**Natural Language Processing**
**Prof. Pawan Goyal**
**Department of Computer Science and Engineering**
**Indian Institute of Technology, Kharagpur**

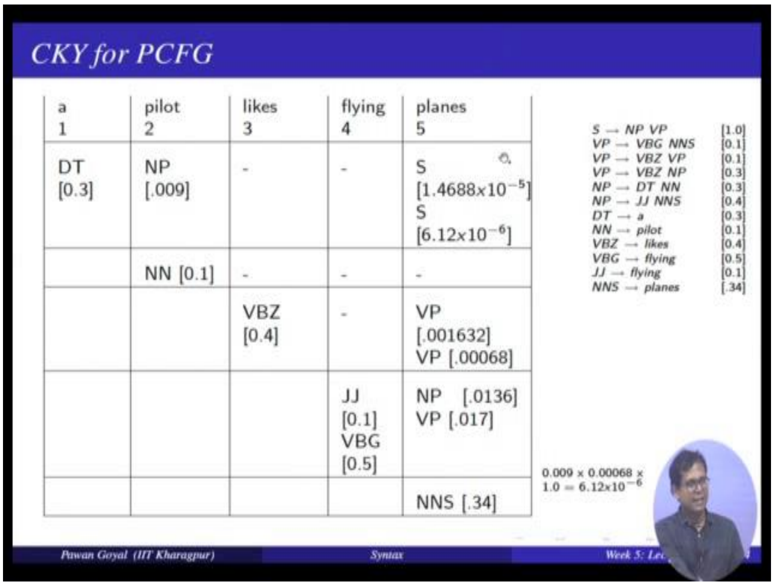
# Lecture - 24
# PCFGs – Inside-Outside Probabilities

Hello everyone, welcome back to the 4th lecture of this week. In last lecture, we had discussed what is PCFGs and we finally, came up with what are the interesting problems that that we would like to answer. So, for example, what is the most likely parse for a given sentence as per my PCFG grammar and what is the probability of the sentence as per the grammar? Finally, we also wanted to see how do I learn my PCFG rule probabilities. If you remember, this was the only difference between context free grammar and the probabilities version is that all the production rules here have a probability value.

So, how do I learn these values? So, this is what we will be covering in this lecture and in the next lecture and we will be using a specific concept called inside outside probabilities and this will be very analogous to what you study in forward backward algorithm in terms of learning parameters for HMM. So, we will go to that.

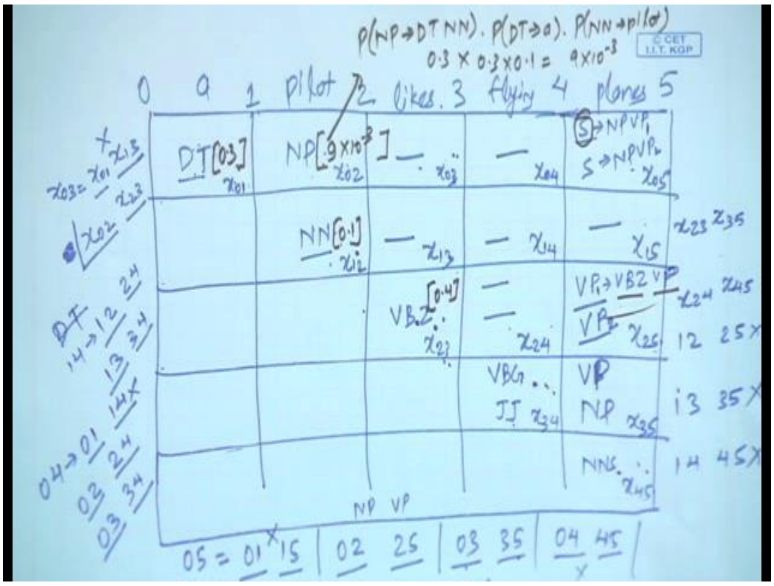
Starting with the first problem, how do I find out the most likely parse for a sentence? And this can be solved simply if we use the CKY algorithm for PCFG. So, from the example that we did in the last class, I hope you understand now how to use CKY algorithm for the context free grammar. So, today I will show, how do you extend that for PCFG?

(Refer Slide Time: 01:43)




Let us take an example. So, we have this PCFG. So, what you are seeing here with each rule, you have a rule probability as well and you can verify that all the probabilities starting from a given right hand side will add up to 1. So, now, I want to find out what is the most likely parse for the sentence, the same sentence the pilot likes flying planes. So, for that I need to find out what is the probability with which I can find a non terminal as in the final position for deriving this whole sentence 0 to 5. So, we proceed nearly in the same manner as we did in the in the case of CKY. So, what I will do? I just take the same example, how we solve CKY in addition and what are the things that you need to do different here.

