter. I'm multiplying two things with N, and then I'm comparing two exponentials.

So if this maximum was at 1, I would have here e to the N. If this one was at 1 minus epsilon, over here I would have e to the N minus N epsilon. And so I can always ignore this compared to that.

And so basically, this is the leading term. And if I were to take its log and divide by N, what do I get? I will get phi of xm. And then I would get from this something like minus 1/2 log of N phi double prime xm over 2 pi. And I divided by N, so this is 1 over N. And the next term would be order of 1 over N squared.

So systematically, in the large N limit, there is a series for the quantity log i divided by N that starts with phi of xm. And then subsequent terms to it, you can calculate. Actually I was kind of hesitant in writing this as asymptotically equal because you may have worried about the dimensions. There should be something that has dimensions of x here. Now when I take the log it doesn't matter that much. But the dimension appears over here. It's really the size of the interval that contributes which is of the order of N to the 1/2. And that's where the log N comes.

Questions?

Now let me do one example of this because we will need it. We can easily show that N factorial you can write as 0 to infinity dx x to N, e to the minus x. And if you don't believe this, you can start with the integral 0 to infinity of dx e to the minus alpha x being one over alpha and taking many derivatives.

If you take N derivatives on this side, you would have 0 to N dx x to the N, e to the minus alpha x, because every time, you bring down a factor of x. On the other side, if you take derivatives, 1 over alpha becomes 1 over alpha squared, then goes to 2 over alpha cubed, then go c over alpha to the fourth. So basically we will N factorial alpha to the N plus 1. So I just set alpha equals to 1.

Now if you look at the thing that I have to integrate, it is something that has a function of x, the quantity that I should integrate starts as x to the N, and then decays exponentially. So over here, I have x to the N. Out here I have e to the

minus x.

It is not quite of the form that I had before. Part of it is proportional to N in the exponent, part of it is not. But you can still use exactly the saddle point approach for even this function. And so that's what we will do. I will write this as integral 0 to infinity dx e to some function of x where this function of x is N log x minus x. And then I will follow that procedure despite this is not being quite entirely proportional to N.

I will find its maximum by setting phi prime to 0. phi prime is N over x minus 1. So clearly, phi prime to 0 will give me that x max is N. So the location of this maximum that I have is in fact N.

And the second derivative, phi double prime, is minus N over x squared, which if I evaluate at the maximum, is going to be minus 1 over N. Because the maximum occurs at the N.

So if I'm were to make a saddle point expansion of this, I would say that N factorial is integral 0 to infinity, dx e to the phi evaluated at x max, which is N log N minus N. First derivative is 0. The second derivative will give me minus 1 over N with a factor of 2 because I'm expanding second order. And then I have x minus this location of the maximum squared. And there would be higher order terms from the higher order derivatives.

So I can clearly take e to the N log N minus N out front. And then the integration that I have is just a standard Gaussian with a variance that is just proportional to N. So I would get a root 2 pi N. And then I would have higher order corrections that if you are energetic, you can actually calculate. It's not that difficult.

So you get this Stirling's Formula that limit of large N, let's do log of N factorial is N log N minus N. And if you want, you can go one step further, and you have 1/2 log of 2 pi N. And the next order term would be order of 1/N.

Any questions?

OK? Where do I need to use this? Next part, we are going to talk about entropy, information, and estimation.

So the first four topics of the course thermodynamics, probability, this kinetic theory of gases, and basic of statistical physics. In each one of them, you will define some version of entropy. We already saw the thermodynamic one as dQ divided by T meaning dS. Now just thinking about probability will also enable you to define some form of entropy. So let's see how we go about it.

So also information, what does that mean? It goes back to work off Shannon. And the idea is as follows, suppose you want to send a message of N characters. The characters themselves are taken from some kind of alphabet, if you like, x1 through xM that has M characters. So, for example if you're sending a message in English language, you would be using the letters A through Z. So you have M off 26. Maybe if you want to include space, punctuation, it would be larger than that.

