Statistical Mechanics

Prof. Ashwin Joy
Department of Physics
Indian Institute of Technology, Madras

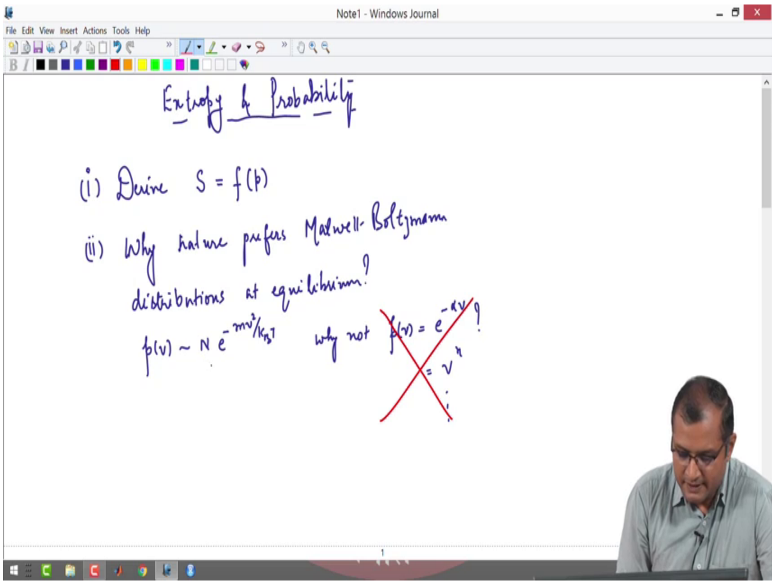
Lecture – 09
Entropy and Probability

So, good morning students, today we will discuss a very important concept infact two concepts which basically connection between Entropy and Probability. And this is the first time I will in this course derive from first principles the connection between entropy and probability. After deriving this connection I will show why nature prefers to order particles at equilibrium to have a velocity distribution which is normal or Maxwell Boltzmann.

So, you know that if you look at a system which is isolated and has been given sufficient amount of time to evolve the velocity distributions for example, for particles in this room will be normally distributed. Also for example, if we have energy levels of for atoms at a given temperature and you ask yourself a question how are these energy levels distributed or how are the particles distributed in energy levels then the distribution turns out to be Maxwell Boltzmann like. Why is it always Maxwell Boltzmann for system z equilibrium and the answer lies in entropy maximization.

We shall see in the lecture that there are these distributions the only distribution which maximizes entropy are the Maxwell Boltzmann distribution. So, we will derive this formally and that is going to be the second half of this lecture. So, I will start with so, I will write down the basically the focus of our present lecture.

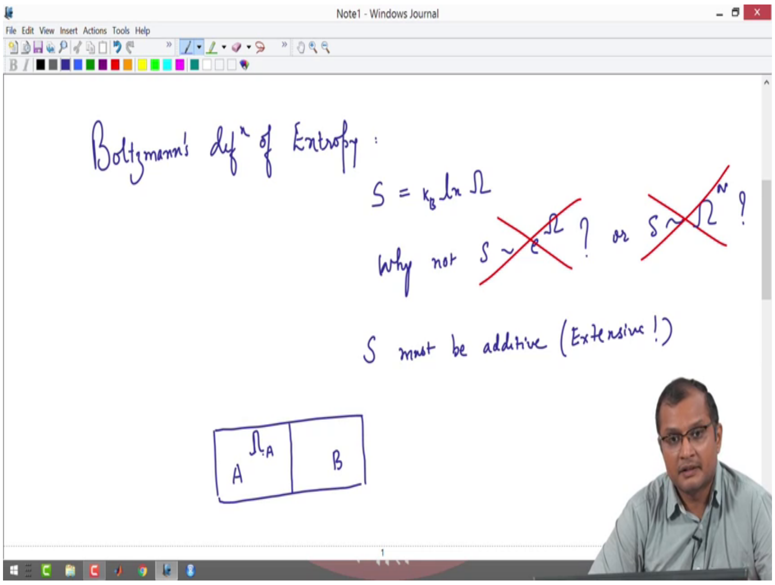
(Refer Slide Time: 02:11)



So, I am going to talk about entropy and probability and here I am going to first the first task is to derive a connection between entropy and probability. So, I will show that entropy is a function of probability ok. So, this is the connection and in the second task towards the second half of the lecture I will show why you know why nature prefers Maxwell Boltzmann distributions at equilibrium ok. For example, why is the distribution of velocity is let us say if we have a one dimensional gas at equilibrium thermal equilibrium t is always you know to some normalization constant always e raise to minus m v square over K BT ok.

Why not something like e raise to minus some alpha times v exponential ok? So, all these distributions you know you can take some v to the power n and so and so forth all these distributions are rejected and what is selected is normal distribution at equilibrium ok. So, the reason why this distribution is selected over other distribution is because a Gaussian distribution is a distribution of maximum entropy ok. So, we shall systematically derive this distribution by maximizing entropy. So, that is the agenda of a lecture. So, we will start off with simple Boltzmann's definition of entropy.

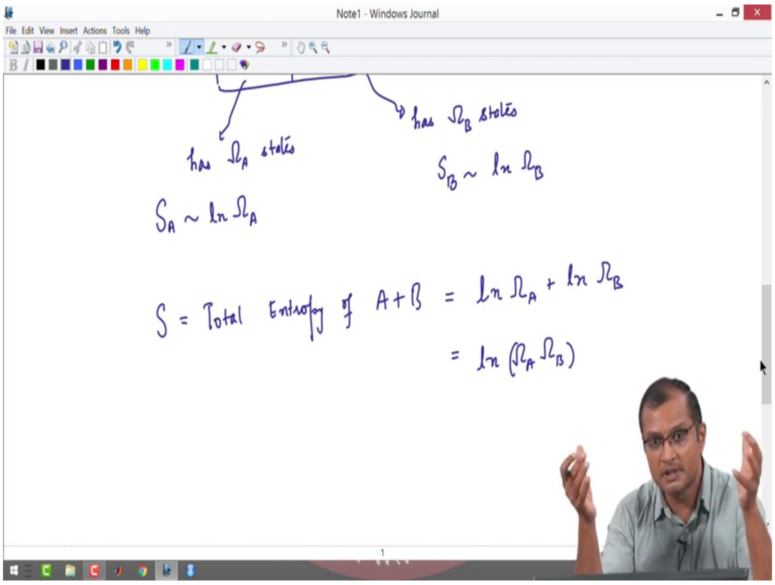
(Refer Slide Time: 04:59)



Remember the task is to connect entropy with probability. So, we will start off with the definitions of entropy given by Boltzmann. So, Boltzmann gave us the prescription of entropy then it is logarithmically related to the number of accessible states. So, this is the famous connection between number of states and the entropy for any system given by Boltzmann ok. Now why not we take why did not Boltzmann take something like e raise to omega or as going as some omega to the power N. What is the a reason why Boltzmann did not take choices like e to power omega or N omega to the power N?

Where the simple reason that he wanted entropy to be additive or extensive and we shall see why logarithmic dependents on number of states makes entropy naturally extensive. Suppose you take a system and call it let us say you have two systems A and B, our system A can exist in let us say omega A number of states ok. So, number of states for our system A has omega A I am going to write it outside so, that the figure does not become busy.