(Refer Slide Time: 02:50)



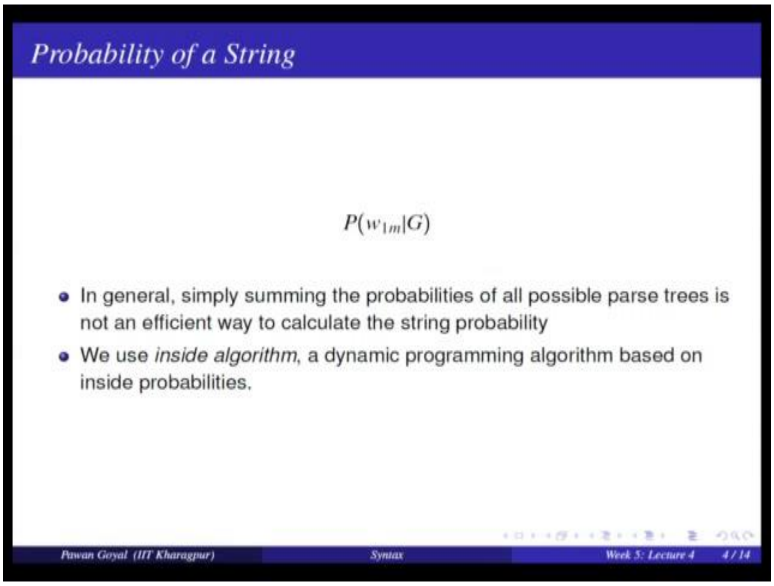

This is what we had seen last time. So, if we have to build, if we have to find out, what is the; whether the sentence can derive this whole thing, I will fill it starting from 0 1, 1 2, 2 3, 3 4, 4 5, then I will go further. So, now, I will just tell, what will differ in the case of PCFG. So, now, with each non terminal, there will be probability with which it can derive that particular terminal or a set of terminals. So, it is starting from the first element. So, this says that the non terminal DA derives a. Now I will write down here, what is the probability? So, if you go to your grammar, the probability DT gives me is 0.3. So, I will put 0.3 here.

Similarly, for this element, what is the probability that NN gives me pilot and if I look at my grammar, this is 0.1, similarly here what is the probability that VBZ derives likes this gives me 0.4 and like that I will fill all these entries now this is ok.

Now, what I want to show you is that how do I fill the next entry that is 0 2, what is the probability with which the non terminal NP derives the sequence a pilot? Now if you remember, a single non terminal can derive 2 non terminals; 2 terminals only via 2 non terminals. So, it has to first derive x 0 1, x 1 2 then each of these will individually derive a N pilot. So, how do I fill in this probability value? So, this would be probability that NP derives DT NN times probability DT derives a times probability, NN derives pilot. This probability I can get from my PCFG. So, this probability if I see in my grammar, this is 0.3, this probability I have already filled in is 0.3 and this probability also I have filled in this is 0.1.

The probability that NP derives a pilot here comes out to be 9 times 10 to the power minus 3 and that is what I fill in here and in the same manner you will fill all the entries if it is 5, it means the probability will be 0 and if there is a non terminal here; that means, there will be certain probabilities and this you can get by finding out probability individually for each of this and multiplying the probability of this rule. So, that way you can incrementally build all the stable from in a bottom up fashion we are starting from the first element then in the next one and so on. So, remember nearly same as we did in the last class only the probabilities and the multifunctional probabilities will change, suppose you do that for this example, this is what you will find out. So, these are the rule probabilities that you will come up with. So, I will encourage all of you that you should try this on your own and see that you can get the same set of values that have been shown in the slide.

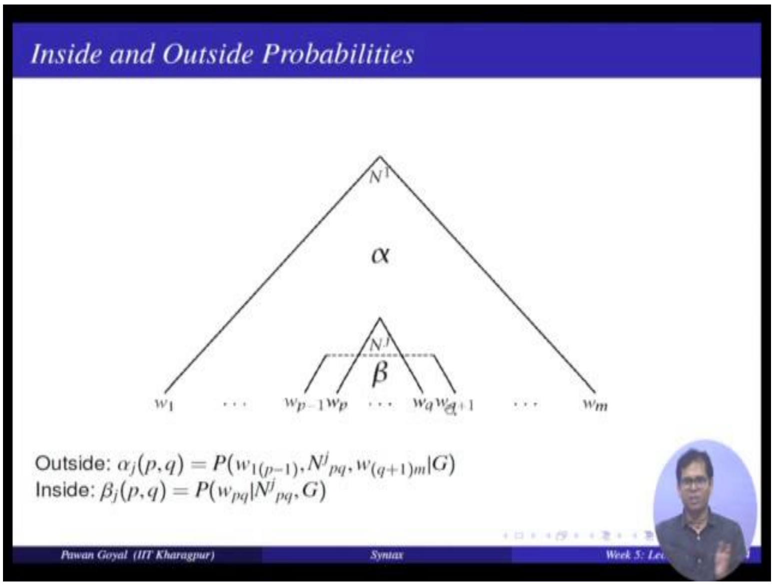
(Refer Slide Time: 06:35)




Now coming to the next question, I am given the PCFG as my grammar, how do I find out, what is the probability of the sentence. So, one simple solution you can always find is that I find out all the possible parse trees, find the probabilities for each of the parse trees and add those up that is one way, but this you might be inefficient if there are very huge number of parses trees, if you remember one of the introductory slides, we said that as you increase the number of words in the sentence or the number of phrases, the number of possible parses increase roughly exponentially, there was something like Peclet number with this their group and it can go to 100 plus parses.

So, you do not want to sum this get the probability individually and then do addition you want to do something more efficient than that and how do we do this and that is why we use this inside algorithm that is one of the main topics of this lecture, so what is the inside algorithm. And this is some dynamic programming algorithm based on the concept of inside outside probabilities.