But let's say if you're dealing with English language, the probabilities of the different characters are not the same. So S and P, you are going to encounter much more frequently than, say, Z or X. So let's say that the frequencies with which we expect these characters to occur are things like P1 through PM. OK?

Now how many possible messages are there? So number of possible messages that's are composed of N occurrences of alphabet of M letters you would say is M to the N. Now, Shannon was sort of concerned with sending the information about this message, let's say, over a line where you have converted it to, say, a binary code. And then you would say that the number of bits that would correspond to M to the N is the N log base 2 of M.

That is, if you really had the simpler case where your selections was just head or tail, it was binary. And you wanted to send to somebody else the outcome of 500 throws of a coin. It would be a sequence of 500 0's and 1's corresponding to head or tails. So you would have to send for the binary case, one bit per outcome.

If it is something like a base of DNA and there are four things, you would have two

per base. So that would be log 4 base 2. And for English, it would be log 26 or whatever the appropriate number is with punctuation-- maybe comes to 32-- possible characters than five per [? element ?]. OK.

But you know that if you sort of were to look at all possible messages, most of them would be junk. And in particular, if you had used this simple substitution code, for example, to mix up your message, you replaced A by something else, et cetera, the frequencies would be preserved. So sort of clearly a nice way to decode this substitution code, if you have a long enough text, you sort of look at how many repetitions they are and match them with their frequencies that you expect for a real language.

So the number of possible messages-- So in a typical message, what you expect Ni, which is Pi N occurrences, of xi. So if you know for example, what the frequencies of the letters in the alphabet are, in a long enough message, you expect that typically you would get that number. Of course, what that really means is that you're going to get correction because not all messages are the same. But the deviation that you would get from getting something that is proportional to the probability through the frequency in the limit of a very long message would be of the order of N to the 1/2.

So ignoring this N to the 1/2, you would say that the typical message that I expect to receive will have characters according to these proportions. So if I asked the following question, not what are the number of all possible messages, but what is the number of typical messages? I will call that g. The number of typical messages would be always of distributing these number of characters in a message of length N. Again there are clearly correlations. But for the time being, forgetting all of the correlations, if [? we ?] [? do ?] correlations, we only reduce this number. So this number is much, much less time M to the N.

Now here is I'm going to make an excursion to so far everything was clear. Now I'm going to say something that is kind of theoretically correct, but practically not so much. You could, for example, have some way of labeling all possible typical messages. So you would have-- this would be typical message number one,

number two, all the way to typical message number g. This is the number of typical message.

Suppose I could point out to one of these messages and say, this is the message that was actually sent. How many bits of information would I have to that indicate one number out of g? The number of bits of information for a typical message, rather than being this object, would simply be log g.

So let's see what this log g is. And for the time being, let's forget the basis. I can always change basis by dividing by log of whatever quantity I'm looking at the basis. This is they log of N factorial divided by these product over i of Ni factorials which are these Pi N's. And in the limit of large N, what I can use is the Stirling's Formula that we had over there. So what I have is N log N minus N in the numerator. Minus sum over i Ni log of Ni minus Ni.

Of course the sum over Ni's cancels this N, so I don't need to worry about that. And I can rearrange this. I can write this as this N as sum over i Ni. Put the terms that are proportional to Ni together. You can see that I get Ni log of Ni over N, which would be log of Pi. And I can actually then take out a factor of N, and write it as sum over i Pi log of Pi.

And just as a excursion, this is something that you've already seen hopefully. This is also called mixing entropy. And we will see it later on, also. That is, if I had initially a bunch of, let's say, things that were of color red, and separately in a box a bunch of things that are color green, and then bunch of things that are a different color, and I knew initially where they were in each separate box, and I then mix them up together so that they're putting all possible random ways, and I don't know which is where, I have done something that is irreversible.