(Refer Slide Time: 07:07)

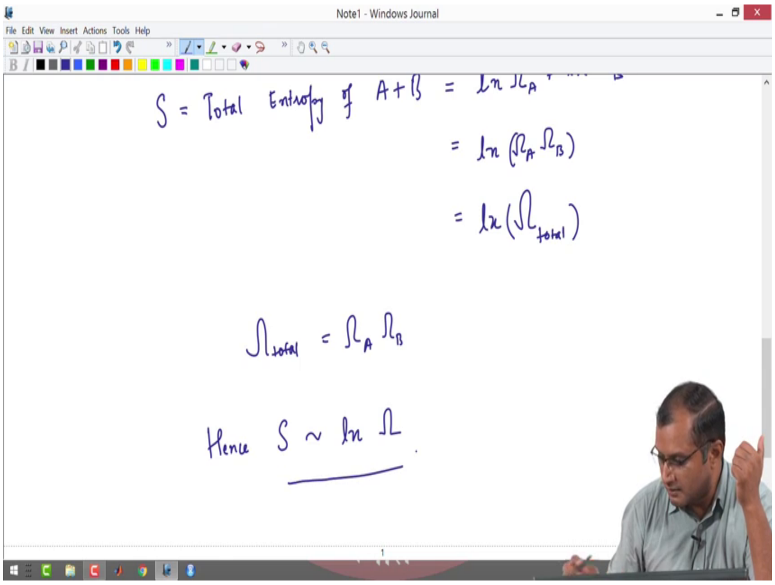


So, this system has these many states and let us assume that the system B has access to omega B states you know number of states ok. So, one can say that the entropy of the system A is like lon of omega A the natural log of omega A and similarly entropy of system B is like lon of omega B fine. So, one would ask a natural question what is the joint entropy or total entropy of the joint system and I am call the joint system as A plus B.

So, this would be definitely this would certainly go as lon of I am dropping the constant factors the Boltzmann constant has been dropped we have taken Boltzmann constant to be 1. So, you can take it as lon of omega A plus lon of omega B and then you can write it as lon of omega A times lon of omega B.

So, I am basically writing down the entropy to be additive because that is what I want and it should be in extensive quantity and it should scale with system size. So, if I double up the system the entropy is expected to double up and so, I have simply added the two entropies.

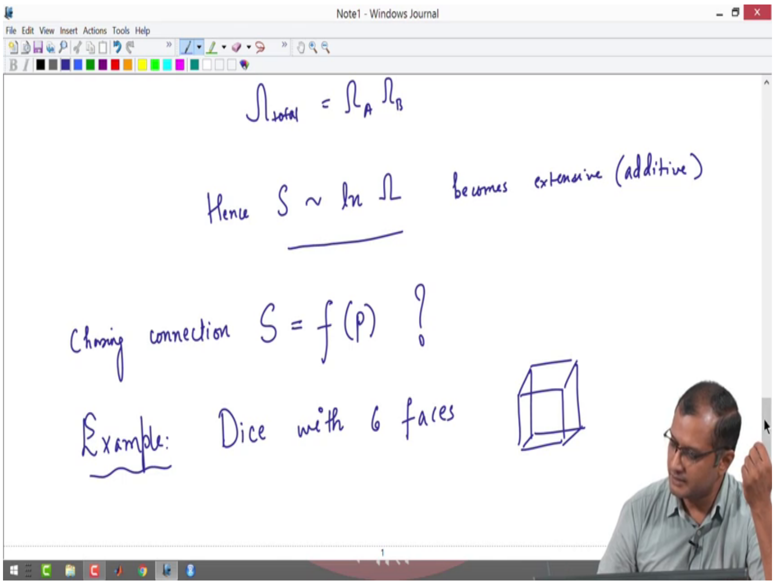
(Refer Slide Time: 09:05)



What it has given me is a logarithm of a product of the number of states in each of the system. And this is precisely nothing, but if I call it as lon of omega total where omega total is the total number of states accessible to joint system.

So, if we have first system existing in 5 states and the next system existing in 6 states the joint system would definitely exist in 30 different states which is precisely what is written here. So, hence you expect S to be logarithmically related to the number of states accessible to a system. So, any other combination any other mathematical formulation for entropy and total states will not lead to additivity.

(Refer Slide Time: 10:27)



So, this is the nice you know this is the reason why Boltzmann has taken entropy to be logarithmically related to the number of states, it makes entropy naturally additive ok. So, I will write down the conclusion. So, this is so, this is the reason why you take entropy to be logarithmically depend on the number of states.

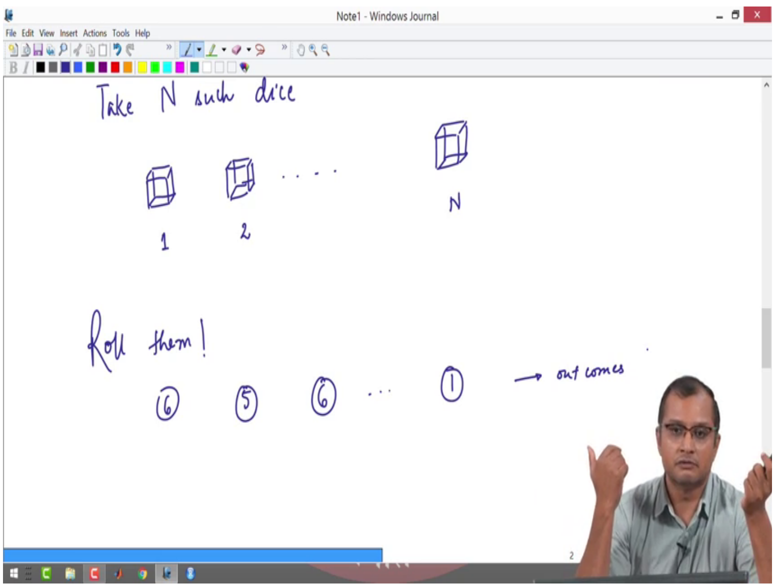
Now, like I said we are chasing a connection between entropy and probability this is the next task. So, I have just derived I have not derived I have just explained why entropy is logarithmically related to number of states; now I am going to develop a connection between probability and entropy ok. So, I know that entropy is a measure of disorder. So, the relationship that I am going to chase between entropy and probability has to respect that physical meaning of that entropy. And so, by disorder I mean higher access to number of states access to a higher number of states larger number of states.

So, certainly if you go into a classroom and you are 5 students let us say and you go to a class room and you see 5 benches. And the benches are wide enough to accommodate only 1 student, what you will do is you will basically go and sit in each one of the benches and all benches will be occupied. So, 5 benches, 5 students each one sitting in 1 bench because, a bench can accommodate only 1 student. Now if a professor walks in he says that 5 benches and 5 students it is in ordered systems ok. But, imagine these 5 students went into a class which had 50 benches; now they can sit anywhere they could sit together or they could sit you know far away from each other.

So, if now somebody walks into the class they may see certain configuration where the students are all sitting you know at the random locations in 50 benches 5 of them sitting at random locations. So, naturally the second configuration is more disordered than the first configuration and here the disorders simply arose because, you give more number of states to these 5 students to sit; if he had given him less number of states they would have been more ordered like the first case.

So, clearly disorder it can be in some sense connected to total number of states available to increase a number of states you increase possibility of disorder ok. So, I am going to take the example of dice ok. So, we all know a dice is has a 6 faces. So, let us say you have unbiased dice like a, dice is like a cube having 6 faces and let us say that you take N such dice ok.

