Probabilistic Context Free Grammars

1 Defining PCFGs

A PCFG G consists of

1. A set of terminals: $\{\mathrm{w}^k\},\ k = 1 \ldots, \mathrm{V}$
2. A set of non terminals: $\{\mathrm{N}^i\},\ i = 1 \ldots, n$
3. A designated Start symbol: N^1
4. A set of rules: $\{\mathrm{N}^i \rightarrow \xi^j\},\ i = 1 \ldots, n$
5. A probability function P which assigns probabilities to rules so that, for all nonterminals N^i

$$\sum_j \mathrm{P}(\mathrm{N}^i \rightarrow \xi^j) = 1$$

Conventions:

Notation	Meaning
G	Grammar (PCFG)
$\mathcal{L}$	Language (generated or accepted by a grammar)
t	parse tree
$\{\mathrm{N}^1, \ldots, \mathrm{N}^n\}$	Nonterminal vocabulary (N^1 start symbol)
$\{\mathrm{w}^1, \ldots, \mathrm{w}^V\}$	Terminal vocabulary
$w_1 \ldots w_m$	Sentence to be parsed
N^j_{pq}	Nonterminal j spans positions p through q in sentence
$\alpha_j(p, q)$	Outside probabilities
$\beta_j(p, q)$	Inside probabilities
$\mathrm{N}^j \overset{*}{\Rightarrow} w_p \ldots w_q$	Nonterminal j dominates words w_p through w_q in sentence (not necessarily directly)

A key property of PCFGs is what we will call the **independence assumption**:

Independence Assumption
The probability of a node sequence N^s N^r depends only on the immediate mother node, not any node above that or outside the current constituent.

We first address the problem of finding the probability of a string given a PCFG grammar. This is directly useful for some applications, such as filtering the hypoethses of a speech recognizer.

But considering some efficient algorithms for calculating string probabilities will have a side-effect. It will provide a line of attack on a somewhat more central problem for this course: finding the most probable parse of a string.

2 Example

Thus the only parameters of a PCFG grammar are the rule probabilities, as in the following simple PCFG:

(1)

S → NP VP	1.0	NP → NP PP	0.4
PP → P NP	1.0	NP → astronomers	0.1
VP → V NP	0.7	NP → ears	0.18
VP → VP PP	0.3	NP → saw	0.04
P → with	1.0	NP → stars	0.18
V → saw	1.0	NP → telescopes	0.1

Notice the constraint

$$\sum_j \mathrm{P}(\mathrm{NP} \rightarrow \xi^j) = 1$$

is met. The sum of the rule probabilities for all the NP rules is 1.

We consider the probability of the sentence

(2) Astronomers saw stars with ears.

with respect to this grammar, which admits the two analyses shown in Figures 1 and 2. The probability of each anaylsis is computed simply by multiplying the probabilities of the rules.

We compute the probability of the string by summing the probabilities of its two analyses:

$$\begin{aligned} \mathrm{P}(w_{15}) &= \mathrm{P}(t_1) + \mathrm{P}(t_2) \\ &= 0.0009072 + 0.0006804 \\ &= 0.0015876 \end{aligned}$$

3 Two algorithms for efficiently computing string probabilities

The naive approach to computing string probabilities: Compute the probability of each parse and add.

But: The number of parses of a string with a CFG grammar in general grows exponentially with the length of the string. Therefore this is inefficient.

3.1 Inside Probabilities

$$\begin{aligned} \mathrm{P}(w_{1,m} \mid G) &= \mathrm{P}(\mathrm{N}^1 \overset{*}{\Rightarrow} w_{1,m} \mid G) \\ &= \mathrm{P}(w_{1,m} \mid \mathrm{N}^1_{1,m}, G) = \beta_1(1,m) \end{aligned}$$

The probability of a wordstring given the grammar equals the probability that the start symbol exhaustively dominates the string, which is the probability of the string $w_{1,m}$ given that the start symbol dominates words 1 through m, which we call the *inside probability* of the string for the span $1m$ and the category N^1. The general definition of an inside probability:

$$\beta_j(p,q) = \mathrm{P}(w_{p,q} \mid \mathrm{N}^j_{p,q}, G)$$

There is an an algotrithm for computing the inside probability of a strinng known as the **inside algorithm**. It breaks down to a base case and an induction:

1. **Base case**: The event that a nonterminal N^j exhaustively dominates the word at position k is N^j_{kk} and the probability that that happens given the grammar is:

$$\begin{aligned} \mathrm{P}(\mathrm{N}^j_{kk}) &= \beta_j(k,k) \\ &= \mathrm{P}(\mathrm{N}^j \rightarrow w_k \mid G) \end{aligned}$$

   The point of the β notation is that it uses only indices, indices of word positions (subscripts) and grammar elements (superscripts). The point of the 2nd line is that this reduces to a fact about the grammar, irrespective of word positions.