It is very easy to take these boxes of marbles of different colors and mix them up. You have to do more work to separate them out. And so this increase in entropy is given by precisely the same formula here. And it's called the mixing entropy. So what we can see now that we sort of rather than thinking of these as particles, we were thinking of these as letters. And then we mixed up the letters in all possible

ways to make our messages.

But quite generally for any discrete probability, so a probability that has a set of possible outcomes Pi, we can define an entropy S associated with these set of probabilities, which is given by this formula. Minus sum over i Pi log of Pi. If you like, it is also this-- not quite, doesn't makes sense-- but it's some kind of an average of log P.

So anytime we see a discrete probability, we can certainly do that. It turns out that also we will encounter in cases later on, where rather than having a discrete probability, we have a probability density function. And we would be very tempted to define an entropy associated with a PDF to be something like minus an integral dx P of x log of P of x. But this is kind of undefined. Because probability density depends on some quantity x that has units.

If this was probability along a line, and I changed my units from meters to centimeters, then this log will gain a factor that will be associated with the change in scale So this is kind of undefined. One of the miracles of statistical physics is that we will find the exact measure to make this probability in the continuum unique and independent of the choice of-- I mean, there is a very precise choice of units for measuring things that would make this well-defined. Yes.

**AUDIENCE:** But that would be undefined up to some sort of [INAUDIBLE].

**PROFESSOR:** After you [INAUDIBLE].

**AUDIENCE:** So you can still extract dependencies from it.

**PROFESSOR:** You can still calculate things like differences, et cetera. But there is a certain lack of definition. Yes.

**AUDIENCE:** [INAUDIBLE] the relation between this entropy defined here with the entropy defined earlier, you notice the parallel.

**PROFESSOR:** We find that all you have to do is to multiply by a Boltzmann factor, and they would become identical. So we will see that. It turns out that the heat definition of entropy,

once you look at the right variables to define probability with, then the entropy of a probability distribution is exactly the entropy that comes from the heat calculation. So up to here, there is a measured numerical constant that we have to define.

All right. But what does this have to do with this Shannon story? Going back to the story, if I didn't know the probabilities, if I didn't know this, I would say that I need to pass on this amount of information. But if I somehow constructed the right scheme, and the person that I'm sending the message knows the probabilities, then I need to send this amount of information, which is actually less than N log M.

So clearly having knowledge of the probabilities gives you some ability, some amount of information, so that you have to send less bits. OK. So the reduction in number of bits due to knowledge of P is the difference between N log M, which I had to do before, and what I have to do now, which is N Pi sum over i Pi log of Pi.

So which is N log M plus sum over i Pi log of Pi. I can evaluate this in any basis. If I wanted to really count in terms of the number of bits, I would do both of these things in log base 2.

It is clearly something that is proportional to the length of the message. That is, if I want to send a book that these twice as big, the amount of bits will be reduced proportionately by this amount. So you can define a quantity that is basically the information per bit. And this is given the knowledge of the probabilities, you really have gained an information per bit which is the difference of log M and sum over i Pi log Pi.

Up to a sign and this additional factor of log N, the entropy-- because I can actually get rid of this N-- the entropy and the information are really the same thing up to a sign.

And just to sort of make sure that we understand the appropriate limits. If I have something like the case where I have a uniform distribution. Let's say that I say that all characters in my message are equally likely to occur. If it's a coin, it's unbiased coin, it's as likely in a throw to be head or tail. You would say that if it's an unbiased

coin, I really should send one bit per throw of the coin. And indeed, that will follow from this.

Because in this case, you can see that the information contained is going to be log M. And then I have plus 1 over M log of 1 over M. And there are M such terms that are uniform. And this gives me 0. There is no information here. If I ask what's the entropy in this case. The entropy is M terms. Each one of them have a factor of 1 over M. And then I have a log of 1 over M. And there is a minus sign here overall. So this is log of M.