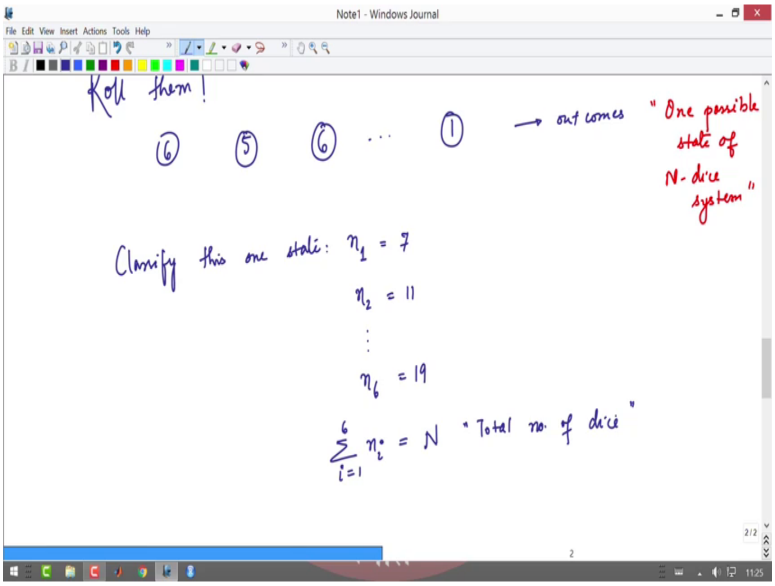
(Refer Slide Time: 14:37)



So, you have taken N such dice. So, you take the first dice and you take the second one and you take the Nth one 1st, 2nd and Nth and you roll them roll all of them ok. So, the first dice may show let us say it may show you know I am going to write only the outcomes here.

So, the 1st dice may show an outcome let us say 6, the 2nd die may show an outcome 5, 3rd die may show gain an outcome 6. So, so on the last die may show let us say an outcome of 1 these are the outcomes ok. So, can you so, is the experiment clear to you have taken N dice and you have rolled all of them. So, the 1st die gave you 6 the 2nd die gave you 5, the 3rd die gave you 6 4 die gave you 4 and the last die gave you 1. So, it is like what I have written in front of you here is 1 state of the N die system ok. So, this is so, these are the outcomes.

(Refer Slide Time: 16:31)



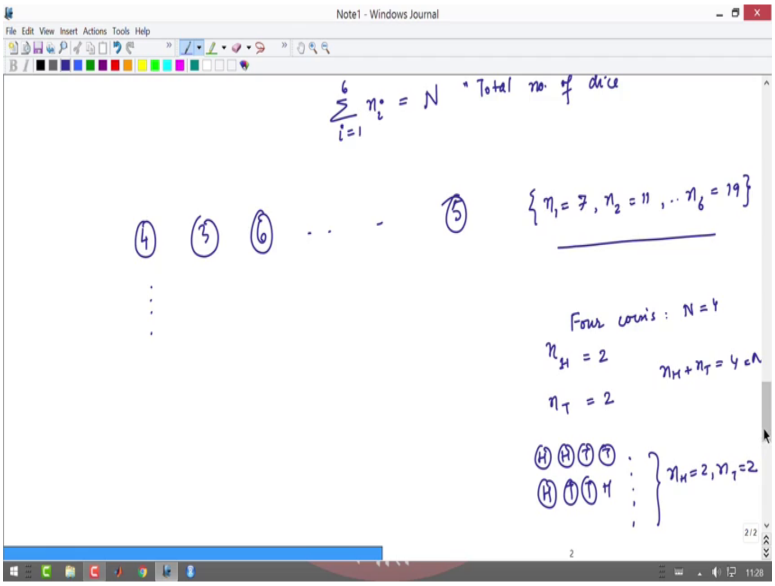
So, I will call this as one possible state of the N dice system is one possible state. So, if you want to classify this state you can classify this state this one state as you can call this as a micro state. So, you can say that this one state can be classified as you can count the number of 6, that you had or count the number 1s count, the number of 2s count, the number of 3s and up to count the number of 6s. So, you can say how many dice gave you the outcome 1?

So, that is n 1 let us say you had you know you had you had take some number you had 7 dices that gave you 1 and how many dice gave you the outcome 2 let us say you had 11 and all the way you count the number of dice that gave you the outcome 6 let us say it is 19 ok. And we know that summation over all the different faces some of these ni's should be the total number of dice this is total number of.

So, what we have done is we have rolled N dice together and we are simply counted how many dice gave you the outcome 1 7 of them gave us the outcome 1, then we counted how many dice gave us the outcome 2 11 of them gave us the outcome 2 and finally, since a die has only 6 faces we do this calculation up to 6. So, number of dice that gave the outcome 6 were 19 and I know the total number of dice that we tossed was big N.

So, summation n i has to be the big N ok, but there are many ways in which I can keep this values of ni’s.

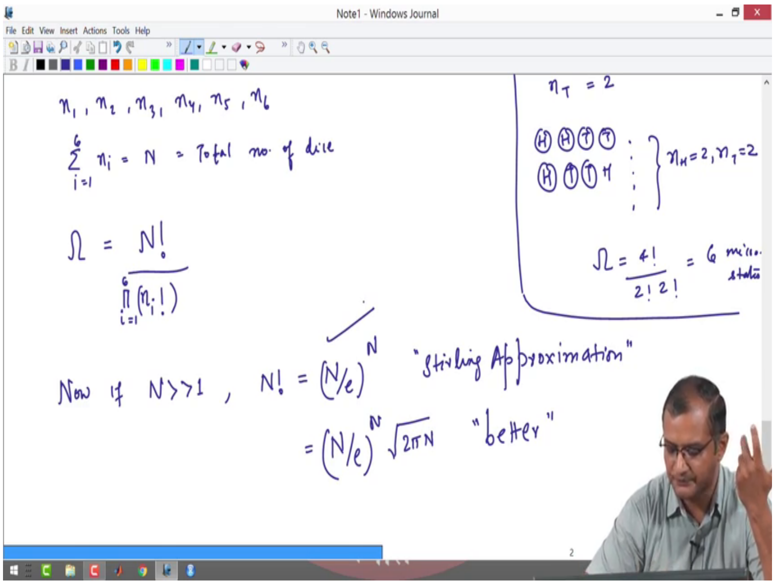
(Refer Slide Time: 19:19)



So, I can I can change my micro state such that the 1st die gave me 4, the 2nd die gave me 3, the 3rd die gave me 6 and so on and the last guy gave me 5 such that I still have n 1 as 7 n 2 as 11 and so on the same macro state. So, it is possible to have many many different types of orderings that correspond to the macro state that I have taken a very simple example would be take a coin toss for example. So, if I take that if I take 4 coins and I say that number of heads. So, I take let us say 4 coins so, n equals to 4 and I say that number of heads are 2 and number of tails are 2.

Such that n H plus n T is 4 how many ways can you have it well you can have you can have let me just write down a few ways and you should be able to tell us naturally. So, you can take a head here and a head here and a tail here and tail here or you can take a head here and tail here tail here and head here. So, all these configurations belong to n H equals to 2 n T equal to 2 2 plus 2 is 4 which is the total number of coins that I have tossed and the answer is clear.

(Refer Slide Time: 21:23)

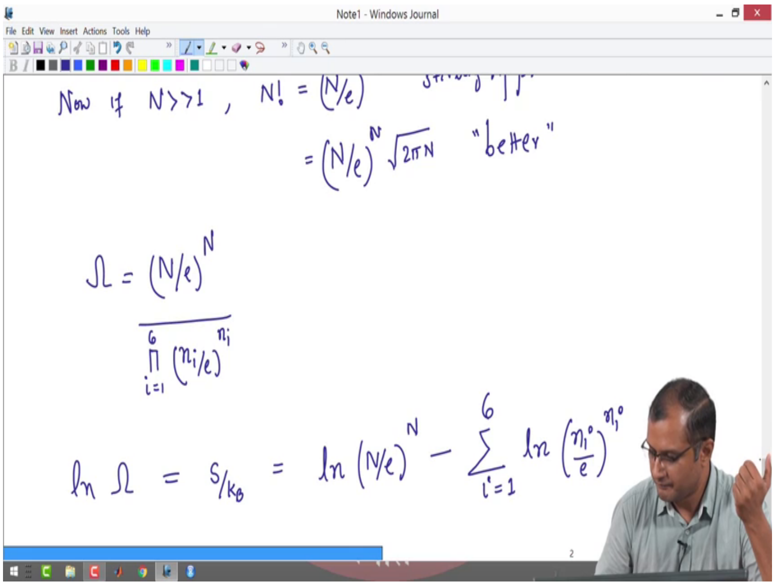


You know the total number of states that are accessible to you is n c 2. So, you can take 4 factorial by 2 factorial into 2 factorial. So, this is basically 6 ways so, we have 6 micro states that are possible or I will write down instead of ways 6 micro states ok. So, that was a simple example from a coin let us go back to our example of the dice. So, like I said if I have tossed the die N dice together and I have simply counted how many times I got the outcome 1 and how many times I got the how many dice gave me the outcome 2 and how many dice gave me the outcome 6?

