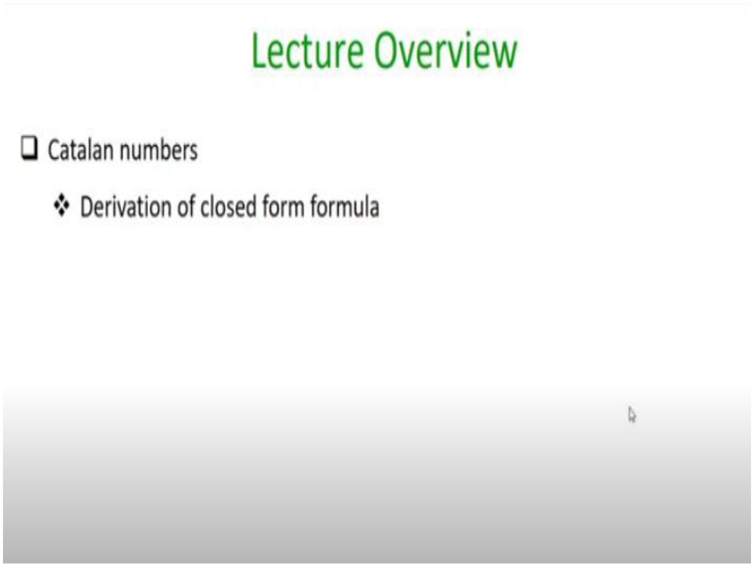
Discrete Mathematics

Prof. Ashish Choudhury
Department of Mathematics and Statistics
International Institute of Information Technology, Bangalore

Lecture -41
Catalan Numbers- Derivation of Closed Form Formula

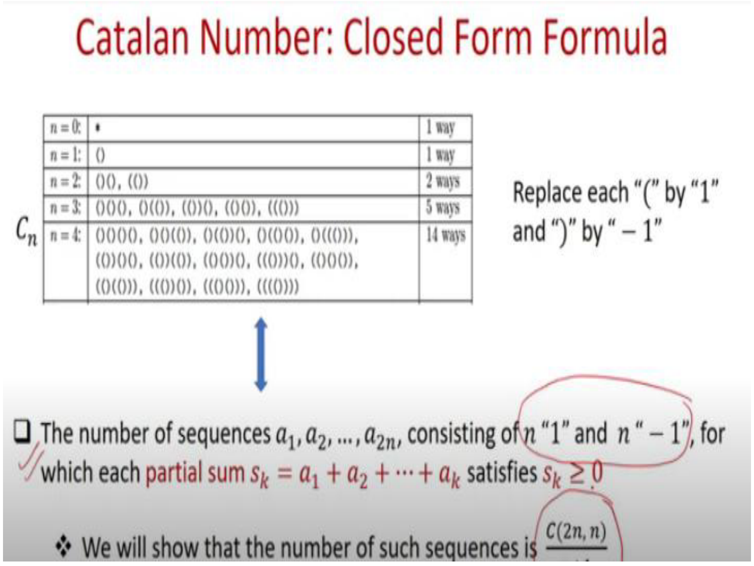
So, hello everyone. Welcome to this lecture, just a quick recap.

(Refer Slide Time: 00:26)



In the last lecture we discussed various problems whose solutions constitute Catalan numbers. So, in this lecture we will continue our discussion on Catalan numbers and we will derive a closed form formula for the recurrence relation for Catalan numbers.

(Refer Slide Time: 00:43)