(Refer Slide Time: 07:41)



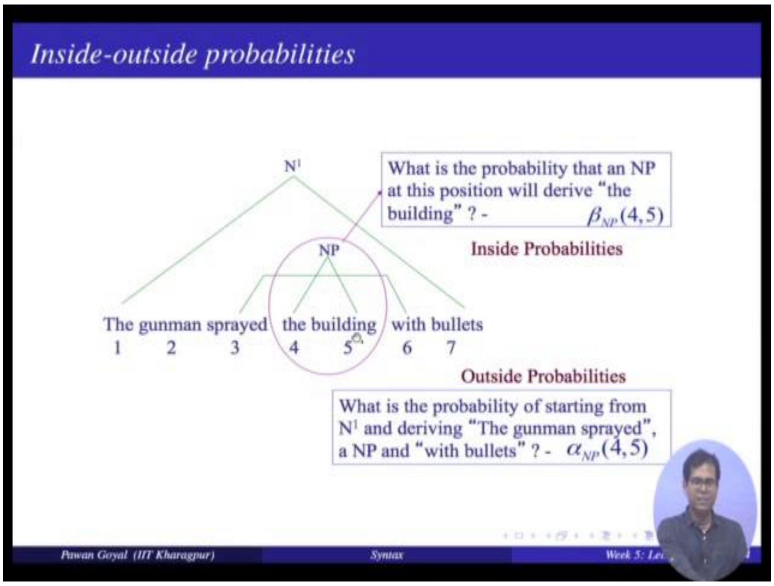

Now I will try to introduce, what is that inside outside probability. So, this is very centre to understanding this lecture as well as the next lecture. So, let us try to understand; what is the concept and just so that it helps you, this is very very similar to the concept of forward backward algorithm that we did in the case of HMFs. So, what is inside outside probabilities? So, I can this is parameterized for certain node of my tree. So, what you see in this figure, I have a sentence with words W 1 to W m and N 1 is the non terminal you can think of it as the sentence non terminal that derives the whole sequence of words using my grammar.

Now, I am putting a parameter here that is N j; N j is one particular non terminal. It can be N p v p or so on. So, what I am saying here assume that there is a non terminal N j that derives the sequence W p to W q, again p and q are arbitrary, you can take any p starting from 1 to m minus 1. So, what I am parameterized here is that this N j derives the sequence W p to W q. So, now, with respect to that I am defining, what is my inside probabilities and what is my outside probabilities. So, you can as such you can guess this is something outside of that and this is something inside of that.

How do I define my inside and outside probabilities? So, outside probability is that it is starting from N 1, I can derive the sequence W 1 to W p minus 1, I can derive this non terminal W sorry, N j and the sequence W q plus 1 to W m, this is my outside probability and inside probability is given this N j the probability that I can derive W p to W q as per my grammar and it will then follow the same sort of ideas that if I can multiply inside and outside probability that will give me something like the probability of the sentence parameterized by this particular j th non terminal. So, formally that is how I can write the outside and inside probabilities.

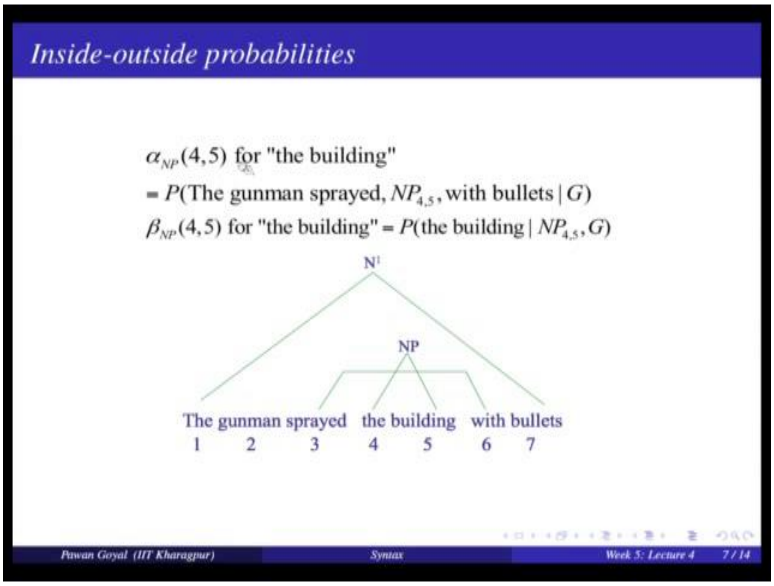
Outside probability - alpha j p q that is standard format which we can write p q denotes the sequence W p W q on which this parameterized and j denotes the particular non terminal N j here. So, the outside probabilities, probability of generating W 1 to W p minus 1 s and j p q and W q plus 1 to m given by grammar and inside probabilities probability of generating W p to W q given N j, so this is also parameterized by N j p q because it is arriving W p to W q and in my grammar. So, that is why I define my inside and outside probabilities.

(Refer Slide Time: 10:42)



Now, let us take a simple example, I have the sentence, the gunman sprayed the building with bullets and here your p q is 4 and 5. So, you are parameterization with respect to this sequence the building. So, accordingly this will be particular non terminal that is NP that is deriving the building. So, now, what do we mean by inside probability? So, inside probabilities; what is the probability that this non terminal NP at this location derives the sequence the building, this is my inside probability.