2. **Induction** We assume a Chomsky normal grammar, so the situation is as pictured in Figure 3.

   Then

$$\begin{aligned} \beta_j(p,q) &= \mathrm{P}(w_{pq} \mid \mathrm{N}^j_{pq}, G) && (1) \\ &= \sum_{r,s}\sum_{d=p}^{q-1} \mathrm{P}(w_{pd}, \mathrm{N}^r_{pd}, w(d+1)q, \mathrm{N}^s_{(d+1)q} \mid \mathrm{N}^j_{pq}, G) && (2) \end{aligned}$$

   We regard the probability of an analysis of w_{pq} as an instance of category N^j as the sum of the probabilities of an analysis using some way of expanding N^j; each such analysis uses some production:

$$\mathrm{N}^j \rightarrow \mathrm{N}^r\ \mathrm{N}^s$$

   Following the picture in Figure 3, we split the word string into two halves at word d, $p \leq d < q$, w_{pd} and $w_{(d+1)q}$, one corresponding to N^r and one to N^s. Thus the jointly occurring events in (2) are wordstring w_{pd} being analyzed as N^r and word string $w_{(d+1)q}$ being analyzed as N^s.

$$\begin{aligned} \beta_j(p,q) = \sum_{r,s}\sum_{d=p}^{q-1} & \mathrm{P}(\ \mathrm{N}^r_{pd},\ \mathrm{N}^s_{(d+1)q} \mid \mathrm{N}^j_{pq}, G) \times \mathrm{P}(w_{pd} \mid \mathrm{N}^j_{pq}, \mathrm{N}^r_{pd}, \mathrm{N}^s_{(d+1)q}, G) \\ & \times \mathrm{P}(w_{(d+1)q} \mid \mathrm{N}^j_{pq}, \mathrm{N}^r_{pd}, \mathrm{N}^s_{(d+1)q}, w_{pd}, G) \end{aligned} \quad (3)$$

   Equation (3) uses the chain rule on the joint events of (2). Now we use the independence assumptions of CFGs to conclude:

$$\begin{aligned} \beta_j(p,q) = \sum_{r,s}\sum_{d=p}^{q-1} & \mathrm{P}(\ \mathrm{N}^r_{pd},\ \mathrm{N}^s_{(d+1)q} \mid \mathrm{N}^j_{pq}, G) \times \mathrm{P}(w_{pd} \mid \mathrm{N}^r_{pd}, G) \\ & \times \mathrm{P}(w_{(d+1)q} \mid \mathrm{N}^s_{(d+1)q}, G) \end{aligned} \quad (4)$$

   And substituting in the definitions of an inside probability and a rule probability, we have the basic equation we can compute with:

$$\beta_j(p,q) = \sum_{r,s}\sum_{d=p}^{q-1} \mathrm{P}(\ \mathrm{N}^j \rightarrow \mathrm{N}^r\ \mathrm{N}^s \mid \mathrm{N}^j_{pq}, G) \times \beta_r(p,d) \times \beta_s(d+1,q) \quad (5)$$

This means we can compute inside probabilities bottum up.

We use a dynamic programming algorithm, and as is customary represent the computations in a table. Cell $(i,\ j)$ for row i and column j contains the result of the β computations for the string beginning with word i and ending with word j. First we do all the word βs:

	1	2	3	4	5
1	$\boldsymbol{\beta_{\text{NP}} = 0.1}$				
2		$\boldsymbol{\beta_{\text{NP}} = 0.04}$ $\boldsymbol{\beta_{\text{V}} = 1.0}$			
3			$\boldsymbol{\beta}_{\text{NP}} = 0.18$		
4				$\boldsymbol{\beta}_{\text{P}} = 0.1$	
5					$\boldsymbol{\beta}_{\text{NP}} = 0.18$
	astronomers	saw	stars	with	ears

We now proceed as with CYK, trying to build constituents of length 2 next:

	1	2	3	4	5
1	$\beta_{\text{NP}} = 0.1$				
2		$\beta_{\text{NP}} = 0.04$ $\beta_{\text{V}} = 1.0$	$\boldsymbol{\beta}_{\text{VP}} = 0.126$		
3			$\beta_{\text{NP}} = 0.18$		
4				$\beta_{\text{P}} = 0.1$	$\boldsymbol{\beta}_{\text{PP}} = 0.18$
5					$\beta_{\text{NP}} = 0.18$
	astronomers	saw	stars	with	ears

The computation of $\beta_{\text{VP}}(2,3)$ and $\beta_{\text{PP}}(4,5)$

$$\begin{aligned}
\beta_{\text{VP}}(2,3) &= \text{P}(\text{VP} \rightarrow \text{V NP}) \times \beta_{\text{V}}(2,2) \times \beta_{\text{NP}}(3,3) && (6)\\
&= 0.7 \times 1.0 \times 0.18 && (7)\\
&= 0.126 && (8)\\
\beta_{\text{PP}}(4,5) &= \text{P}(\text{PP} \rightarrow \text{P NP}) \times \beta_{\text{P}}(4,4) \times \beta_{\text{NP}}(5,5) && (9)\\
&= 1.0 \times 1.0 \times 0.18 && (10)\\
&= 0.18 && (11)
\end{aligned}$$

And in the next round:

	1	2	3	4	5
1	$\beta_{\text{NP}} = 0.1$		$\boldsymbol{\beta_{\text{S}} = 0.0126}$		
2		$\beta_{\text{NP}} = 0.04$ $\beta_{\text{V}} = 1.0$	$\beta_{\text{VP}} = 0.126$		
3			$\beta_{\text{NP}} = 0.18$		$\boldsymbol{\beta_{\text{NP}} = 0.01296}$
4				$\beta_{\text{P}} = 0.1$	$\beta_{\text{PP}} = 0.18$
5					$\beta_{\text{NP}} = 0.18$
	astronomers	saw	stars	with	ears

The computation of $\beta_{\text{S}}(1,3)$ and $\beta_{\text{NP}}(3,5)$

$$\begin{aligned}
\beta_{\text{S}}(1,3) &= \text{P}(\text{S} \rightarrow \text{NP VP}) \times \beta_{\text{NP}}(1,1) \times \beta_{\text{VP}}(2,3) && (12)\\
&= 1.0 \times 0.1 \times 0.126 && (13)\\
&= 0.0126 && (14)\\
\beta_{\text{NP}}(3,5) &= \text{P}(\text{NP} \rightarrow \text{NP PP}) \times \beta_{\text{NP}}(3,3) \times \beta_{\text{PP}}(4,5) && (15)\\
&= 0.4 \times 0.18 \times 0.18 && (16)\\
&= 0.1296 && (17)
\end{aligned}$$

And in the only round with an ambiguous constituent:

	1	2	3	4	5
1	$\beta_{\text{NP}} = 0.1$		$\beta_{\text{S}} = 0.0126$		
2		$\beta_{\text{NP}} = 0.04$ $\beta_{\text{V}} = 1.0$	$\beta_{\text{VP}} = 0.126$		$\boldsymbol{\beta_{\text{VP}} = 0.015876}$
3			$\beta_{\text{NP}} = 0.18$		$\beta_{\text{NP}} = 0.01296$
4				$\beta_{\text{P}} = 0.1$	$\beta_{\text{PP}} = 0.18$
5					$\beta_{\text{NP}} = 0.18$
	astronomers	saw	stars	with	ears

The computation of $\beta_{\text{VP}}(2,5)$:

$$\begin{aligned}
\beta_{\text{VP}}(2,5) &= (\text{P}(\text{VP} \rightarrow \text{V NP}) \times \beta_{\text{V}}(2,2) \times \beta_{\text{NP}}(3,5)) + && (18)\\
&\qquad (\text{P}(\text{VP} \rightarrow \text{VP PP}) \times \beta_{\text{VP}}(2,3) \times \beta_{\text{PP}}(4,5)) &&\\
&= (0.7 \times 1.0 \times 0.01296) + (0.3 * 0.126 * 0.18) && (19)\\
&= 0.009072 + 0.006804 && (20)\\
&= 015876 && (21)\\
& && (22)
\end{aligned}$$

The **probability table** (or chart) may be completed in the next step:

	1	2	3	4	5
1	$\beta_{\text{NP}} = 0.1$		$\beta_{\text{S}} = 0.0126$		$\boldsymbol{\beta_{\text{S}} = 0.0015876}$
2		$\beta_{\text{NP}} = 0.04$ $\beta_{\text{V}} = 1.0$	$\beta_{\text{VP}} = 0.126$		$\beta_{\text{VP}} = 0.015876$
3			$\beta_{\text{NP}} = 0.18$		$\beta_{\text{NP}} = 0.01296$
4				$\beta_{\text{P}} = 0.1$	$\beta_{\text{PP}} = 0.18$
5					$\beta_{\text{NP}} = 0.18$
	astronomers	saw	stars	with	ears

The computation of $\beta_{\text{S}}(1,5)$ is left as an exercise.

3.2 Outside Probabilities

We define $\alpha_j(p,q)$, the outside probability for a category N^j and a span w_{pq} as

$$\alpha_j(p,q) = \mathrm{P}(w_{1(p-1)}, \mathrm{N}^j_{pq}, w_{(q+1)m} \mid G)$$

This is the probability of generating N^j covering the span (p,q) (whatever the words in it) along with the all the words outside it.

Using the notion outside probability, we can view the problem of computing the probability of a string from the point of view of any word in it. Pick a word w_k. The following is true:

$$\begin{array}{rcl}
\mathrm{P}(w_{1,m} \mid G) & = & \sum_j \mathrm{P}(w_{1,(k-1)}, w_k, w_{(k+1)m}, \mathrm{N}^j_{kk} \mid G) \\
& = & \sum_j \ \mathrm{P}(w_{1,(k-1)}, w_{(k+1)m}, \mathrm{N}^j_{kk} \mid G) \\
& & \qquad \times \mathrm{P}(w_k \mid w_{1,(k-1)}, \mathrm{N}^j_{kk}, w_{(k+1)m}, G) \\
& = & \alpha_j(k,k) \times \mathrm{P}(\mathrm{N}^j \rightarrow w_k)
\end{array}$$

We start with the observation that any assignment of a category to w_k fixes a probability for analyses of the entire string compatible with that assignment. So we sum up such analysis probabilities for each category N^j in the grammar. We then apply the chain rule, and the independence assumptions of PCFGs, and the definition of $\alpha_j(p,q)$ to derive an expression for the probability of the entire string.

So if we knew the $\alpha_j(k,k)$ values, computing the string probability would be easy.

We now show the **outside algorithm**, a method for solving the more general problem of computing $\alpha_j(p,q)$. This is done topdown. We break it down as before into a base case and an induction:

1. **Base case**:

$$\begin{array}{rcl}
\alpha_1(1,m) & = & 1 \\
& = & 0 \text{ for } j \neq 1
\end{array}$$

2. **Induction** The situation is as pictured in Figure 4 or Figure 5

   Then

$$\begin{array}{rcl}
\alpha_j(p,q) & = & [\sum_{f,g\neq j} \sum_{e=q+1}^{m} \mathrm{P}(w_{1(p-1)}, w(q+1)m, \mathrm{N}^f_{pe}, \mathrm{N}^j_{pq}, \mathrm{N}^g_{(q+1)e})] \\
& & +[\sum_{f,g} \sum_{e=1}^{p-1} \mathrm{P}(w_{1(p-1)}, w(q+1)m, \mathrm{N}^f_{eq}, \mathrm{N}^g_{e(p-1)}, \mathrm{N}^j_{pq})]
\end{array} \qquad (23)$$

   The first double sum is the case of Figure 4. The second is the case of Figure 5. In the first case, we restrict the two categories not to be equal, so as not to count rules of the form:

$$\mathrm{X} \rightarrow \mathrm{Y\ Y}$$

   twice.