So, basically I have these numbers n 1, n 2, n3, n 4, n 5 and n 6 because there are only 6 faces and I know that summation i going from 1 to 6 n i is equal to the total number of dice that were that were tossed together ok. So, if I want to know how many states this these N dice can stay together with the specified values of these small n 1 n 2 all the way to n 6 then this is nothing, but factorial N N been the total number of dice divided by the product of these ni's now if n is large then I can approximate N factorial then I can approximate N factorial as N by e to the power N this is called as a Stirling approximation ok.

So, you may actually have a better approximation of Stirling you may take N by e to the power N into square root of 2 pi N this is better, but for the purpose so, both will give you same result I am going to choose this one slightly lesser accurate, but its suffices. So, I know that my N factorial is very large so, I can approximate it as a Stirling approximation.

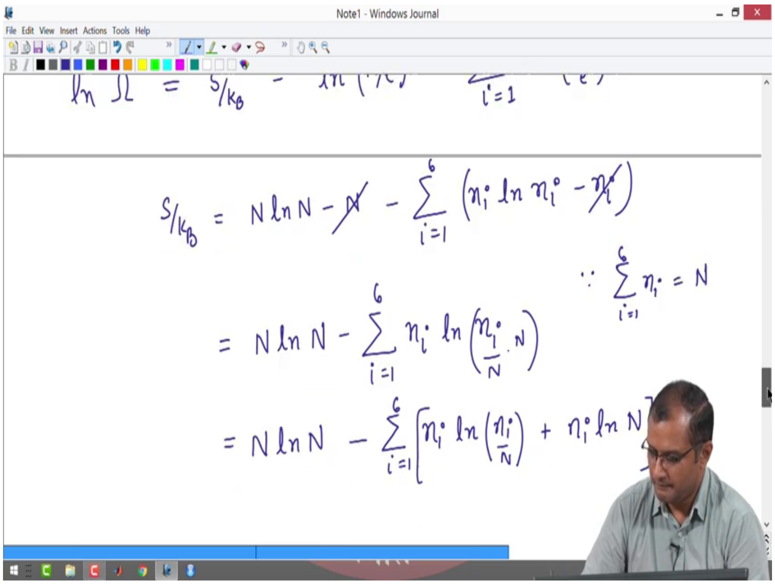
(Refer Slide Time: 24:57)



And this will give me N by e to the power N the denominators are also you know all the small n I's are also large. So, I will approximate these also by Stirling’s formula.

So, I am going to write it as n i by e to the power n i and then I am going to take logarithm on both sides ok. So, we are now taking logarithm on both sides what you will see is basically and I know that this is nothing, but entropy over KB from the Boltzmann’s prescription. So, if I take the logarithm of n the expression splits up into 2 logs and the denominator becomes summation over i and you can write now this as S over k B.

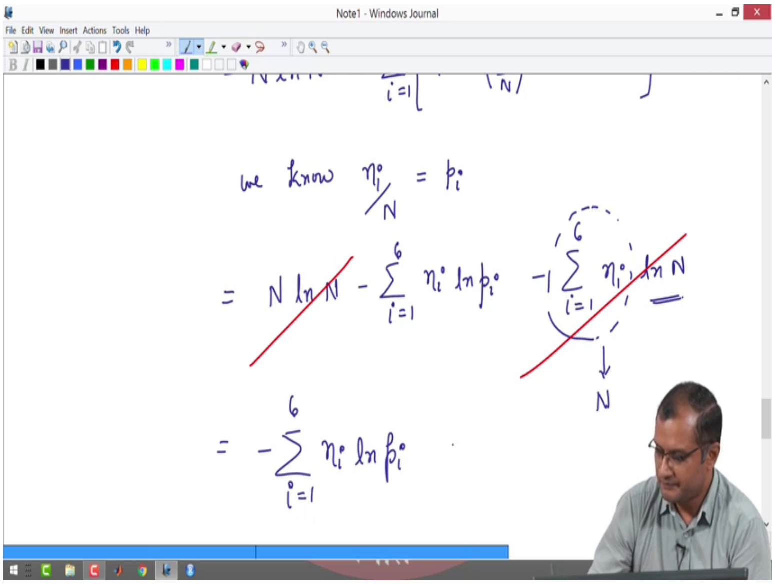
(Refer Slide Time: 26:25)



Now, it is just a matter of simplifying things and log e to the base e is taken as 1. So, this is nothing, but N log N minus N that is the first term minus summation i going from 1 to 6 I am going to write it as n i log lon n i minus n i and what you can see. So, this entire thing is summed over i and since I now that summation over all n i's is my total number of dice. So, basically this term cancels with this term.

So, this is the summation outside the bracket and what I have is basically N log N minus summation i going from 1 to 6 n i log lon n i because summation n i is big N right. So, I need to simplify this further which is very simple what I will do is basically divide this by N and multiply this by N ok. So, let us see what this gives us this will give us n i log n i by n plus n i log N fine.

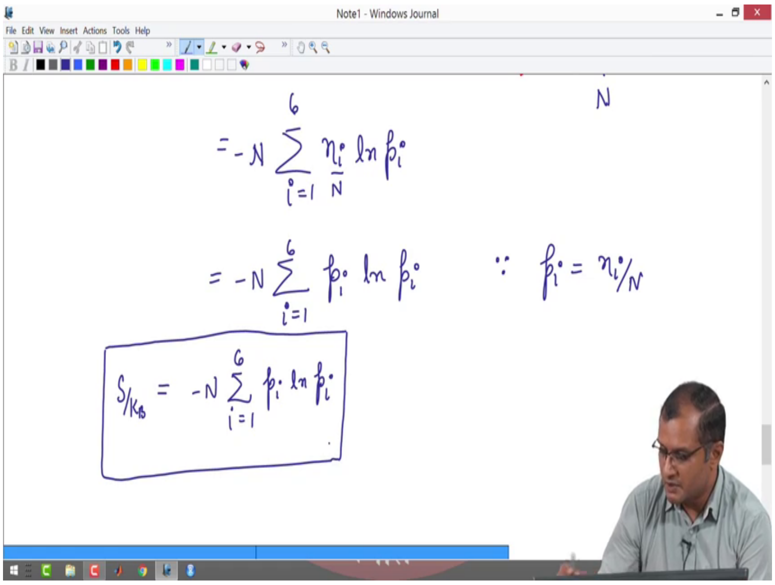
(Refer Slide Time: 28:47)



And we know that so, this is the birth of our appearance over probability we know that number of times the ith outcome came divided by the total number of dice is nothing, but the probability of getting the ith outcome ok. So, I am just going to re write the previous expression as minus summation i going from 1 to 6 n i l n p i and the second expression is simply minus I am just going to split the summation into two summations ok.

Now look at the first term and the last term this is the constant lon N N and summation i going from 1 to 6 n i is nothing, but N ok. So, you can knock off the first term and the third term and what you are left with is just minus summation 1 going i going from 1 to 6 n i lon p i; now we are just 1 step far away from our final expression.