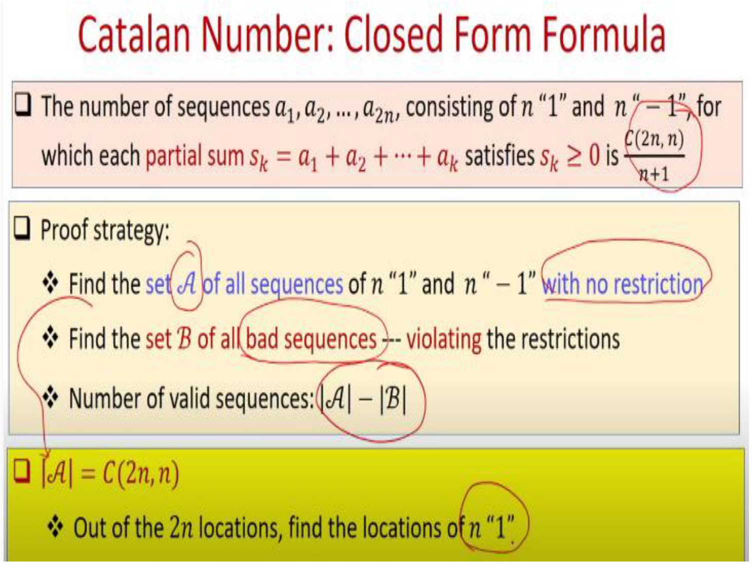


So, we recall from the last lecture that we discussed two problems whose solutions are the Catalan numbers. The first problem was that of coming up with the number of strings with n pairs of opening and closing parenthesis where a string is called valid, if whenever we scan the string from left to right then at any point of time or at any position in the string the number of; each instance of an opening parenthesis has a closing parenthesis.

And the second problem that we saw in the last lecture is that of coming up with a number of sequences consisting of n number of 1s and n number of –1s such that if we scan the string from the first position to the last position then each partial sum should be greater than equal to 0. And we saw a bijection between the number of sequences of n 1s and n –1s where each partial sum is greater than equal to 0.

And we saw a bijection between the set of sequences consisting of n 1s and n –1s where each partial sum is greater than equal to 0 and the set of all strings all valid strings with n pairs of opening and closing parenthesis. So, in this lecture, we will consider the set of sequences of n 1s and n –1s where each partial sum is greater than equal to 0 and we will show that the number of such sequences of length n of length 2n is this quantity namely $\frac{C(2n,n)}{n+1}$. That will show that the closed form formula for the Catalan number is $\frac{C(2n,n)}{n+1}$.

(Refer Slide Time: 02:48)



So, this is the statement which we want to prove. We want to prove that the number of sequences consisting of n 1s and n −1s, where in each sequence the partial sum at any position is greater than equal to 0 is $\frac{C(2n,n)}{n+1}$. So, for that the proof strategy will be the following. We will first find out the cardinality or the number of sequences consisting of n 1s and n −1s without any restriction.

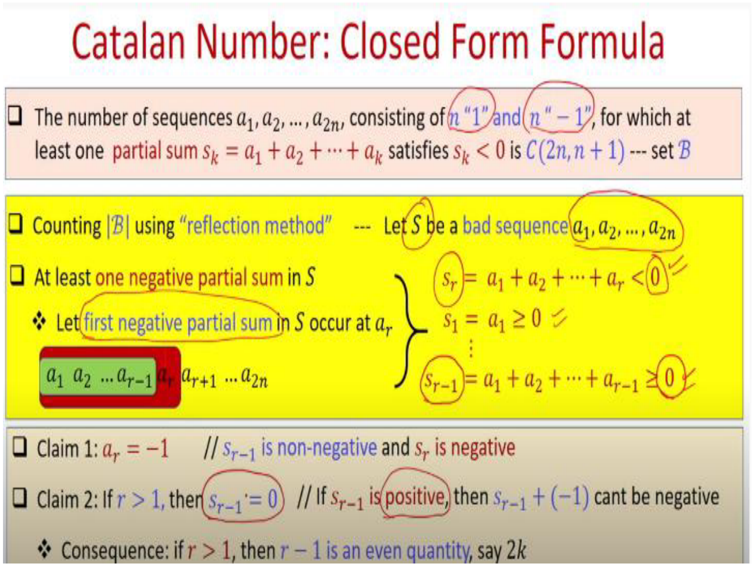
So, that means let A denote the set of sequences of n 1s and n −1s with no restriction. Then the set A has all the sequences where the partial sums are greater than equal to 0, as well as it has all the sequences where the partial sum at every k may not be greater than equal to 0. And then we will find out the set B of all bad sequences and by bad sequences I mean the sequences consisting of n number of 1s and n number of −1s which violate the restrictions.

And of course then it is easy to see that the required value or value of the required number of sequences of n 1s and n −1s where the partial sums are greater than equal to 0 will be the difference of the cardinality of the A set and the B set. So, that is the proof strategy. So, it is easy to see that the cardinality of the A set is C(2n, n). This is because, what is the set A? It is the set of all sequences with n number of 1s and n number of −1s where we do not put any restriction whatsoever over the partial sums in the sequences.

