

Context-Free languages (part II)

Prof. Dan A. Simovici

UMB

- 1 Leftmost Derivations and Ambiguity
- 2 Chomsky Normal Form
- 3 Derivation Trees

Leftmost Derivations

Definition

Let $G = (A_N, A_T, S, P)$ be a context-free grammar.

A *leftmost derivation* is a derivation $\gamma_0 \Rightarrow \cdots \Rightarrow \gamma_n$ in G such that, if the production applied in deriving γ_{k+1} from γ_k is $X_k \rightarrow \beta_k$, then

$\gamma_k = \gamma'_k X_k \gamma''_k$, $\gamma_{k+1} = \gamma'_k \beta_k \gamma''_k$ and $\gamma'_k \in A_T^*$.

- The words γ_k (for $0 \leq k \leq n$) are referred to as *left sentential forms*.
- If $\gamma_k = \gamma'_k X_k \gamma''_k$, where $\gamma'_k \in A_T^*$, then γ'_k is the *closed part* of γ_k , while $X_k \gamma''_k$ is the *open part* of γ_k .
- In a context-free grammar G ,

$$\gamma_0 \Rightarrow \gamma_1 \Rightarrow \dots \Rightarrow \gamma_n$$

is a leftmost derivation if, at every step of this derivation, we always rewrite the leftmost nonterminal symbol.

Notations

- The existence of a leftmost derivation of length n in the context-free grammar G , $\gamma_0 \Rightarrow \gamma_1 \Rightarrow \dots \Rightarrow \gamma_n$, will be denoted by $\gamma_0 \xRightarrow[n]{G, \text{left}} \gamma_n$.
- The existence of a leftmost derivation of any length of γ' from γ in the same grammar will be denoted by $\gamma \xRightarrow[G, \text{left}]{*} \gamma'$.
- The existence of a leftmost derivation of positive length of γ' from γ will be denoted by $\gamma \xRightarrow[G, \text{left}]{+} \gamma'$.

Example

Let $G = (A_N, A_T, S_0, P)$ be a context-free grammar, where $A_N = \{S_0, S_1, S_2\}$, $A_T = \{a, b\}$, and P contains the following productions:

$$\begin{aligned} S_0 &\rightarrow aS_2, S_0 \rightarrow bS_1, S_1 \rightarrow a, S_1 \rightarrow aS_0, \\ S_1 &\rightarrow bS_1S_1, S_2 \rightarrow b, S_2 \rightarrow bS_0, S_2 \rightarrow aS_2S_2. \end{aligned}$$

(Example cont'd)

The derivation

$$\begin{array}{lclcl}
 S_0 & \Rightarrow & bS_1 & \Rightarrow & bbS_1S_1 & \Rightarrow & bbS_1aS_0 \\
 & & \Rightarrow & & bbS_1aaS_2 & \Rightarrow & bbaaaS_2 & \Rightarrow & bbaaab
 \end{array}$$

is not leftmost since in deriving bbS_1aaS_2 from bbS_1aS_0 we do not replace the leftmost nonterminal S_1 .

(Example cont'd)

We can transform this derivation into a leftmost derivation by changing the order in which nonterminals are replaced. Namely, in grammar G , we have the leftmost derivation

$$\begin{aligned} S_0 &\Rightarrow bS_1 \Rightarrow bbS_1S_1 \Rightarrow bbaS_1 \\ &\Rightarrow bbaaS_0 \Rightarrow bbaaaS_2 \Rightarrow bbaaab. \end{aligned}$$

Theorem

Let $G = (A_N, A_T, S, P)$ be a context-free grammar. For every complete derivation d of length n in G , $X \Rightarrow \gamma_1 \Rightarrow \cdots \Rightarrow \gamma_n$, where $\gamma_n = u \in A_T^$, there is a complete leftmost derivation of length n , using the same productions as d , that allows us to derive γ_n from X .*

Proof

The argument is by strong induction on $n \geq 1$ for leftmost derivations. For $n = 1$, the statement is trivially true, since any derivation $X \Rightarrow w_1$ is a leftmost derivation.