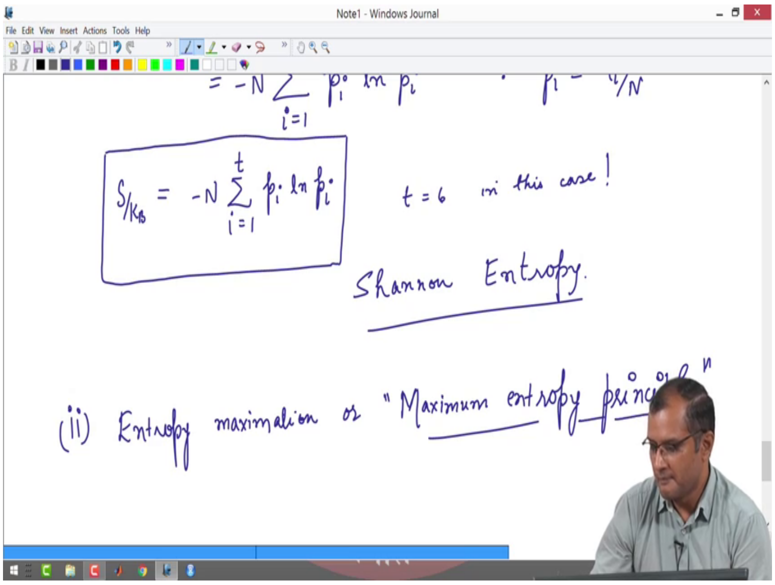
(Refer Slide Time: 30:33)



So, what I am going to do is multiply by N and divide by N. So, add a minus N here. So, I have multiplied the entire thing by N and I have divided each term in the summation by N same thing the entire thing has been multiplied by N and the entire thing is nothing, but sum of N terms or 6 term so, I divide each term by N.

So, this is nothing, but minus N summation i going from 1 to 6 p i l n p i; since we have this condition of probability the probability is defined as n i over N ok. So now, we have this connections so, the left hand side was S over K B as minus N summation i going from 1 to 6 p i l n p i. Now this is derived for a dice where you know each die can have 6 possible outcomes, but in general you can do it for any system where the number of state accessible to each particle is n. So, this is the definition of. So, I could in principle take a t face die instead of taking 6.

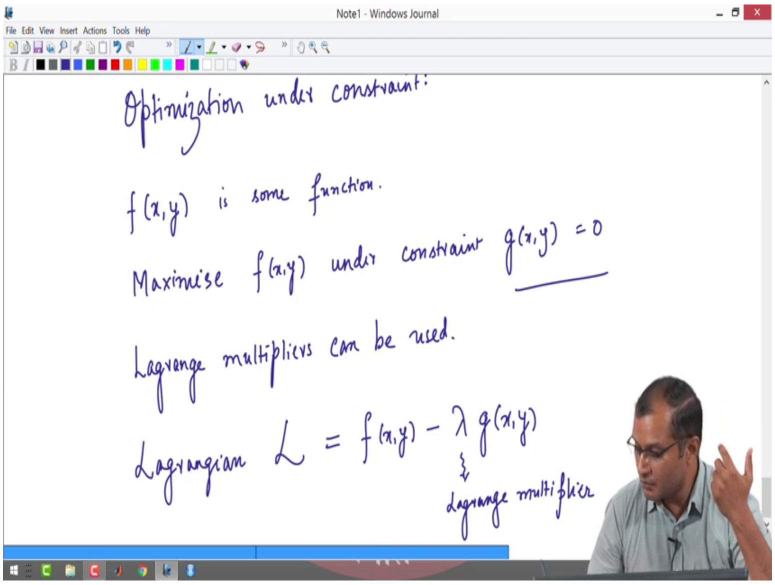
(Refer Slide Time: 32:37)



I could have taken a t faced die and taken the sum up to t ok. So, t was 6 in this case in principle it could be any number ok. So, this is also called some times that the Shannon's form of entropy that is beautiful relationship between probability and entropy ok.

So, this is the connection between probability and entropy. Now, we were focusing on our second agenda was to basically maximize the entropy and as a function of probability and find out which probability density is the maxima of the entropy. So, the next topic is the entropy maximization or maximum entropy principle ok. So, you can name it whatever you want to the agenda is nothing, but fine distributions probability distribution that maximize the entropy; now this is the subject of optimization under a constraint.

(Refer Slide Time: 34:23)

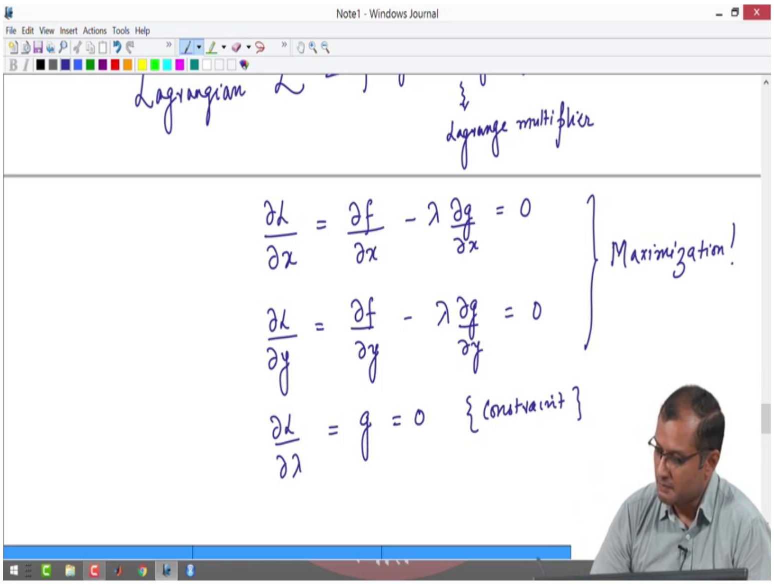


I am optimizing something I am optimizing entropy under certain constraint and I will basically highlight what is mean meant by these constraints ok. So, basically mathematically this is the problem of optimization. So, I am going to basically take an example and explain this point ok.

So, instead of directly optimizing entropy I will take a random function let us a say I have a function f of 2 variables x and y. And a some function and I want to maximize f under a constraint that some g of x y equals to 0 this is my constraint ok. So, this is where you can use a method of Lagrange multipliers ok. So, let me demonstrate this method.

So, the method of Lagrange multipliers simply states that you construct a Lagrangian some sort of a Lagrangian let us say that is L this is written as our function f minus lambda which turns out to be a Lagrange multiplier into g which is our constraint ok. So, this is the Lagrange multiplier and I am going to maximize my SL my Lagrangian I will take one very cute example.

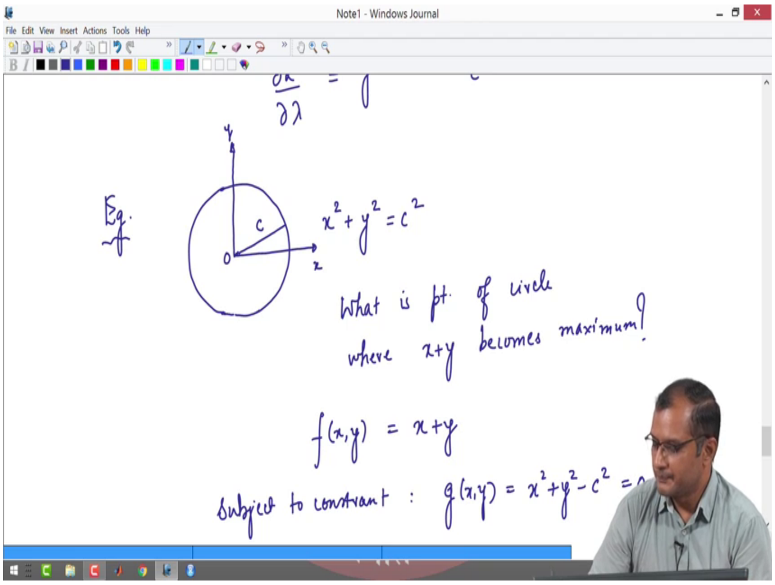
(Refer Slide Time: 37:13)



So, what you do is basically you say that I will take derivative since L is extremized I am going to take derivative with respect to x and this will be del f by del x minus lambda times del g by del x. And I am going to take derivative of L with respect to y and get del f by del y minus lambda times del g by del y and naturally I can also take derivative with respect to lambda which in this case will just give me L g which is 0 and that is my constraint ok.