So, any sequence in this set will have n number of 1s and n number of –1s. So, it is easy to see that the cardinality of A is nothing but the number of ways in which we can find out n locations out of 2n locations where we can put the 1s. Because once we find out the n locations where we can put the 1 the remaining locations are of course has to be occupied by –1.

So, that is why the cardinality of the set A is C(2n, n). Now what we will show is that the cardinality of the set B is C(2n, n+1) and if we subtract the cardinality of B from the cardinality of A then we will get our required answer.

(Refer Slide Time: 05:38)



So, now for the rest of our discussion our focus will be to find out the cardinality of the set of bad sequences and what is a bad sequence? A sequence is a bad sequence if it consists of n number of 1s, n number of –1s such that in such sequence there is at least one occurrence of a partial negative sum. That means if I parse the string from a_1 to a_{2n} at least at some position k, some index k the values are such that if I just take the sum of a_1 to a_k then the partial sum is negative.

There might be multiple positions or multiple such indices k in that bad sequence but at least one such bad index or the index case is there. So, that is the definition of an invalid sequence. So, what we will do is we will introduce a very nice method called as reflection method or why it is called reflection method will be clear very soon. So, we will find the cardinality of the set B using the reflection method. So, for that let us consider an arbitrary bad sequence and we know that this bad sequence has n number of 1s and n number of -1s and at least one partial negative sum, where exactly the partial negative sum is appearing we do not know.

But we know that this is a bad sequence and I call this bad sequence as S. So, let r be the index or let r be the index at which the first negative partial sum occurs in the sequence S. So, that means if the values in the bad sequences are a_1 to a_{2n} then it is at the index r that the first instance of partial negative sum occurs. So, that means pictorially you can imagine that if I take the sum s_r which denotes the summation of the values a_1 to a_r it is negative.

And if I consider all other partial sums up to the $r - 1$th positions namely the partial sum s_1, the partial sum s_2 and partial sum s_{r-1}, all of them are greater than equal to 0. This is because of our assumption that the index r is the index where the first negative partial sum is occurring in the bad sequence S. So, now we will do the following. We will derive or conclude some properties regarding the values that are there in our bad sequence and based on that we will complete our proof.

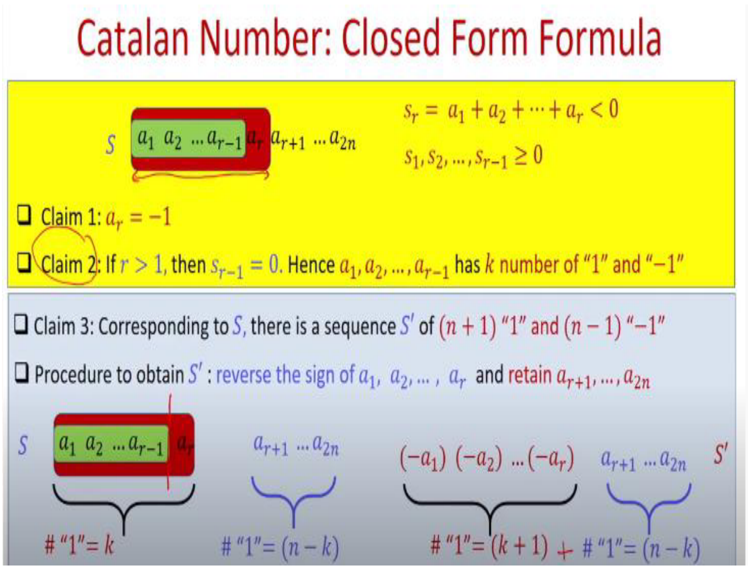
So, our first claim is that in this bad sequence S the value at the position r is definitely -1. And this is because as per our assumption the partial sum namely the summation of the first $r - 1$ values is greater than equal to 0. So, that means that if a_r; so remember by the way that at rth position we can have either $+ 1$ or -1. So, if s_{r-1} is greater than equal to 0 and if my rth value the number at rth position is $+ 1$ then definitely s_r will also be positive.

But that goes against the assumption that s_r namely the partial sum at rth position is negative. This is an easy claim. The second claim is that if the index r is greater than one then the partial sum till the $r - 1$th position is 0. And again this is because of the fact that r is the index, where the first negative partial sum is occurring. So, if r is equal to 1 definitely this claim is not true because $r - 1$ is not there if r is equal to 1.