Suppose that the statement holds for derivations whose length is no more than n , and let d

$$X \Rightarrow \gamma_1 \Rightarrow \cdots \Rightarrow \gamma_{n+1}$$

be a derivation of length $n + 1$. If the first production used in this derivation is $X \rightarrow w_0 X_{i_1} w_1 \cdots X_{i_k} w_k$, where $w_i \in A_T^*$ for $0 \leq i \leq k$, then we can write $\gamma_{n+1} = w_0 u_1 w_1 \cdots u_k w_k$, where d_j is a complete derivation $X_{i_j} \xRightarrow[G]{*} u_j$ of length no greater than n , for $1 \leq j \leq k$.

(Proof cont'd)

By the inductive hypothesis, for each of these derivations d_j , we obtain the existence of the leftmost derivation $d'_j: X_{i_j} \xRightarrow{*}_{G, \text{left}} u_j$ for $1 \leq j \leq k$, which uses the same set of productions as d_j . Now, we obtain the existence of the leftmost derivation d' :

$$\begin{aligned}
 X &\Rightarrow w_0 X_{i_1} w_1 X_{i_2} \dots X_{i_k} w_k \\
 &\xRightarrow{*} w_0 u_1 w_1 X_{i_2} \dots X_{i_k} w_k \text{ (using derivation } d'_1) \\
 &\xRightarrow{*} w_0 u_1 w_1 u_2 \dots X_{i_k} w_k \text{ (using derivation } d'_2) \\
 &\vdots \\
 &\xRightarrow{*} w_0 u_1 w_1 u_2 \dots u_k w_k \text{ (using derivation } d'_k),
 \end{aligned}$$

which concludes our argument.

The Theorem may fail if the derivation is not complete, that is, the final word is not in A_T^* .

Example

Let

$$G = (\{S, X, Y, U, V\}, \{a, b\}, S, \{S \rightarrow XY, Y \rightarrow UV, \\ X \rightarrow a, U \rightarrow b, V \rightarrow b\})$$

be a context-free grammar. Consider the derivation

$$S \Rightarrow XY \Rightarrow XUV$$

This derivation is not leftmost, and there is no leftmost derivation in G such that $S \xRightarrow[G]{*} XUV$.

Corollary

Let $G = (A_N, A_T, S, P)$ be a context-free grammar. For every complete derivation d of length n in G , $\gamma_0 \Rightarrow \gamma_1 \Rightarrow \cdots \Rightarrow \gamma_n$, where $\gamma_0 \in (A_N \cup A_T)^+$ and $\gamma_n \in A_T^$, there is a complete leftmost derivation of length n , using the same productions as d , that allows us to derive γ_n from γ_0 .*

Proof

Suppose that $\gamma_0 = s_0 \dots s_{k-1}$, where $s_i \in A_N \cup A_T$ for $0 \leq i \leq k-1$. We can write $\gamma_n = u_0 \dots u_{k-1}$ such that $s_i \xRightarrow[G]{*} u_i \in A_T^*$ for $0 \leq i \leq k-1$.

Thus, we obtain the existence of the leftmost derivations $s_i \xRightarrow[G, \text{left}]{*} u_i$ for $0 \leq i \leq k-1$ that use the same productions as the corresponding previous derivations. Starting from these derivations we obtain the leftmost derivation:

$$\begin{array}{ll}
 \gamma_0 = s_0 s_1 \dots s_{k-1} & \\
 \xRightarrow[G, \text{left}]{*} & u_0 s_1 \dots s_{k-1} \\
 \xRightarrow[G, \text{left}]{*} & u_0 u_1 \dots s_{k-1} \\
 \vdots & \\
 \xRightarrow[G, \text{left}]{*} & u_0 u_1 \dots u_{k-1} = \gamma_n.
 \end{array}$$

Definition

A context-free grammar $G = (A_N, A_T, S, P)$ is *ambiguous* if there exists a word $w \in A_T^*$ such that there are at least two leftmost derivations from S to w in G . Otherwise, G is *unambiguous*.