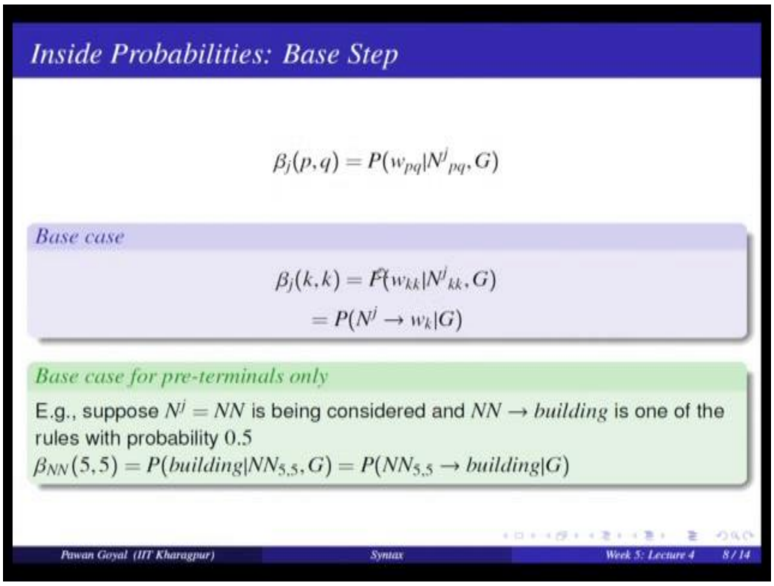
(Refer Slide Time: 11:38)




And what will be the outside probability here that is starting from N 1, I can derive the sequence, the gunman is sprayed this NP and with bullets this is my outside probability. So, this is what we have defined the outside probability alpha NP, 4 5 would be the probability of this sequence the gunman is sprayed NP 4 5 and with bullets given my grammar and the inside probability beta NP 4 5 would be the probability for deriving the building given NP 4 5 and grammar.

Now, this is the definition of inside and outside probabilities. Now how do I actually compute this inside and outside probability given this sentence and suppose our grammar is also given to me how do I compute that?

(Refer Slide Time: 12:20)

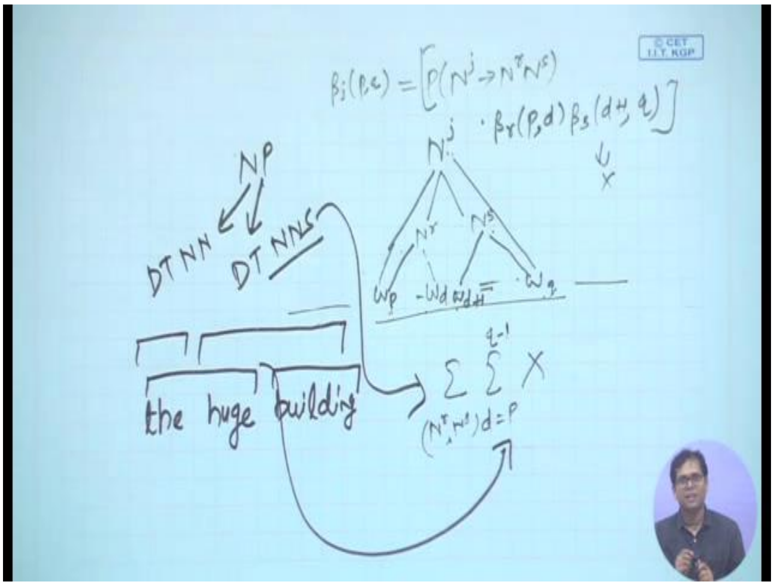



For inside probability, so let us see this is the definition of inside probability beta j p q is probability of deriving the sequence W p to W q given non terminal N j p q in my grammar now. So, I will be deriving both these probabilities in iterative manner. So, what will be the base case here? So, the base case here if you see would be when p and q are same that is I am deriving a single terminal from a non terminal and that you can find out very easily if your grammar is in the Chomsky normal form. So, what is the particular non terminal that derives this term? So, this would be the base case for deriving a single term base and then I will go in a bottom up fashion derive the inside probabilities for higher and higher sequences. So, the base case here is when P is equal to q. So, beta j k k would be probability of W k k given N j k k and G, Now W k k is a single word sequence from k to k and I can immediately find out what is the role in my grammar that derives this word W k k and I put this rule probability here and this will give me the this probability beta j k k.

Now, as in example, suppose here my word is building. So, this is the 5th word. So, I am computing beta for 5 5 and suppose I take N j is equal to NN that works in non terminal derives this word. So, you can compute beta NN 5 5 as the probability with which the non terminal NN can derive this word building and this is easily given by my grammar. So, this is my base case.

Now, what will be the inductive case? So, I have to go bottom up. So, let us take a generic beta j p q how do I derive it in terms of the smaller values. So, let us try to do that.