But if r is greater than 1 then definitely I know that s_{r-1} is equal to 0. Because if r – 1 would have been positive say + 1, + 2 or + 3 or + 4 even if you take the least positive value namely + 1. That means the summation of the first r – 1 value is say + 1 or greater than + 1, then even if we are putting –1 at the rth position that positive value added with –1 would have given the partial sum s_r to be 0 or more than 0.

But that goes against the assumption that s_r or the partial sum at rth position is negative. This shows that if r is greater than 1 then r – 1 is an even quantity. Because if the partial sum at r – 1th position is 0 that means by the time we have reached r – 1th position we have encountered equal number of 1s and –1s. So, that is why r – 1 will be an even quantity and it will have and in the first r – 1 positions we would have encountered k number of 1s and k number of –1s where k is greater than equal to 1.

(Refer Slide Time: 11:47)



So, I have retained the summary of the claims regarding the various values that we have in the bad sequence S. Now what we are going to do is corresponding to the bad sequence S; so this is our bad sequence S. We will find another sequence S’ which will have n + 1 number of 1s and n –1 number of –1s. Namely the number of 1s will be two more than the number of –1s. Remember in the bad sequence S we had an equal number of 1s and –1s.

But now we are going to define a sequence S' corresponding to the sequence S which will have two more 1s than the number of -1s. And this is done as follows; So, let me first demonstrate how exactly we construct the sequence S' corresponding to the sequence S for the case where n is equal to 2 and then we will see the general method for any n.

So, for n equal to 2 we have 4 possible bad sequence S consisting of 2 1s and 2 -1s. And now for each of this bad sequence S I have highlighted the first occurrence of partial negative sum in that sequence. So, for instance if I consider the first sequence my r is equal to 1. Because at r equal to 1 I have an occurrence of partial negative sum. For the second string also my r is equal to 1, for my third bad sequence also r is equal to 1.

But for my fourth bad sequence r is equal to 3, because if I consider the partial sums s_1 for this bad sequence then it is positive. If I consider the partial sum at position 2 then it is 0, still it is not negative. And only when I consider the partial sum at position 3 it becomes negative. So, that is why r is equal to 3 for the fourth bad sequence. Now the corresponding string S' for each of these bad sequences S is as follows.

If you see here what I have done basically is for each of the bad sequence S' the remaining portion of that bad sequence which is occurring after the first instance of the partial negative sum is retained as it is. So, we had this -1, 1, 1 they are retained as it is. And whatever partial sequence we had here the first occurrence of partial negative sum is occurring, I am just converting each -1 to + 1 and each + 1 to -1.

Well in this case there is only one value in the sequence. So, that -1 gets converted into + 1. In the same way for the second bad sequence the remaining portion of the sequence after the first occurrence of partial negative sum is retained as it is and then in the sequence which has the first occurrence of partial negative sum we replace -1 to + 1 and so on. If you take the third sequence then this unhighlighted portion remains as it is.

And now you see the highlighted portion, namely the sequence which has the first partial negative sum; there we convert each 1 to –1 and vice versa. So, if this is my S, this is the S'. If this is my S, this is my S'. This is my S and this is my S' and for this fourth S this is my S'. And now you can see in S' we have the number of 1s exceeding the number of –1s by 2.

So, now let us see the general process. The process to obtain S' from S is as follows. We reverse the sign of a_1 to a_r. So, remember as per our assumption r is the first index such that the partial sum s_r is negative. So, what we do is we convert a_1 to $-a_1$, a_2 to $-a_2$ and so on and the remaining portion of the bad sequence S is retained as it is, namely a_{r+1} is retained as it is, a_{r+2} is retained as it is and so on in S'.

So, now let us count the number of 1s and –1s in the sequence S'. So, if I consider the portion where the occurrence of partial negative sum is there namely if I focus on the portion of the sequence till the rth position then I know that the number of –1s is more than the number of 1s by one position. This is because as per our claim 2, till the r –1th position the sequence S has equal number of 1s and –1s , namely k number of 1s and –1s.