A context-free language can be generated by both ambiguous and unambiguous grammars.

Example

Consider the context-free grammars

$$G_1 = (\{S\}, \{a\}, S, \{S \rightarrow SS, S \rightarrow a\})$$

and

$$G_2 = (\{S\}, \{a\}, S, \{S \rightarrow aS, S \rightarrow a\}).$$

They both generate the language $\{a^n \mid n \geq 1\}$.

(Example cont'd)

They both generate the language $\{a^n \mid n \geq 1\}$. Note that in G_1 we have distinct leftmost derivations:

$$\begin{array}{ccccccc}
 S & \Rightarrow_{G_1} & SS & \Rightarrow_{G_1} & SSS & \Rightarrow_{G_1} & aSS \\
 & & \Rightarrow_{G_1} & & aaS & \Rightarrow_{G_1} & aaa
 \end{array}$$

and

$$\begin{array}{ccccccc}
 S & \Rightarrow_{G_1} & SS & \Rightarrow_{G_1} & aS & \Rightarrow_{G_1} & aSS \\
 & & \Rightarrow_{G_1} & & aaS & \Rightarrow_{G_1} & aaa.
 \end{array}$$

Thus, G_1 is an ambiguous grammar.

(Example cont'd)

On other hand, the equivalent grammar G_2 is unambiguous, since for every a^n , $n \geq 1$, we have exactly one derivation:

$$S \xRightarrow{G_2} aS \xRightarrow{G_2} a^2S \cdots \xRightarrow{G_2} a^n.$$

Since a language may have both an ambiguous and an unambiguous grammar, it may not be sufficient to examine one grammar to determine whether or not a language is ambiguous.

Definition

Let L be a context-free language. L is *unambiguous* if there is an unambiguous context-free grammar G such that $L = L(G)$.

L is *inherently ambiguous* if every context-free grammar G such that $L(G) = L$ is ambiguous.

The language $\{a^n \mid n \geq 1\}$ is unambiguous.

Definition

A context-free grammar $G = (A_N, A_T, S, P)$ is in *Chomsky normal form* if all productions are either of the form $X \rightarrow YZ$ or of the form $X \rightarrow a$, where $X, Y, Z \in A_N$ and $a \in A_T$.

If G is in Chomsky normal form, then G is λ -free, so $\lambda \notin L(G)$.

Theorem

For every context-free grammar G such that $\lambda \notin L(G)$ there is an equivalent grammar in Chomsky normal form.

Proof.

We can assume that G is a λ -free grammar, G has no chain productions and that every production that contains a terminal symbol is of the form $X \rightarrow a$.

Thus, the productions of G have either the form $X \rightarrow a$ or the form $X \rightarrow X_{i_0} \cdots X_{i_{k-1}}$ with $k \geq 2$. □

(Proof cont'd)

Productions of the form $X \rightarrow a$ or $X \rightarrow X_{i_0} X_{i_1}$ already conform to Chomsky normal form. If $\pi : X \rightarrow X_{i_0} \cdots X_{i_{k-1}}$ is a production of P with $k \geq 3$, consider $k - 2$ new nonterminals $Z_0^\pi, \dots, Z_{k-3}^\pi$ and the productions

$$X \rightarrow X_{i_0} Z_0^\pi, Z_0^\pi \rightarrow X_{i_1} Z_1^\pi, \dots, Z_{k-3}^\pi \rightarrow X_{i_{k-2}} X_{i_{k-1}}$$

Define the grammar $G' = (A_N \cup A', A_T, S, P')$, where A' consists of all symbols Z_ℓ^π , and P' consists of all productions of the form $X \rightarrow a$ or $X \rightarrow X_{i_0} X_{i_1}$, and of productions obtained from productions of P having the form $X \rightarrow X_{i_0} \cdots X_{i_{k-1}}$ with $k \geq 3$, by applying the method described above. It is easy to see that G' is equivalent to G and that G' is in Chomsky normal form.