$$\begin{array}{rcl}
\alpha_j(p,q) & = & [\sum_{f,g\neq j} \sum_{e=q+1}^{m} \mathrm{P}(w_{1(p-1)}, w(q+1)m, \mathrm{N}^f_{pe}) \times \mathrm{P}(\mathrm{N}^j_{pq}, \mathrm{N}^g_{(q+1)e} \mid \mathrm{N}^f_{pe}) \\
& & \times \mathrm{P}(w(q+1)m \mid \mathrm{N}^g_{(q+1)e})] + [\sum_{f,g} \sum_{e=1}^{p-1} \mathrm{P}(w_{1(e-1)}, w(q+1)m, \mathrm{N}^f_{eq}) \\
& & \times \mathrm{P}(\mathrm{N}^g_{e(p-1)}, \mathrm{N}^j_{pq} \mid \mathrm{N}^f_{eq}) \times \mathrm{P}(w_{e(p-1)} \mid \mathrm{N}^g_{e(p-1)})] \\
& = & [\sum_{f,g\neq j} \sum_{e=q+1}^{m} \alpha_f(p,e) \mathrm{P}(\mathrm{N}^f \rightarrow \mathrm{N}^j\ \mathrm{N}^g) \beta_g(q+1,e)] \\
& & +[\sum_{f,g} \sum_{e=1}^{p-1} \alpha_f(e,q) \mathrm{P}(\mathrm{N}^j \rightarrow \mathrm{N}^g\ \mathrm{N}^j) \beta_g(e,p-1)]
\end{array}$$

4 Finding the most probable parse

We define $\delta_j(p,q)$ as the highest probability parse of w_{pq} as N^j

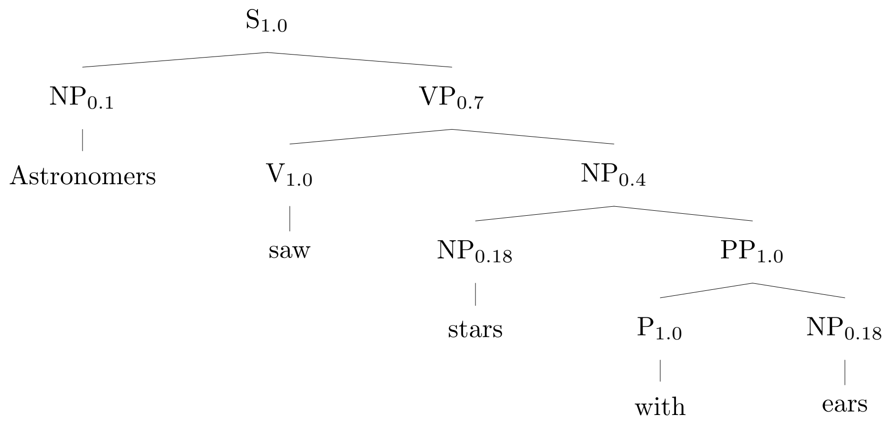
What follows is a Viterbi like algorithm for finding the best parse. Corresponding to Viterbi valkues we will have δ values. Corresponding to a Viterbi bnacktrace we will have poointers to the subtrees that help us produce our highest probvability parse for each (category,start, end) triple.

1. **Initialization**:
$$\delta_i(p,p) = \mathrm{P}(\mathrm{N}^i \rightarrow w_p)$$
2. **Induction**:
$$\delta_i(p,q) = \max_{\substack{1 \leq j,k \leq n \\ p \leq r < q}} \mathrm{P}(\mathrm{N}^i \rightarrow \mathrm{N}^j\ \mathrm{N}^k)\delta_j(p,r)\delta_k(r+1,q)$$

   Store backtrace:
$$\Psi_i(p,q) = \operatorname*{argmax}_{j,k,r} \mathrm{P}(\mathrm{N}^i \rightarrow \mathrm{N}^j\ \mathrm{N}^k)\delta_j(p,r)\delta_k(r+1,q)$$
3. **Termination**: The probability of the max probability tree $\hat{t}$ is:
$$\mathrm{P}(\hat{t}) = \delta_1(1,m)$$

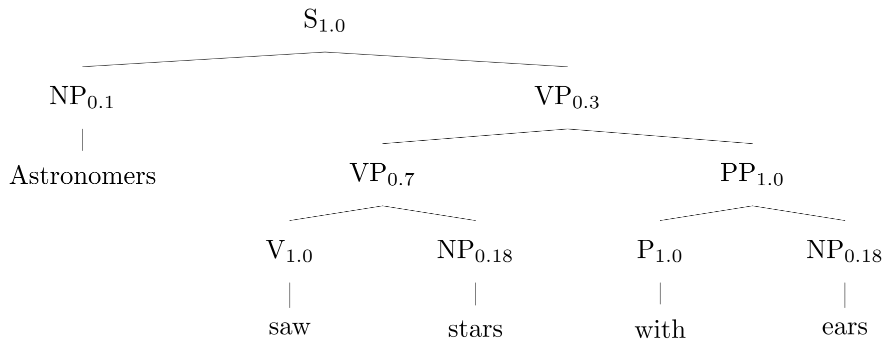
   In addition we know the mother node and span of the highest probability parse: N^1 and (p,q).