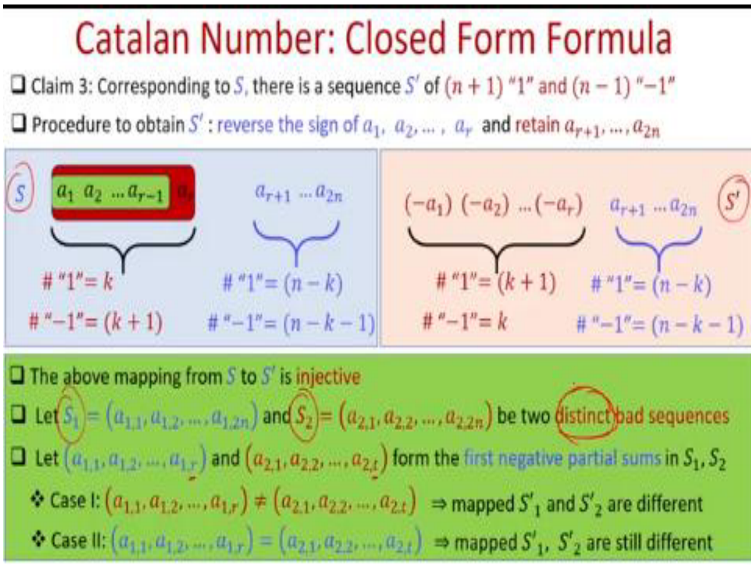
And since at rth position the partial sum becomes negative. That is because at the rth position we have a –1. So, that is why we have one more –1 compared to the number of 1s till the rth position and since my overall sequence S has n number of 1s and n number of –1s that means in the remaining portion of the sequence S the number of 1s will be n - k and a number of –1s will be n - k + 1 which is this one.

So, now what we can say about the number of 1s and –1s in S'. So, the number of 1s and the number of –1s in this half; in this portion of S' is same as the number of 1s and number of –1s in this portion of S. No change in the number of 1s and –1s in this portion of S', the later portion of S', the remaining portion. Whereas if I consider the first portion of S' namely which is obtained by reversing the signs of a_1 to a_r then the number of 1s and –1s are as follows.

The number of 1s in this portion, it will be the same as the number of –1s. Because due to the reversal of the signs all the –1s they will be converted into 1s. And due to the same reason; due to the reversal of the 1s to –1s and –1s to 1s the number of –1s in this portion of S’ will be the same as the number of 1s in the highlighted portion of S which is k.

So, that tells you that if I sum the total number of 1s in S’ then it will be k + 1 + n - k which is n + 1. And if I find the number of –1s it will be k + n - k –1 which is the same as n –1. So, that means the number of 1s is 2 more than the number of –1s in S’.

(Refer Slide Time: 20:56)



So, that and now you might have understood why we are calling the method of finding the cardinality of the set B as the reflection method. If you see closely this process of reversing the sign of a_1 to a_r, it is like reflecting the 1s to –1s and –1s to +1s. So, that is why the method is called as the reflection method. So, what we have done till now is we have converted, we have found a new sequence S’ for each bad sequence S. My claim is this process of getting the sequence S’ from the sequence S is an injective process.

That means the S’ that are obtained from S are obtained in an injective fashion. That means it cannot be the case that there are two bad sequences S’ with equal number of 1s and –1s and each of which has an occurrence of partial negative sum such that if we find the corresponding sequences S’ for these two bad sequences they are the same. How do we prove that? It is very simple. So, imagine S_1 and S_2 are two distinct bad sequences.