So you've probably seen this version of the entropy before. That if you have M equal possibilities, the entropy is related to log M. This is the case where all of outcomes are equally likely. So basically this is a uniform probability. Everything is equally likely. You have no information. You have this maximal possible entropy.

The other extreme of it would be where you have a definite result. You have a coin that always gives you heads. And if the other person knows that, you don't need to send any information. No matter thousand times, it will be thousand heads. So here, Pi is a delta function. Let's say i equals to five or whatever number is. So one of the variables in the list carries all the probability. All the others carry 0 probability.

How much information do I have here? I have log M. Now when I go and looked at the list, in the list, either P is 0, or P is one, but the log of 1 and M is 0. So this is basically going to give me 0. Entropy in this case is 0. The information is maximum. You don't need to pass any information.

So anything else is in between. So you sort of think of a probability that is some big thing, some small things, et cetera, you can figure out what its entropy is and what is information content is. So actually I don't know the answer. But presume it's very easy to figure out what's the information per character of the text in English language. Once you know the frequencies of the characters you can go and calculate this. Questions. Yes.

**AUDIENCE:** Just to clarify the terminology, so the information means the [INAUDIBLE]?

**PROFESSOR:** The number of bits that you have to transmit to the other person. So the other person knows the probability. Given that they know the probabilities, how many fewer bits of information should I send to them? So their knowledge corresponds to a gain in number of bits, which is given by this formula.

If you know that the coin that I'm throwing is biased so that it always comes heads, then I don't have to send you any information. So per every time I throw the coin, you have one bit of information.

Other questions?

**AUDIENCE:** The equation, the top equation, so natural log [INAUDIBLE] natural log of 2, [INAUDIBLE]?

**PROFESSOR:** I initially calculated my standing formula as log of N factorial is N log N minus N. So since I had done everything in natural log, I maintained that. And then I used this symbol that log, say, 5 2 is the same thing that maybe are used with this notation. I don't know.

So if I don't indicate a number here, it's the natural log. It's base e. If I put a number so log, let's say, base 2 of 5 is log 5 divided by log 2.

**AUDIENCE:** So [INAUDIBLE]?

**PROFESSOR:** Log 2, log 2. Information.

**AUDIENCE:** Oh.

**PROFESSOR:** Or if you like, I could have divided by log 2 here.

**AUDIENCE:** But so there [INAUDIBLE] all of the other places, and you just [? write ?] all this [INAUDIBLE]. All right, thank you, [? Michael. ?]

**PROFESSOR:** Right. Yeah. So this is the general way to transfer between log, natural log, and any log. In the language of electrical engineering, where Shannon worked, it is common to express everything in terms of the number of bits. So whenever I'm expressing

things in terms of the number of bits, I really should use the log of 2.

So I really, if I want to use information, I really should use log of 2. Whereas in statistical physics, we usually use the natural log in expressing entropy.

**AUDIENCE:** Oh, so it doesn't really matter [INAUDIBLE].

**PROFESSOR:** It's just an overall coefficient. As I said that eventually, if I want to calculate to the heat version of the entropy, I have to multiply by yet another number, which is the Boltzmann constant. So really the conceptual part is more important than the overall numerical factor. OK?

I had the third item in my list here, which we can finish with, which is estimation.

So frequently you are faced with the task of assigning probabilities. So there's a situation. You know that there's a number of outcomes. And you want to assign probabilities for these outcomes. And the procedure that we will use is summarized by the following sentence that I have to then define.

The most unbiased-- let's actually just say it's the definition if you like-- the unbiased assignment of probabilities maximizes the entropy subject to constraints. Known constraints. What do I mean by that?

So suppose I had told you that we are throwing a dice. Or let's say a coin, but let's go back to the dice. And the dice has possibilities 1, 2, 3, 4, 5, 6. And this is the only thing that I know. So if somebody says that I'm throwing a dice and you don't know anything else, there's no reason for you to privilege 6 with respect to 4, or 3 with respect to 5. So as far as I know, at this moment in time, all of these are equally likely. So I will assign each one of them for probability of 1/6.