Example

Let $G = (\{S_0, S_1, S_2\}, \{a, b\}, S_0, P)$ be the context-free grammar, where P contains the following productions:

$$\begin{aligned} S_0 &\rightarrow aS_2, S_0 \rightarrow bS_1, S_1 \rightarrow a, S_1 \rightarrow aS_0, S_1 \rightarrow bS_1S_1, \\ S_2 &\rightarrow b, S_2 \rightarrow bS_0, S_2 \rightarrow aS_2S_2. \end{aligned}$$

By introducing the new nonterminal symbols X_a, X_b we obtain the grammar $G_1 = (\{S_0, S_1, S_2, X_a, X_b\}, \{a, b\}, S_0, P_1)$, where P_1 consists of

$$\begin{aligned} S_0 &\rightarrow X_aS_2, S_0 \rightarrow X_bS_1, S_1 \rightarrow a, S_1 \rightarrow X_aS_0, S_1 \rightarrow X_bS_1S_1, \\ S_2 &\rightarrow b, S_2 \rightarrow X_bS_0, S_2 \rightarrow X_aS_2S_2, X_a \rightarrow a, X_b \rightarrow b. \end{aligned}$$

(Example cont'd)

G_1 is equivalent to G , has no chain productions and every production that contains a terminal symbol is of the form $X \rightarrow a$. This grammar has two productions, $S_1 \rightarrow X_b S_1 S_1$ and $S_2 \rightarrow X_a S_2 S_2$, that violate Chomsky normal form, so we introduce the new nonterminals Z_0, Z_1 .

Applying the technique introduced before to these productions results in the set of productions P' given by:

$$\begin{aligned} S_0 &\rightarrow X_a S_2, S_0 \rightarrow X_b S_1, S_1 \rightarrow a, S_1 \rightarrow X_a S_0, \\ S_1 &\rightarrow X_b Z_0, Z_0 \rightarrow S_1 S_1, S_2 \rightarrow b, S_2 \rightarrow X_b S_0, \\ S_2 &\rightarrow X_a Z_1, Z_1 \rightarrow S_2 S_2, X_a \rightarrow a, X_b \rightarrow b. \end{aligned}$$

The resulting grammar $G' = (\{S_0, S_1, S_2, X_a, X_b, Z_0, Z_1\}, \{a, b\}, S_0, P')$ is in Chomsky normal form and is equivalent to G .

Using Chomsky normal form we can prove an important decidability result for the class $\mathcal{L}_2$. To this end, we need the following technical result relating the length of a word to the length of its derivation.

Lemma

Let $G = (A_N, A_T, S, P)$ be a context-free grammar in Chomsky normal form. Then, if $S \xRightarrow[\alpha]{} x$ we have $|\alpha| \leq 2|x| - 1$.*

Proof

We prove a slightly stronger statement, namely that if $X \xRightarrow[\alpha]{*} x$ for some $X \in A_N$, then $|\alpha| \leq 2|x| - 1$.

The argument is by induction on $n = |x| \geq 1$. If $n = 1$, we have $x = a$ for $a \in A_T$ and the derivation $X \xRightarrow[\alpha]{*} x$ consists in the application of the production $\pi : X \rightarrow a$. Therefore, $|\alpha| = 1$ and the inequality is satisfied.

(Proof cont'd)

Suppose that the statement holds for words of length less than n , and let $x \in L(G)$ be a word such that $|x| = n$, where $n > 1$. Let the first production applied be $X \rightarrow YZ$; then we can write $x = uv$, where $Y \xRightarrow[\beta]{*} u$ and $Z \xRightarrow[\gamma]{*} v$ and $|\alpha| = |\beta| + |\gamma| + 1$, because the productions used in the last two derivations are exactly the ones used in $X \xRightarrow[\alpha]{*} x$. Applying the inductive hypothesis we obtain

$$|\alpha| = |\beta| + |\gamma| + 1 \leq 2|u| - 1 + 2|v| - 1 + 1 = 2(|u| + |v|) - 1 = 2|x| - 1.$$

Theorem

There is an algorithm to determine for a context-free grammar $G = (A_N, A_T, S, P)$ and a word $x \in A_T^$ whether or not $x \in L(G)$.*

Proof.