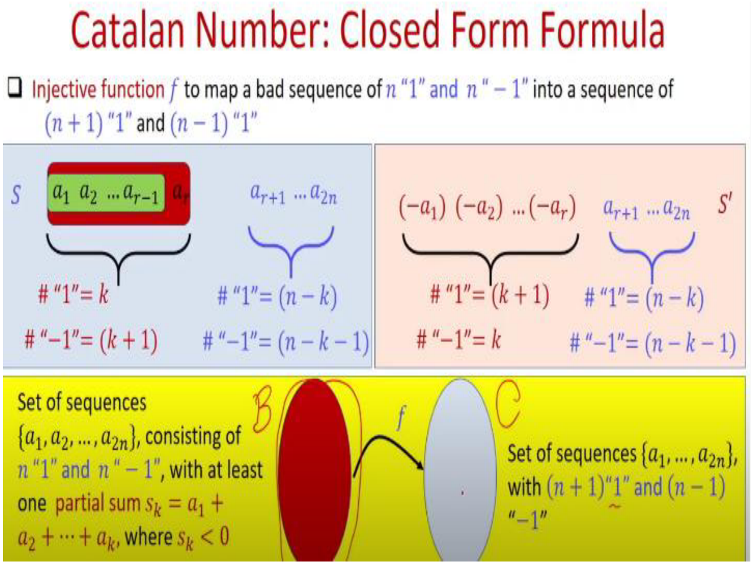
So, each of them has equal n number of 1s and –1s and each of them has an occurrence of negative sum somewhere. So, what I am doing here is that I focus on the first index in the sequence S_1 and the first index in the sequence S_2 where we have an occurrence of partial negative sum. So, let for S_1 the index r is the first index where we have an occurrence of partial negative sum and in the same way for sequence S_2, let the first partial negative sum occur at the position t. We do not know whether our index r is the same as index t or not. So, there are two possible cases. If the portion of S_1 and S_2 with respect to the rth index and the tth index they could be either same or they could be either different. So, let us take the case one. If the subportion of S_1 till the rth position and the subportion of S_2 till the tth position they are different then I do not care what is the remaining portion of S_1 and S_2.

Because of the reflection method the corresponding sequences S_1' and S_2' which are obtained by the reflection method they will be different. Because in the reflection method in S_1' the signs of $a_{1,1}$, $a_{1,2}$, $a_{1,r}$ will be reversed and in S_2' the signs of $a_{1,1}$, $a_{2,2}$, $a_{2,t}$ will be reversed. Now since in S and S' the portion till the rth position and the portion till the tth position in S_1 and S_2 were different, if I reverse their signs then I know that till the rth position and the tth position the portion till the rth position and the portion till the tth position in the reflected strings in S_1' and S_2' also will be different. Whereas case 2 is when r is equal to t basically. So, if that is the case then in S_1' and S_2' the values till rth position and the values till the tth position also will be the same because of the reversal of the sign. But then what I can say is that in S_1 and S_2 the remaining portions of the strings were different.

This is because as per our assumption S_1 and S_2 are two distinct strings. So, if till the rth position and the tth position S_1 and S_2 were respectively the same, since S_1 and S_2 are overall different. It means that the remaining portion of S_1 that means the portion of S_1 from (r + 1)th position to the (2n)th position and the portion of S_2 from the (t + 1)th position to the (2n)th position they are different. Because if they are also the same that means S_1 and S_2 are the same string. But that goes against the assumption that S_1 and S_2 are two distinct strings.

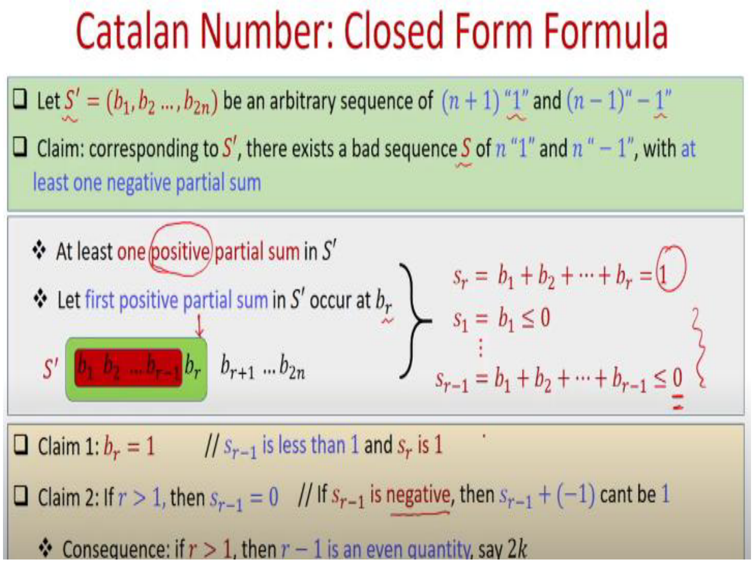
So, in this case even if the reflected portions in S_1' and S_2' are same remaining portion of S_1' and S_2' which are copied as it is from S_1 and S_2 respectively they will be different. So, that shows that the above process of a mapping from S to S' is an injective mapping.