The derivatives with respect to x and y are also 0 because that is the condition for maximization of l ok. So, these derivatives so, the values of x and y that gives me these derivatives as 0 are the values that maximize f let us take a simple example.

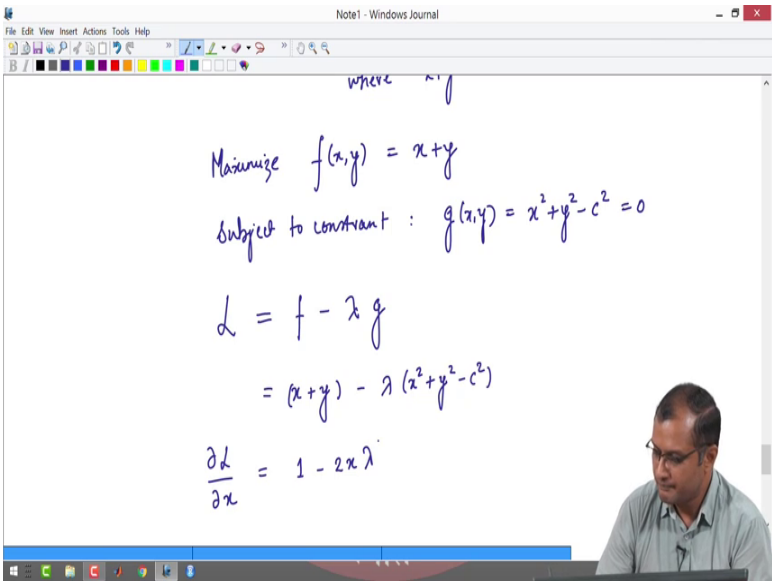
(Refer Slide Time: 38:31)



So, if I asked you if you take a circle. So, let this be a circle of some radius c. So, the equation of circle is x square plus y square equals to c square. And, I ask you simple question let us say the circle is centered at the origin and ask you simple question what is the point on the circle where x plus y becomes maximum ok.

So, the question is very clear; stated mathematically this mean that I have a function x plus y that I want to maximize subject to the constraint my constraint is g of x y which is equal to x square plus y square minus c square equals to 0 ok.

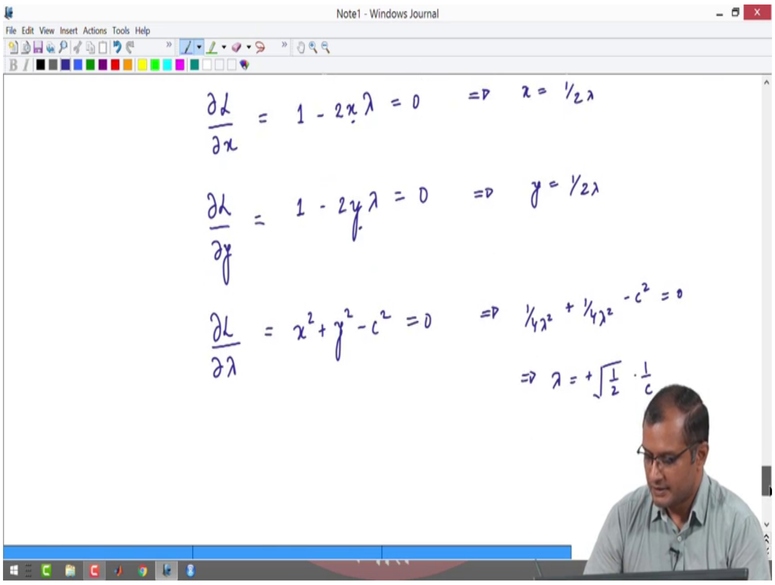
(Refer Slide Time: 40:31)



So, basically this problem means you have to maximize this function x plus y on the circle. So, the constraint is that x square minus y square minus c square is 0 which is my g of x y and I can use the Lagrange multiplier methods.

So, I will say that this is my Lagrangian as I defined earlier f minus lambda times g ok. So now, you can write it as x plus y our function minus lambda times x squared plus y squared minus c square ok, take the derivatives with respect to x and this will give you 1 minus twice x into lambda.

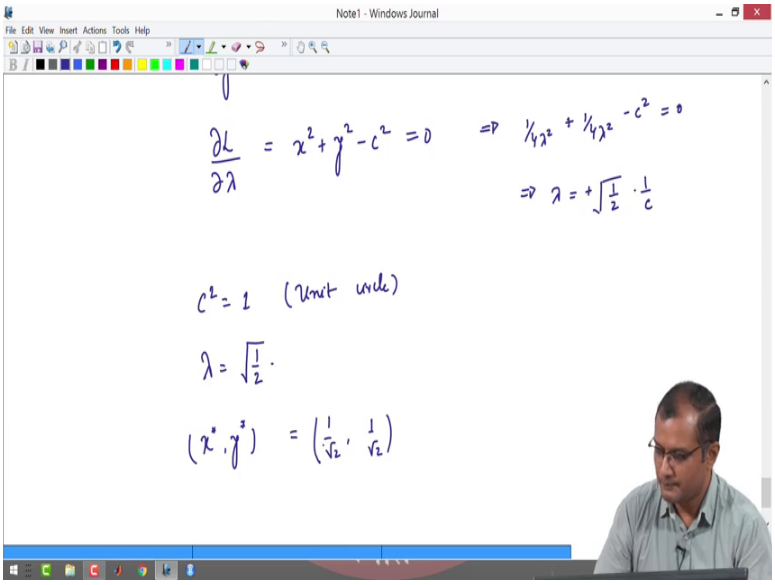
(Refer Slide Time: 41:25)



Take the derivative with respect to y is a partial derivatives. So, you will get again 1 minus twice y lambda and the 3rd equation basically is del L by del lambda which is nothing, but our constraint x square plus y square minus c square equals to 0 ok. So, you have to solve these equations and the x y that you get are the values of x and y where the function x plus y is maximum. Because, you already said that at these values of x and y my Lagrangian is extremized because you have said the derivatives is 0.