   In general given a mother node and a span (i,p,q), we can construct a highest probability parse tree below it by following the Ψ pointers:
$$\begin{array}{ll} \text{Let} & \Psi_i(p,q) = (j,k,r) \\ \text{Then} & \text{left daughter} = \mathrm{N}^j_{pr} \\ & \text{right daughter} = \mathrm{N}^k_{(r+1)q} \end{array}$$



$$\begin{aligned} \mathrm{P}(t_1) &= 1.0 \times 0.1 \times \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 \\ &= 0.0009072 \end{aligned}$$

Figure 1: t_1



$$\begin{aligned} \mathrm{P}(t_2) &= 1.0 \times 0.1 \times \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 \\ &= 0.0006804 \end{aligned}$$

Figure 2: t_2

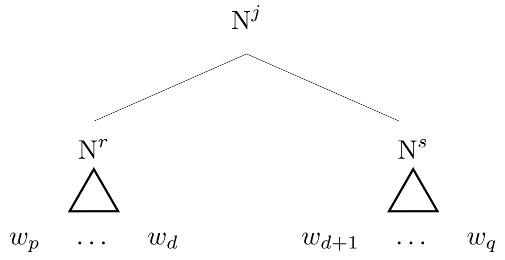


Figure 3: Computation of inside probability: $\beta_j(p, q)$

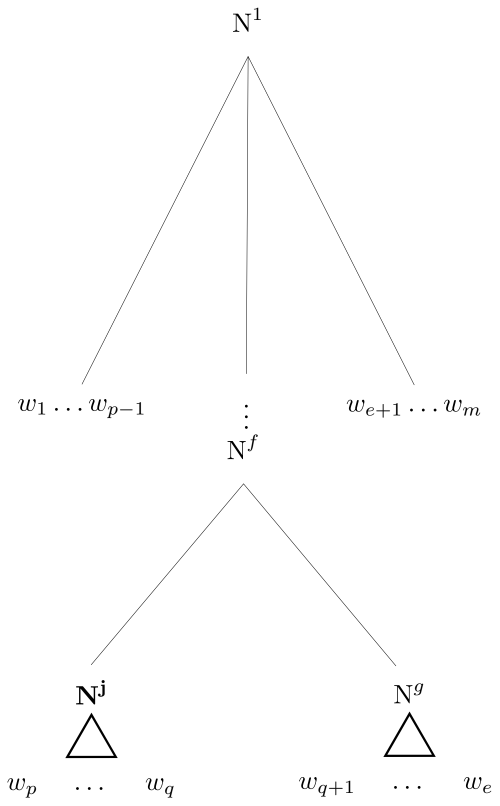


Figure 4: Computation of outside probability (left category): $\alpha_j(p, q)$

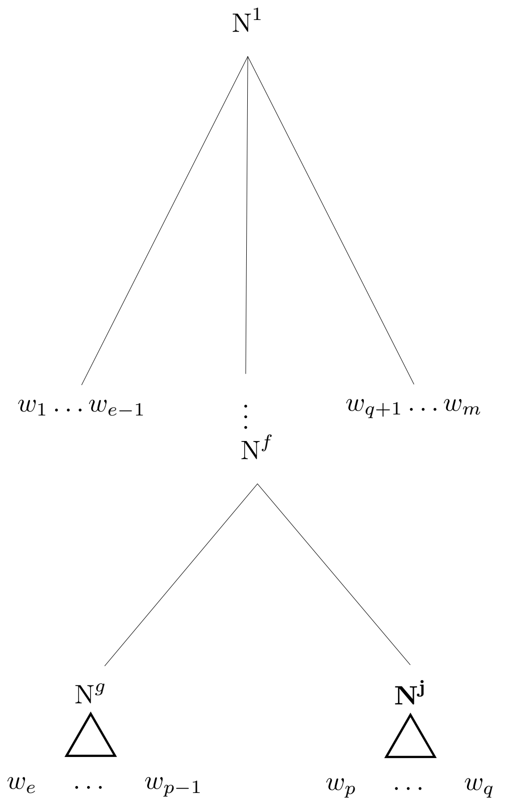


Figure 5: Computation of outside probability (right category): $\alpha_j(p, q)$