(Refer Slide Time: 14:32)

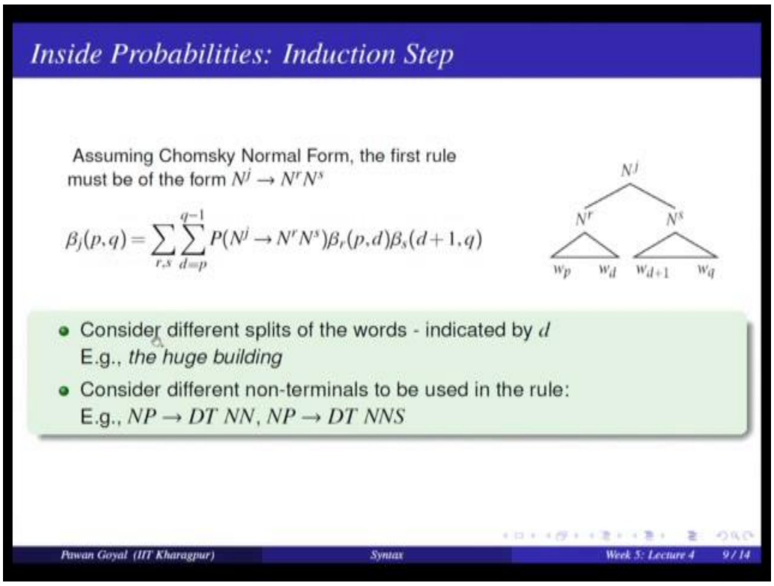



If I am given I have to compute beta j p q, what is that mean? It means that I have a non terminal N j and that is deriving a sequence W p up to W q and I have to compute this probability given N j the probability of the sequence W p to W q because I am conducting in a I am doing it in a bottom fashion and I have to use the proper lower values.

I will first say N j can give me some N r and N s, now N r will derive from W p to some W d. So, very generic case, so this can be any possible N r and N s that my N j can derive and again I can go from W p to any W d, so you can see what can d vary from to so, d can vary from P up to q minus 1 and this N s gives me W d plus 1 up to W q this is my recursive case. Now how do I write down this probability? So, I will say this is first the probability that N j derives the rule the non terminal the N r and N s.

Probability N j derives N r N s times this probability now here N r is given. So, what is the probability is that N r given N r; it derives from W p to W d again, you can compute this using. So, you can express this using inside probability that is beta r p d, similarly this 1 beta s d plus 1 to q, now this is for the particular variation that I have shown, but in general, d can vary from p to q minus 1 and they can be any possible N r and N s. So, I will have to write down this whole thing, let me call it X. So, I have to say this can vary from p to q minus 1 and all possible N r N s and I will put x here. So, this is my inductive step.

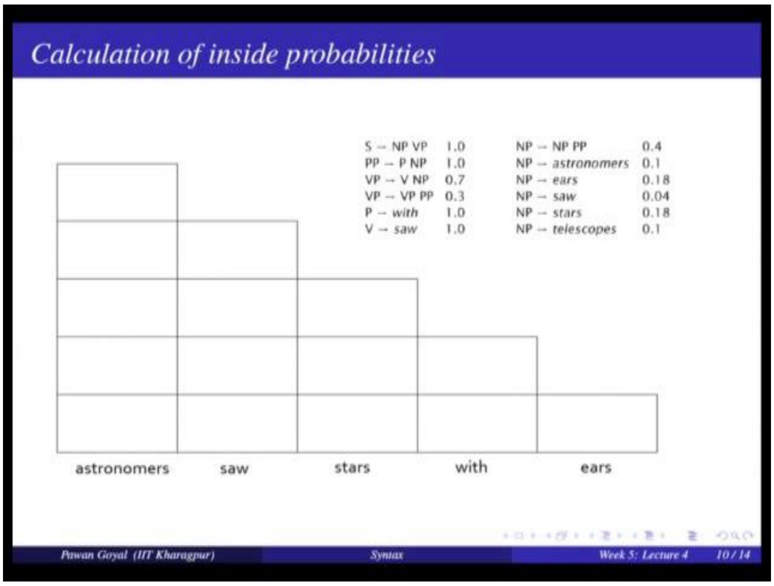
(Refer Slide Time: 17:13)




Now, what do I mean by parameterization over by all possible r s and all possible d. So, if you take the case of this simple phrase, the huge building. So, what do I mean by various d? So, if I have this is the huge building by variation of d, I mean whether I am composing it by the huge and building or the and huge building 2 possible variables of d and what do I mean by parameterization over all possible N r and N s. So, the particular non terminal that derives this whole thing can be support this NP, it can come via say NP gives me DT NN, this is 1 possibility, this is DT NN or there can make another rule, NP gives me DT NNS. This is another possibility. So, this is the parameterization over all possible N r N s and this is over out possible d’s.

Now we have seen that you can compute it for the only the terminals first and then the bottom up to compute it for other values.

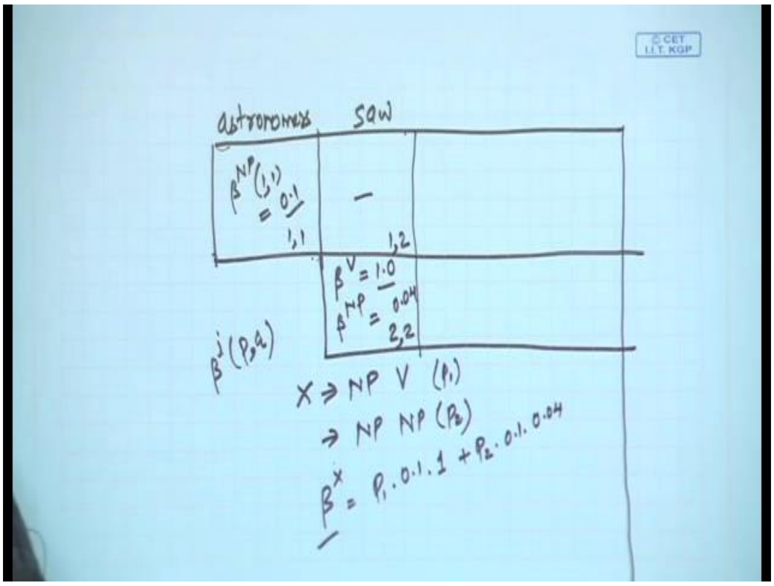
(Refer Slide Time: 18:40)




Now if you do a simple example, so this is again similar to what we have seen earlier. So, I am given this PCFG and this is my sentence and I want to compute the inside probabilities here. So, how do I go about out it and you looking at the table you can see that we will doing something very similar to what we did in the case of CKY algorithm.