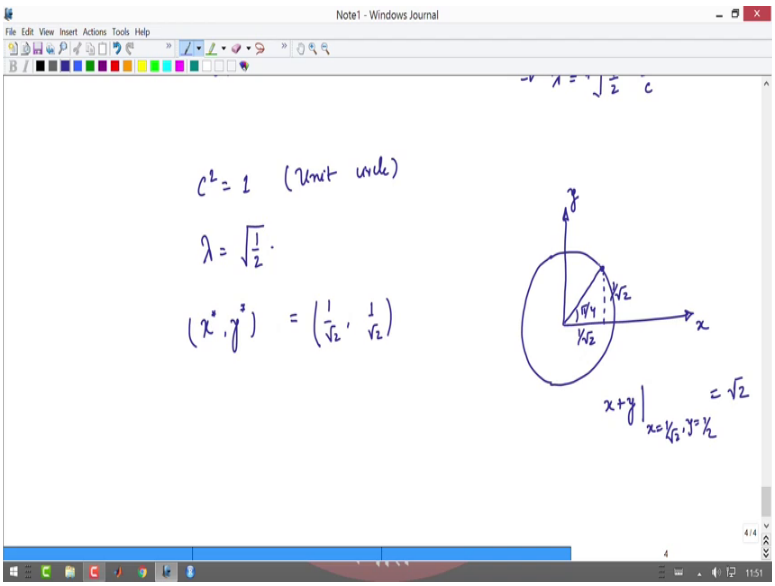
So, from here you can easily say that my x should be 1 upon 2 lambda and from here you can say that my y should be 1 upon 2 lambda and from here you can find out the values of lambda in terms of c. So, you can substitute 1 upon 4 lambda square plus 1 upon 4 lambda square minus c square equals to 0 and that will give you lambda as plus or minus square root of 1 by 2 into 1 by c. So now, we can see the answer.

(Refer Slide Time: 43:15)



So, you can you can take if the if the circle was unit radius circle; then c square would have been 1 for a unit circle and if you could have taken lambda as 1 upon square root 2 and that would have given you the point of maxima you know where x plus y is maximized as simply 1 upon square root 2 by 1 upon square root 2. So, put the values of lambda is 1 upon square root 2 you get these you get these value. So, basically on a on a unit circle you would have been you would have maximized x plus y here at pi by 4.

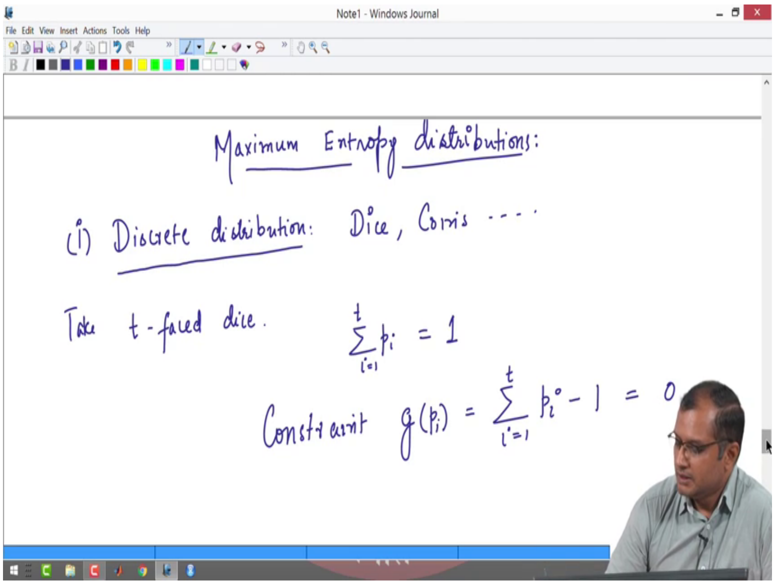
(Refer Slide Time: 43:59)



So, this was the x axis this was the y axis this would have been 1 upon square root 2 this would have been 1 upon square root 2 the sum of x plus y at x equals to 1 upon square root 2 and y to 1 upon square root 2 is a maximum which is nothing, but square root 2 which is more than 1.

So, if you had taken point here it would given you x plus y as just 1 if you had taken a point here your x plus y would have been 1 because their y is 1 x is 0, but if you take it here at theta equals to pi by 4 x plus y comes out to be a square root 2 which is a maximum right. So, let us extend this procedure to find out maximum entropy distributions.

(Refer Slide Time: 45:13)

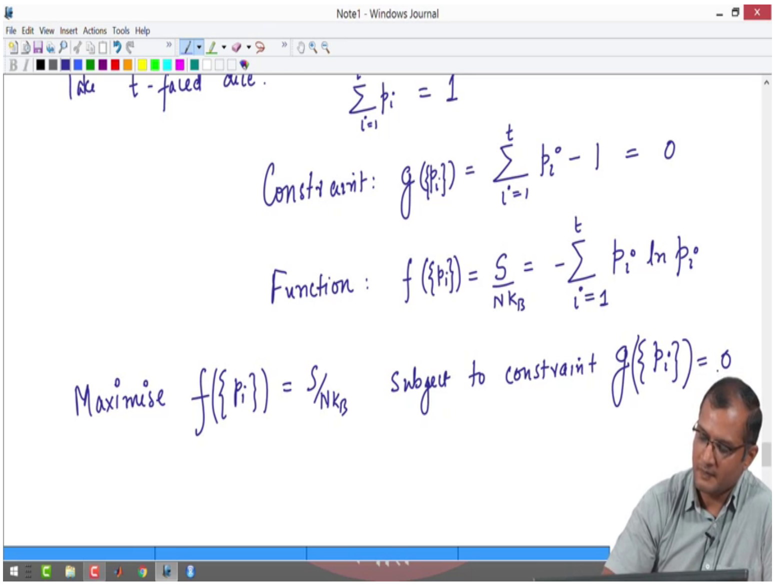


Now this is the first time probably you are introduced to regress derivation of the fact that if you leave something at equilibrium with certain constraints the probability distribution is always Maxwell Boltzmann like ok. So, you may want to pay attention to this to this derivation.

So, I am going to take such an example of a discrete a discrete distribution where the outcomes are discrete like the coins or the dice ok. So, I am going to take systems of dice coins so on so forth. So, you take for example, a t faced like before you take a t faced dice, a dice which has t faces.

And now you say that I know only 1 thing that the sum of probabilities of each face is unity ok. So, in some sense you have a constraint here I call it as g of p pi in fact, as simply summation i going from 1 to 6 pi minus 1 right. So, we have a t face die so, we are generalizing it correct ok, you can take t to be 6 I am leaving it as a variable. So, this is my constraint ok.

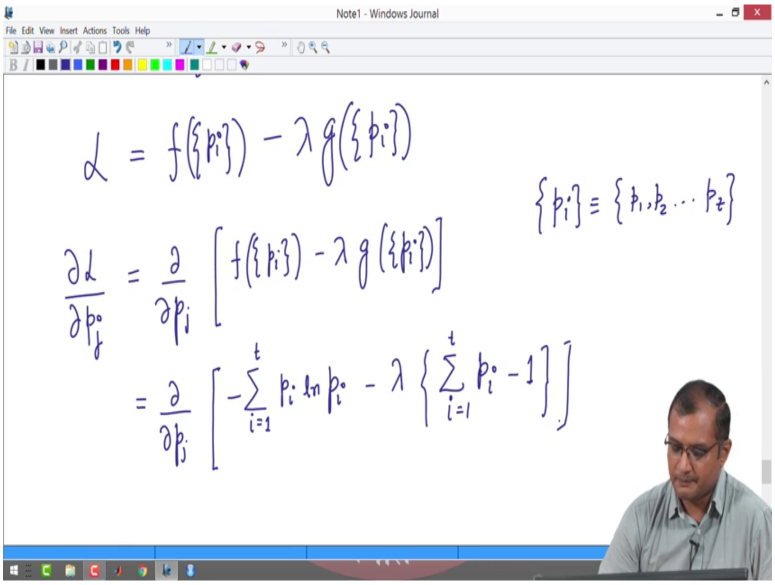
(Refer Slide Time: 47:53)



And my function that I want to extremize or maximize is basically the entropy. So, the function of probabilities are nothing, but the entropy which is I know entropy over I will say I will take it as N K B as upon N K B is the function that I want to maximize it is nothing, but minus of summation I going from 1 to t p i l n p i. And here because this is a function of all the variables p 1 p 2 all the way to p of t I am going to say this is a entire set p of i right this is a suppose it if the die was only 6 faces this would have been a function of p 1 p 2 to p 3 all the way to p 6 this means that this is multi valued so, no its a multivariable function both f and g.