Construct a grammar G' equivalent to G such that one of the following two cases occurs:

- 1 if $\lambda \notin L(G)$ then G' is λ -free;
- 2 if $\lambda \in L(G)$ then G' contains a unique erasure production $S' \rightarrow \lambda$, where S' is the start symbol of G' and S' does not occur in any right member of any production of G' .



(Proof cont'd)

If $x = \lambda$, then $x \in L(G)$ if and only if $S \rightarrow \lambda$ is a production in G' . Suppose that $x \neq \lambda$. Let G_1 be a context-free grammar in Chomsky normal form such that $L(G_1) = L(G') - \{\lambda\} = L(G) - \{\lambda\}$. We have $x \in L(G_1)$ if and only if $x \in L(G)$. By the previous Lemma, if $S \xRightarrow[\alpha]{*} x$, then $|\alpha| \leq 2|x| - 1$, so we can decide if $x \in L(G)$ by listing all derivations of length at most $2|x| - 1$.

- As an alternative to writing a sequence of derivation steps, we consider describe context-free derivations using labeled ordered trees, so-called derivation trees.
- The labels of the leaves of an A -labeled ordered tree, when read from left-to-right, spell out a word in A^* .

Definition of Derivation Trees

Definition

Let $G = (A_N, A_T, S, P)$ be a λ -free context-free grammar, and let $d = (\gamma_0, \dots, \gamma_m)$ be a derivation in G , where $\gamma_0 = X \in A_N$ and $\gamma_i \in (A_N \cup A_T)^*$ for $0 \leq i \leq m$. Let $A = A_N \cup A_T$.

The *derivation tree of the derivation d* is an A -labeled, ordered tree T_d defined inductively as follows:

Def. cont'd

- ① If $m = 0$, then T_d consists of only one node labeled by $(0, X)$.
- ② Suppose that $m \geq 1$ and that $\gamma_1 = X_0 \dots X_{n-1}$, where $X_0 \dots X_{n-1} \in (A_N \cup A_T)^*$. Let T_i be the A -labeled ordered tree that corresponds to the derivation $(X_i, \dots, \alpha_i)$ for $0 \leq i \leq n-1$, where $\alpha = \alpha_0 \dots \alpha_{n-1}$. Then, T_d is $\langle T_0, \dots, T_{n-1}; X \rangle$.

The *set of derivation trees of G* is the set

$$\text{TREES}(G) = \{T_d \mid d \text{ is a derivation in } G\}.$$

A derivation tree $T_d \in \text{TREES}(G)$ is *complete* if $\text{word}(T_d) \in A_T^*$, i.e. if all its leaves are labeled by terminal symbols of the grammar.
The set of complete derivation trees of G is denoted by $\text{TREES}_c(G)$.

Example

Let

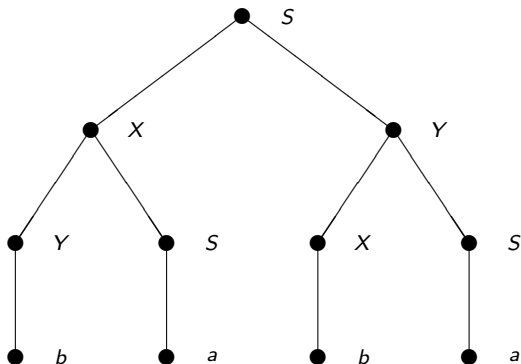
$$G = (\{S, X, Y\}, \{a, b\}, S, \{S \rightarrow XY, S \rightarrow a, X \rightarrow YS, \\ Y \rightarrow XS, X \rightarrow b, Y \rightarrow b\})$$

be a context-free grammar in Chomsky normal form. The derivation tree of

$$S \Rightarrow XY \Rightarrow YSY \Rightarrow YSXS \Rightarrow bSXS \Rightarrow baXS \Rightarrow babS \Rightarrow baba$$