But now we will focus on filling up what is the inside probability of deriving this from the particular non terminal. So, let us just try to filling some entries here.

(Refer Slide Time: 19:36)



Let us try only the first 2, this is astronomers and this is saw and I have to fill in the inside probabilities. So, this would be, so now, in the way we defined inside probability this would be 1 1, this would be 2 2 sequence from word 2 to word 2, this would be sequence from word 1 to word 2.

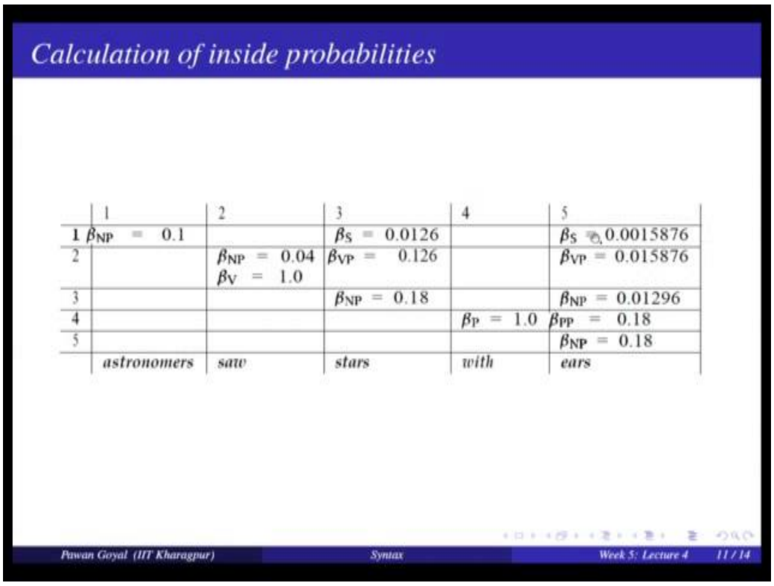
Now when I define beta j p q, so, here p and q are both 1. So, I have to find out what is the non terminal that derives these word astronomers, now if I see a grammar and these derive astronomers with the probability of 0.1. So, I can fill the inside probability here beta NP 1, 1; 1, 1, you need not to write and I am just writing here is equal to 0.1 because 1 1 is already defined by the particular cell where we are, similarly let us go to 2, 2; 2, 2, is saw if I see the grammar NP derives saw NP derives saw. So, there are 2 different. So, that is where you can see there are 2 different j that are possible here. So, one is beta V, another is beta NP and this has the probability of one and this has the probability of 0.04 using my grammar.

Now, suppose I have to fill this one 1 2, what is the particular non terminal that will derive this whole thing again similar to what we were doing in CKY is the something that derives NP and NP if I see my grammar nothing derives NP and NP in the right hand side. Now is there something that derives NP followed by V, again there is nothing here NP followed by V. So, this will come out to be 0. Suppose there was something that derives NP followed by V. So, I would simplify find out what is the non terminal that is deriving this and put the probability as this times this times the rule probability.

Similar to what we were doing in the case of CKY for probability context free grammars the only difference here is that if there are multiple splits for same non terminal, I will add all these. So, for example, the same non terminal so I am taking a hypothetical case, suppose the same non terminal X derives NP followed by V and also NP followed by NP if the probability of P 1 is the rule probability, this is a probability of P 2. So, I will put here beta x and this will be P 1 times 0.1 times 1 plus P 2 times 0.1 times 0.04 and that is why I fill all the entries.

I will take care of all the possible ways in which this particular non terminal, we can derive this.

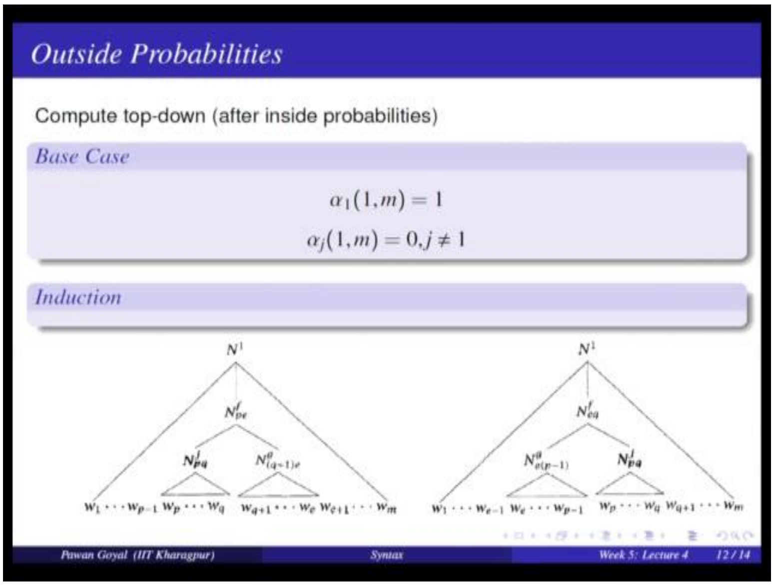
(Refer Slide Time: 23:41)




Again, so, if you keep on doing that this is what you will see. So, we had filled in this entry and this entry and this entry, but I will encourage you to fill in all the entries of this table and see if you can get to this particular term what is the probability of the sentence generating this whole thing. So, now, if you see this table you also get the answer for the second question is how do I find out the probability of my sentence use the same probabilities and compute beta as for 1 to m or 1 to 5 in this case that will give you the probability of this whole sentence. So, this was about inside probabilities.