(Refer Slide Time: 26:08)



So, if I consider this red circle which is the set of all bad sequences; so this is our set B as per our construction. Namely it has all sequences with equal number of 1s and –1s which have at least one occurrence of partial negative sum. And we have another set which I call as say the set C which is the set of all sequences where the number of 1s is two more than the number of –1s.

And we have established an injective mapping namely the reflection method from the set B to the set C. Now what we will prove is that the above process of converting any bad sequence S to a corresponding sequence S', that mapping is also a surjective mapping. And that will show that the cardinality of the set B and the cardinality of the set C are the same. **(Refer Slide Time: 27:16)**



So, for proving that our mapping f is surjective mapping what we have to do is, we have to take any arbitrary sequence in the set C and we have to show corresponding to that there is a bad sequence. So, let us do that. So, imagine I take an arbitrary bad sequence S' where the number of 1s is two more than the number of -1s. Now corresponding to that, our goal will be to show the existence of a bad sequence which has equal number of 1s and -1s and which has at least one negative partial sum.

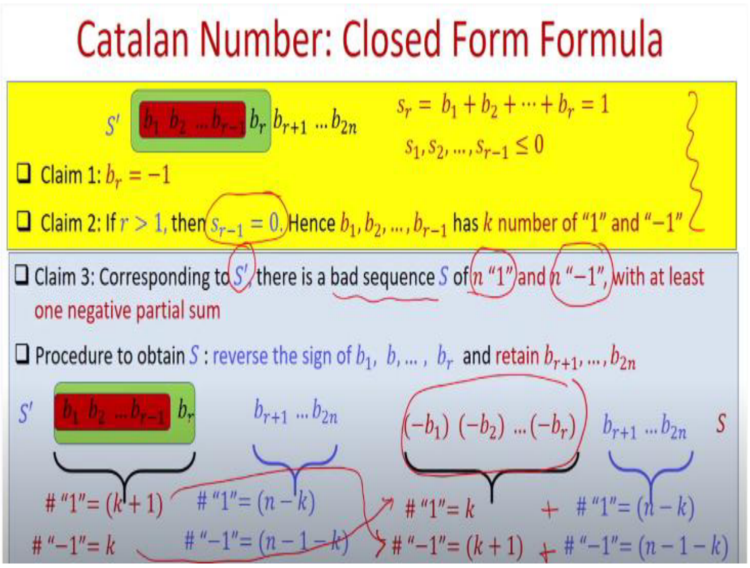
So, for that intuitively what we will do is we will just reverse the process that we followed for getting the string S' from the string S. So, what we can say about the string S' is that since it has more number of 1s than -1s, it has number of 1s is two more than number of -1s, definitely it has one positive partial sum not negative partial sum. So, there could be multiple positions in S' where we have positive partial sum.

But let us focus on the first occurrence of positive partial sum in S' and suppose it occurs at the rth position. So, again pictorially you can imagine that if I scan the sequence S' then at rth position if I take the sum of all the values till the rth position then the sum becomes 1. But till the (r – 1)th position if I take the partial sums they were either 0 or negative. Then again we make similar claims as we did when we converted the sequence S to S'.

So, we can say that the value at rth position in S' will be definitely 1. It cannot be –1, because your s_{r-1} namely the partial sum till the (r – 1)th position was either 0 or negative. And if at the rth position also we put a –1 then we get that at rth position the partial sum is still 0 or negative. But that goes against the assumption that the partial sum at rth position is positive namely 1.

Similarly we can claim here that if the index r is greater than 1 then the partial sum at (r – 1)th position is exactly 0. It cannot be negative because if it would have been negative then if in that negative sum if we add a 1, namely even if we put 1 at rth position then the partial sum at rth position would have stayed 0 or negative it cannot become 1. That shows that if r is greater than 1 then till (r – 1)th position we have equal number of 1s and –1s, say k number of 1s and k number of –1s.

(Refer Slide Time: 30:39)



So, these are the summary of whatever we have concluded till now about S'. Now what we do is we will show a method we will follow the reflection method and what we will do is that corresponding to S' we will show a bad sequence consisting of equal number of 1s and –1s and which will have one negative partial sum. I stress in S' we do not have equal number of 1s and –1s.