Now if this is your constraint that the probabilities are normalized and you want to maximize your entropy. So, we are going to maximize S over N K B which is our function. So, I am going to write it as before to be consistent I am going to maximize my function which is nothing, but S over N K B subject to the constraint and the constraint is g then I can use the Lagrange multiplier method.

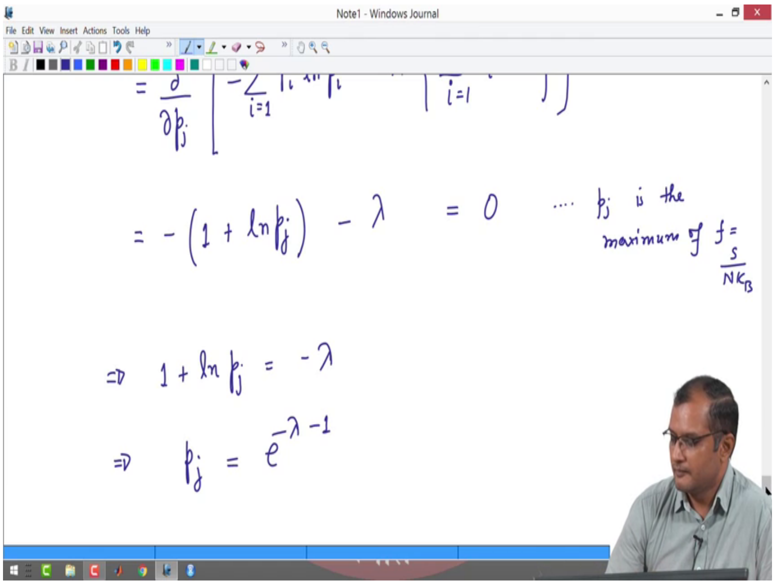
(Refer Slide Time: 50:23)



So, I can define a Lagrangian which is my function minus lambda lambda being the Lagrange multiplier f minus lambda times g prescription remains the same ok. And you know that my variables are these are my variables on which my function is you know these are my variables. So, I take some variables p j so, they are there are basically n equations I will write only 1 of them ok, in like we had x and y now I have p 1 p 2 p up to p t.

So, there are n equations here subject to the value of j that I pick up this would be nothing, but d over d p j of the entire right hand side ok. So, you plug the values here. So, f of p i is nothing, but our function S upon N KB and which I had taken as minus summation i going from 1 to t p i l n p i and lambda times the function g was taken as summation i going from 1 to t p of i minus 1 that was my function g fine.

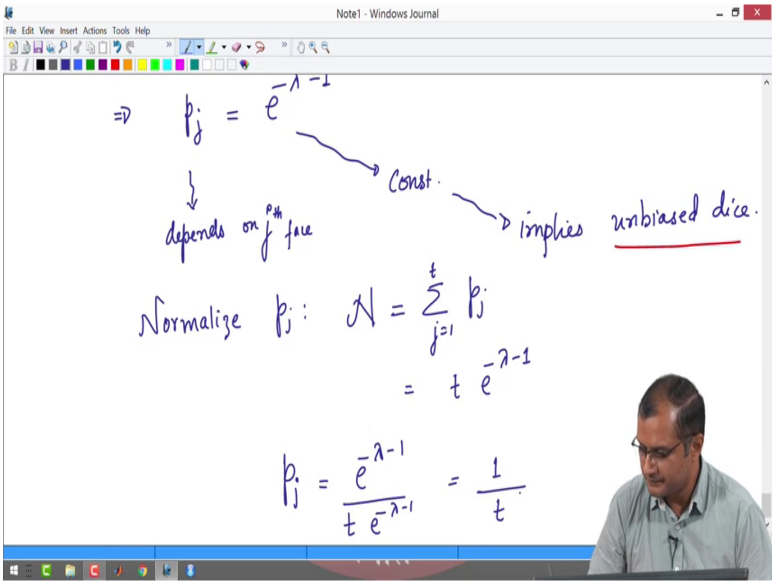
(Refer Slide Time: 52:39)



So, if you take the derivatives with respect to p j see only one term in the summation where p i is equal to p j will contribute. So, that is what I am going to take. So, this would be all other i's will go to the derivatives of all other terms where i is not equal to j will go to 0, only the term i equal to j will contribute here. And, that will simply give me 1 plus l n p j and j here it will give me lambda times 1.

And, this is equal to 0 because this is derivative of the Lagrangian with respect to p j that is our condition is p j is the maximum of our f which is nothing, but S over N K B ok. So, if you solve this for p j what you get is basically 1 plus lon times p j which gives you minus lambda. So, you can write this as p j e equals to e raise to the power minus lambda minus 1.

(Refer Slide Time: 54:37)

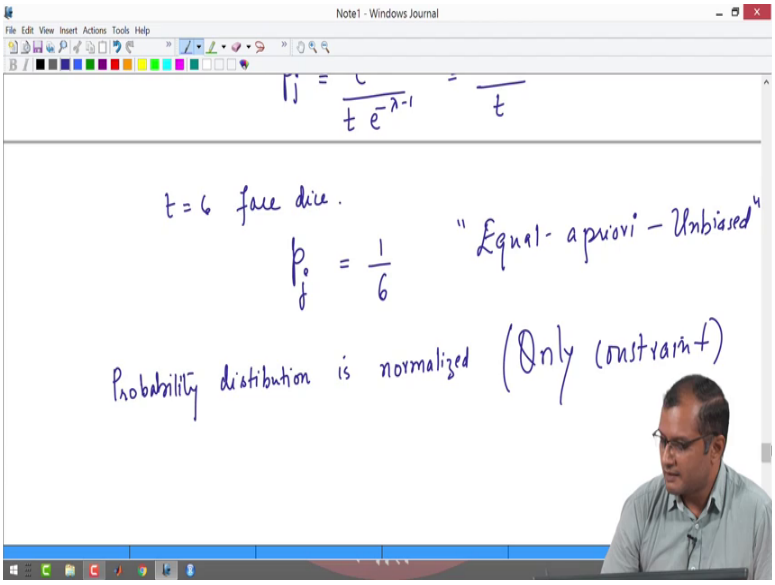


So, the left hand side depends on j where as the right hand side is a constant. So, what can you say about this? What can you say you pick up any face of the die its probability is the same. So, what can you talk what can you tell about such a dice which gives you same probability for all the faces.

So, the dice is basically implies that it is an unbiased dice each face has equal probability. So, the probability of the jth face comes out to be a constant, pick up any j first face, second face, third face, fourth face, tth face they are all the same e raised to lambda minus 1, lambda is a constant. So, the probability of any face is a constant. So, this is an unbiased die. So, that is an important result we just go to normalize it. So, what I am going to do is normalize this probability that is the only task.

So, I can normalize probability by simply computing a norm which is nothing, but summation over all values of p j. And that is nothing, but if i sum over all p j s what I get is simply t times e raised to minus lambda minus 1 because each one them is constant. So, the final normalized probability of a jth face is n normalized probability divided by the norm which is simply 1 upon t; this is your answer and this you know.

(Refer Slide Time: 56:49)



So, if you take t equal to 6 faced die then the probability of any face let us say gth face would simply turn out to be 1 over 6. So, what we saw here is that if you leave the probability distribution without any constraint apart from the fact that the probability distribution is normalized, if that is the only constraint you have I am going to say this is the only constraint you have. Then the PDF that you get is equal apriori distribution or unbiased distribution which means if I have, if I was introduced a constraints then the probability distribution would come out to be biased ok.

So, the next lecture, we will see how keeping constraints in your distribution will lead to probability distributions that are biased ok. So, we will carry out from here and meet in the next class and discuss things further.