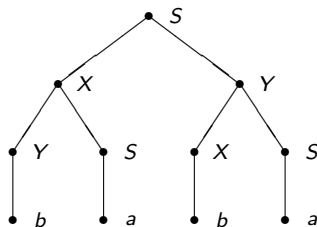
is given next:

$$S \Rightarrow XY \Rightarrow YSY \Rightarrow YSXS \Rightarrow bSXS \Rightarrow baXS \Rightarrow babS \Rightarrow baba$$



Every derivation in a context-free grammar $G = (A_N, A_T, S, P)$ is described by a derivation tree. Conversely, if T is a derivation tree such that $\text{word}(T) = x \in A_T^*$ then, in general, several distinct derivations exist for the word x .

Example



In the same grammar $G = (\{S, X, Y\}, \{a, b\}, S, \{S \rightarrow XY, S \rightarrow a, X \rightarrow YS, Y \rightarrow XS, X \rightarrow b, Y \rightarrow b\})$ this derivation tree also describes the derivation:

$$S \Rightarrow XY \Rightarrow XXS \Rightarrow XXa \Rightarrow YSXa \Rightarrow bSXa \Rightarrow baXa \Rightarrow baba$$

Theorem

Let $G = (A_N, A_T, S, P)$ be a context-free grammar, and let $T \in \text{TREES}_c(G)$ be a complete derivation tree whose root is labeled by X , where the word spelled by T , $\text{word}(T) = u \in A_T^$. There is a unique leftmost (rightmost) derivation $X \xRightarrow[G, \text{left}]{*} u$. Moreover, the lengths of the leftmost and the rightmost derivations equal the number of internal nodes of T .*

Proof

The argument for leftmost derivations is by induction on the height of T . If $\text{height}(T) = 1$, then the derivation that corresponds to T is (X, u) , which is an one-step leftmost derivation.

Suppose that the statement holds for complete derivation trees of height less than n , and let T be a complete derivation tree in G such that $\text{height}(T) = n$. Then, $T = \langle T_0, \dots, T_{k-1}; X \rangle$, where $\text{height}(T_i) < n$ for $0 \leq i \leq k-1$. Also, the root of T_i is labeled by the symbol $X_i \in A_N \cup A_T$ and its leaves are labeled by the terminal word u_i for $0 \leq i \leq k-1$, where $u_0 \cdots u_{k-1} = u$.

(Proof cont'd)

By the inductive hypothesis, for each of the trees T_i , there is a unique leftmost derivation d_i :

$$X_i \Rightarrow w_{i0} \Rightarrow \cdots \Rightarrow w_{i\ell_i-1} = u_i$$

and the length of d_i is equal to the number of internal nodes of T_i for $0 \leq i \leq k-1$.

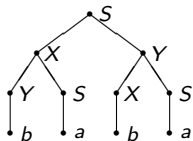
Then, we obtain the following leftmost derivation that corresponds to T :

$$\begin{aligned} X &\Rightarrow X_0 X_1 \cdots X_{k-1} \\ &\Rightarrow w_{00} X_1 \cdots X_{k-1} \Rightarrow \cdots \Rightarrow u_0 X_1 \cdots X_{k-1} \\ &\Rightarrow u_0 w_{10} \cdots X_{k-1} \Rightarrow \cdots \Rightarrow u_0 u_1 \cdots X_{k-1} \\ &\vdots \\ &\Rightarrow u_0 u_1 \cdots w_{k-1 0} \Rightarrow \cdots \Rightarrow u_0 u_1 \cdots u_{k-1}. \end{aligned}$$

(Proof cont'd)

If d is a leftmost derivation for T , then it must expand the nonterminals symbol $X_{i_0}, \dots, X_{i_p}$ that occur in $X_0 \cdots X_{k-1}$. Thus, the derivation d must use the productions that occur in the leftmost derivations $d_{i_0}, \dots, d_{i_{k-1}}$, respectively, in that order. This shows that the leftmost derivation is unique and the length of this derivation equals the number of internal nodes of T .