But we will convert S' to another sequence S which will have equal number of 1s and –1s and at least one negative partial sum. And the idea is just to do the reflection method here. We just reverse the sign of the numbers b_1 to b_r and retain the remaining portion of the sequence S' as it is. So, b_1 gets converted into - b_1. So, if it is + 1 it becomes –1, if it is –1 it becomes + 1.

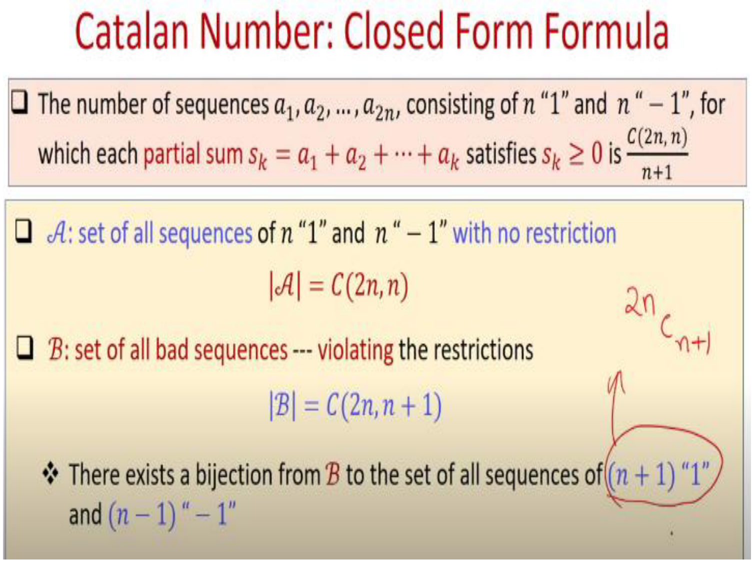
Similarly b_2 becomes $-b_2$ and like that b_r becomes $-b_r$. The remaining portion of S' is left untouched in S and then again we can use similar counting argument to find out the number of 1s and –1s in S. So, we know that in S' in the rth position the number of 1s is exactly one more than the number of –1s. This is because the partial sum at rth position is exactly 1 and the partial sum till the (r – 1)th position is 0.

And since the number of 1s overall in S' is n + 1. That means the number of 1s in the remaining portion will be n + 1 – (k + 1) which is same as this n - k and the number of –1s will be; overall we will have n – 1 number of –1s in S'. We already had k of them till the rth position. So, in the remaining portion this will be the number of –1s : (n – 1 – k). So, these statistics regarding the number of 1s and –1s will be carried over in S as well.

And what we can say about the number of 1s and –1s in S? Well the number of –1s will now become the same as the number of 1s. Because each + 1 has been converted into – 1, whereas the number of –1s will now become the number of 1s. Because each –1 has been converted into + 1. So, that shows that the number of 1s overall will be n - k - k which is n and the number of –1s will be k + 1 + n –1 - k which is again n.

And it is easy to see that if we take the partial sum till the rth position it will be negative. Because till the rth position in S' the sum was positive and the sign of every + 1 and –1 has been reversed. Because of that the partial sum till the rth position S will now become negative. So, we have shown that the mapping is a surjective mapping as well.

(Refer Slide Time: 34:12)



So, going back to the proof of finding the cardinality of the set of bad sequences, we have the set A which is the set of all sequences of equal number of 1s and –1s, without any restriction. We know there are C(2n, n) such strings and we just established that the number of bad sequences which violates the restriction will be C(2n, n + 1). Because we just established a bijection from the set of all bad sequences with equal number of 1s and –1s, and violating the restrictions to the set C of all sequences which has n + 1 number of 1s and n – 1 number of –1s. And the cardinality of this set will be C(2n, n + 1) because it is equivalent to saying that out of 2n positions we have to identify the n + 1 positions where +1s will be there. Automatically the remaining positions will be occupied by –1s. And that shows that the number of actual sequences the valid sequences which we are interested to find out is the difference between the cardinality of the A set and B set.

And if we find the difference of the cardinality of the A set and B set we get the result of the nth Catalan number which is $\frac{C(2n,n)}{n+1}$. So, that brings me to the end of this lecture. Just to summarize in this lecture we extensively derived the closed form formula for the Catalan number and for that we introduced the reflection method, thank you.