Now, how do I compute the outside probabilities? So, if you remember, this is analogous to what we did in the case of forward backward algorithm. So, now, inside probabilities, we are computing bottom up and this will be computed top down and what will be the base case? Base case would be very very simple that is can I generate the whole sequence 1 to m and now because this is the grammatical sentence we are assuming. So, only the first non terminal that is sentence or you can write it as N 1 can derive this whole sentence nothing else can derive. So, that is why the base case will be alpha 11 to m is 1 if the so this j is 1 otherwise it will be 0.

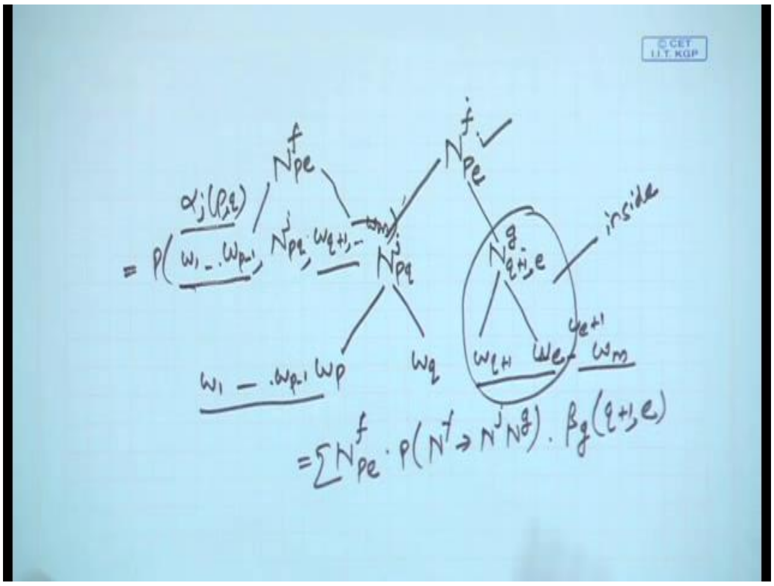
(Refer Slide Time: 25:06)




If I write down alpha 1, 1 to m will be 1 and alpha j 1 to m will be 0, if j is not equal to 1, if j is equal to 1 denotes the starting non terminal like s or N 1 whatever we were using the notation form. So, now, this is the base case, now what will be the inductive case. So, inductive case would be I have to compute it in a top down fashion. So, can you think of inductive case here?

Now what I am showing here what are the 2 possibilities of this inductive case. So, let me just show it quickly on the paper.

(Refer Slide Time: 25:54)

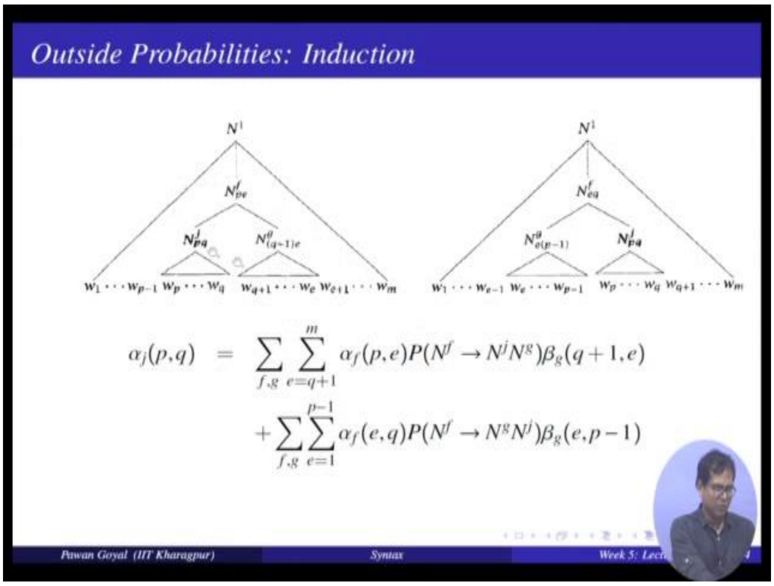


I have to compute alpha j p q and I am going top down, now what is the alpha j p q? I have a non terminal N j that derives W p to W q and everything else W 1 to W p minus 1 and W q plus 1 to W m is there in the sentence. Now I have to do it in a top down manner. So, I have to use the probability from the upper label. So, I would say what is the non terminal that derives this N j with some other non terminal and again this can be either left child or right child. Let us take the case where it is the left child. So, I would say there is some non terminal N f, it is starting from P up to you have to take some W e P e that derives N j and some N g and what is the parameters here this is p q this would be q plus 1 to e. So, this is deriving q plus 1 to e.