Example



For the derivation tree

$$\begin{aligned}
 S &\Rightarrow XY \Rightarrow YSY \Rightarrow bSY \Rightarrow \\
 &\quad baY \Rightarrow baXS \Rightarrow babS \Rightarrow baba
 \end{aligned}$$

is a leftmost derivation.

(Example cont'd)

The derivation

$$\begin{aligned} S &\Rightarrow XY \Rightarrow XXS \Rightarrow XXa \Rightarrow Xba \\ &\Rightarrow YSba \Rightarrow Yaba \Rightarrow baba \end{aligned}$$

is the rightmost derivations.

If G is a context-free grammar and $x \in L(G)$, several distinct derivation trees may exist for x . In some cases, a considerable number of such distinct trees may exist.

Example

Let $G = (\{S\}, \{a\}, S, \{S \rightarrow SS, S \rightarrow a\})$ be a context-free grammar. It is not difficult to see that the language generated by G is $L(G) = \{a^m \mid m \geq 1\}$. Denote by $C(n)$ the number of derivation trees that describe derivations of the form $S \xRightarrow[G]{*} a^{n+1}$. We have $C(0) = 1$, and

$$C(n) = \sum_{j=0}^{n-1} C(j)C(n-1-j),$$

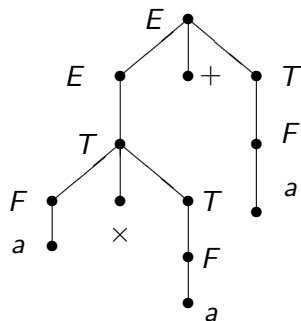
It is possible to prove that $C(n) = \Theta\left(\frac{4^n}{n^{1.5}}\right)$.

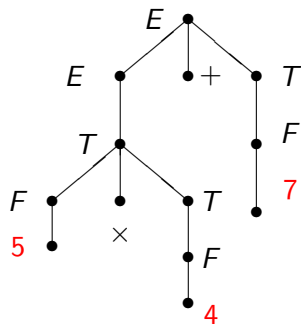
Derivation trees for arithmetic expressions reflect implicitly the priority order of arithmetic operations.

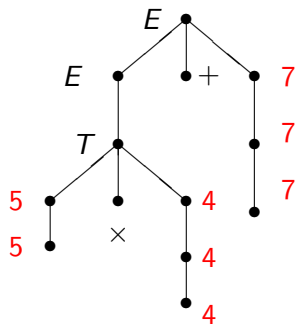
Consider the context-free grammar

$$G = (\{E, T, F\}, \{+, \times, (,)\}, E, \{E \rightarrow T, E \rightarrow E + T, T \rightarrow F, T \rightarrow F \times T, F \rightarrow a, F \rightarrow (E)\}).$$

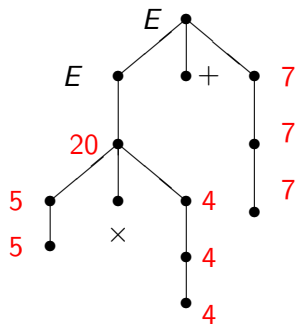
Derivation Tree for $a \times a + a$

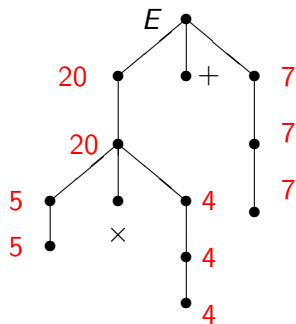


Computation of $5 \times 4 + 7$ 

Computation of $5 \times 4 + 7$ 

Computation of $5 \times 4 + 7$



Computation of $5 \times 4 + 7$ 

Computation of $5 \times 4 + 7$