But we also saw over here what was happening. The uniform probability was the one that had the largest entropy. If I were to change the probability so that something goes up and something goes down, then I calculate that formula. And I find that the-- sorry-- the uniform one has the largest entropy. This has less entropy compared to the uniform one.

So what we have done in assigning uniform probability is really to maximize the entropy subject to the fact that I don't know anything except that the probabilities should add up to 1. But now suppose that somebody threw the dice many, many times. And each time they were throwing the dice, they were calculating the number. But they didn't give us the number and frequency is what they told us was that at the end of many, many run, the average number that we were coming up was 3.2, 4.7, whatever. So we know the average of M.

So I know now some other constraint. I've added to the information that I had. So if I want to reassign the probabilities given that somebody told me that in a large number of runs, the average value of the faces that showed up was some particular value. What do I do? I say, well, I maximize S which depends on these Pi's, which is minus sum over i Pi log of Pi, subjected to constraints that I know.

Now one constraint you already used previously is that the sum of the probabilities is equal to 1. This I introduce here through a Lagrange multiplier, alpha, which I will adjust later to make sure that this holds. And in general, what we do if we have multiple constraints is we can add more and more Lagrange multipliers. And the average of M is sum over, let's say, i Pi. So 1 times P of 1, 2 times P of 2, et cetera, will give you whatever the average value is.

So these are the two constraints that I specified for you here. There could've been other constraints, et cetera. So then, if you have a function with constraint that you have to extremize, you add these Lagrange multipliers. Then you do dS by dPi. Why did I do this? dS by dPi, which is minus log of Pi from here. Derivative of log P is 1 over P, with this will give me minus 1. There is a minus alpha here. And then there's a minus beta times i from here.

And extremizing means I have to set this to 0. So you can see that the solution to this is Pi-- or actually log of Pi, let's say, is minus 1 plus alpha minus beta i. So that Pi is e to the minus 1 plus alpha e to the minus beta times i.

I haven't completed the story. I really have to solve the equations in terms of alpha and beta that would give me the final results in terms of the expectation value of i as

well as some other quantities. But this is the procedure that you would normally use to give you the unbiased assignment of probability.

Now this actually goes back to what I said at the beginning. That there's two ways of assigning probabilities, either objectively by actually doing lots of measurement, or subjectivity. So this is really formalizing what this objective procedure means. So you put in all of the information that you have, the number of states, any constraints. And then you maximize entropy that we defined what it was to get the best maximal entropy for the assignment of probabilities consistent with things that you know.

You probably recognize this form as kind of a Boltzmann weight that comes up again and again in statistical physics. And that is again natural, because there are constraints, such as the average value of energy, average value of the number of particles, et cetera, that consistent with maximizing their entropy, give you forms such as this. So you can see that a lot of concepts that we will later on be using in statistical physics are already embedded in these discussions of probability. And we've also seen how the large N aspect comes about, et cetera.

So we now have the probabilistic tools. And from next time, we will go on to define the degrees of freedom. What are the units that we are going to be talking about? And how to assign them some kind of a probabilistic picture. And then build on into statistical mechanics. Yes.

**AUDIENCE:** So here, you write the letter i to represent, in this case, the results of a random die roll, that you can replace it with any function of a random variable.

**PROFESSOR:** Exactly. So I could have, maybe rather than giving me the average value of the number that was appearing on the face, they would have given me the average inverse. And then I would have had this. I could have had multiple things. So maybe somebody else measures something else. And then my general form would be e to the minus beta measurement of type one, minus beta 2 measurement of type two, et cetera. And the rest of thing over here is clearly just a constant of proportionality that I would need to adjust for the normalization.

OK? So that's it for today.