Now, I am doing it top down. So, I have already completed this that is the probabilities of this sequence this and W e plus 1 up to m, this I have already obtained. Now what I have to obtain in alpha j p q is the probability of W 1 to W p minus p N j p q and W q plus 1 to W m, this is what I have to obtain. Now here I have already obtained this probability and N f p e and somewhere W q plus 1 to W m this I have obtained. So, how do I write it in this in terms of N f p e? I will say this is N f p e although they may be different possible N f p e, we will do the summation over that later times probability of this will probability N f gives you N j N g already we have come here and this is deriving W q plus 1 to W e because I need these probabilities also.

Now, how do I, how do I write down this expression this we just did in the previous case. So, this is nothing but the inside probability. So, this would be beta g q plus 1 to e and that gives me the probability for this particular extraction although there are various summations that we will have to do, same case we can do for the when this is the right child then N g will come here and this is a huge very very similar.

(Refer Slide Time: 29:24)

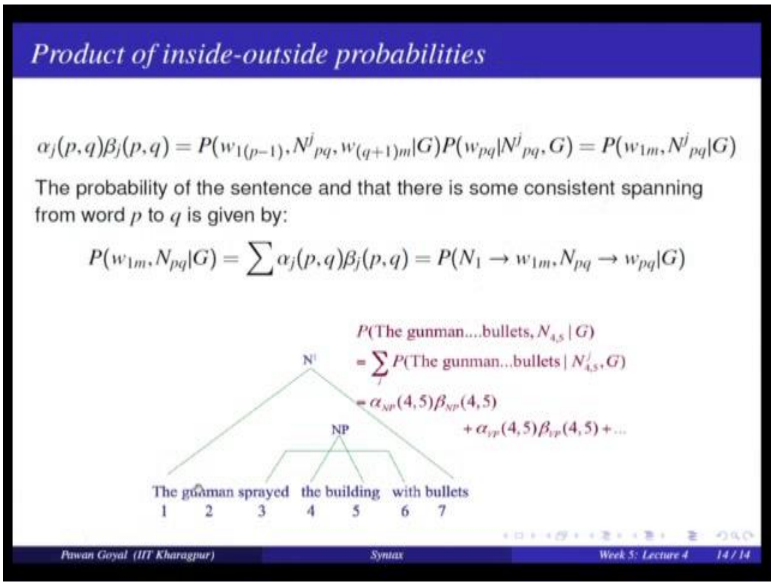



If you look at them in this case so this is where it is a left child, this is where it is a right child, in this case as we did it is alpha f p e probability of this rule and inside probability of this whole sequence; and what are these summing? Over all the possible f and g's because it can be any non terminal that is arriving this N along with any other non terminal; and what is something else that can vary? My e can vary from q plus 1 up to m. So, I am summing over all the possibilities and you can similarly see for the second case.

Now, so what is important here? You can compute this outside probabilities by using inside probabilities on a; that means, you have to first compute the inside probabilities use that to compute the outside probabilities although computing outside probabilities much much more difficult than computing inside probability if you want to do that on paper that is why we did a example only for the inside probabilities.

(Refer Slide Time: 30:31)

.




Now, coming to this point so, remember when I were saying that we have done this parameterization of inside and outside probabilities such that if we multiply that we had something interesting that we can use further and the goal here is also to compute the rule probabilities that is my third question. I was asking that how do I compute my rule probabilities of PCFG? We will come to that in the next lecture.

Let us quickly see what do I get if I multiply the alpha j p q with beta j p q? So, if I do that so, alpha j p g is nothing but probability of W 1 to W p minus 1 N j p q and W q plus 1 to m given my grammar and beta j p q is probability of W p q given and N j p q in the grammar. So, you multiply both of these, you will see that W p q can come here and NP q is also retained here because here I need N j p q to be given that is already given to me before. So, this is the multiplication, now what does this say? So, it is saying if I multiply alpha j p q with beta j p q, I get the probability of generating this whole sequence of words W 1 to W m and that a non terminal N j derives W p to W q given my grammar.

Now, suppose I want to find out the probability that any possible non terminal derives from W p to W q, there is some consistent spanning from W p W q, in that case, I will have to sum over all possible js that I can achieve by summing over all possible js for alpha j p q into beta j p q. This is the probability of the sentence and that there is some consistent is spanning from the W word p to word q and this is you can write in this manner this is the same thing N 1 derives W 1 m and NP q derives W p q this and this are the same thing except this particular index of N j.

Now, if we go back to the previous example that we took what does that mean. So, probability is that this whole sentence; the gunman sprayed the building with bullets and that there is some consistent spanning from word 4 to word 5; that means, there is some non terminal, they derive this sequence the word 4 to word 5, how do I obtain it? This should be summation over all the possible non terminals N j such that this happens the gunman sprayed the building with bullets N j p q given g and this is nothing, but alpha NP 4 5 beta NP 4 5. So, here I am taking all possible j's, j can be NP V p and so on.

Whatever non terminals so I multiply alpha beta for a particular j and add overall the possible j that gives me this probabilities, so, even here it may not be very very clear what is the actual use of obtaining this particular probability value, how do I use it further and that is what we will see in the next lecture that how do I use this inside outside probability for learning the rule probabilities of my probabilistic context free grammatical. So, I hope this concept of inside outside probabilities is clear to you and then we can see how do use that further in the next class.

Thank you.