# Context-free Grammars

(known from DM565)

$$G = (V, \Sigma, R, S)$$

- $V$ = variables
- $\Sigma$ = alphabet
- $R$ = Rules
- $S$ = startsymbol

Derivation $\quad S \Rightarrow u_1 A_1 v_1 \Rightarrow u_2 A_2 v_2 \Rightarrow \dots \Rightarrow u_k A_k v_k \Rightarrow w \in \Sigma^*$

each step replaces a variable $A_i$ by some right hand side of a rule in $R$. We write $S \overset{*}{\Rightarrow} w$ if $S$ can derive $w$ in one or more steps.

$$L(G) = \{ w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w \}$$

why context-free?

replacing $A$ in $uAv$ by $A \to \gamma \in R$ does not depend on $u$ or $v$

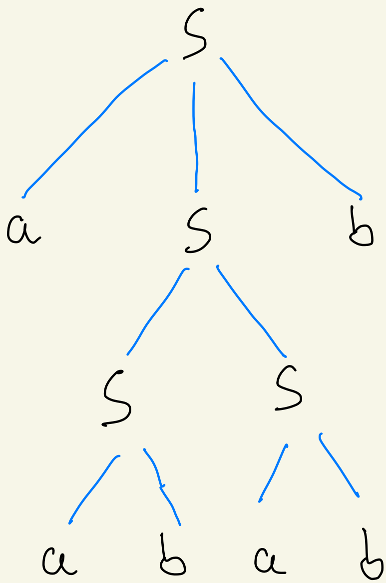
# Example 1

$G: S \to aSb \mid SS \mid ab$

So $R = \{S \to aSb, S \to SS, S \to ab\}$

Parse tree for

$$S \Rightarrow aSb \Rightarrow aSSb \Rightarrow aabSb \Rightarrow aababb$$



$G: S \to aSb \mid SS \mid ab$

What is $L(G)$?

Claim: $L(G)$ is the set of strings with the same # of a's and b's in which every prefix has at least as many a's as b's $= L'$

Clearly we have $L(G) \subseteq L'$ as every derivation preserves the property

Suppose $w$ has the property above

if $|w| = 2$ then $S \to ab = w$ derives $w$

Suppose that every string $w \in L'$ with $|w| \leq 2k$ can be derived and look at $w' \in L'$ with $|w'| = 2k+2$

Case 1 every proper prefix of $w'$ has more $a$'s than $b$'s

Then $w' = aw''b$ when $w'' \in L'$ and $|w''| = 2k$

By induction $S \overset{*}{\Rightarrow} w''$ and then

$S \Rightarrow aSb \overset{*}{\Rightarrow} aw''b$ so $S \overset{*}{\Rightarrow} w'$

Case 2 $w' = w_1 w_2$ when $w_i \in L'$ for $i = 1, 2$

By induction $S \overset{*}{\Rightarrow} w_1$ and $S \overset{*}{\Rightarrow} w_2$ so

$S \to SS \overset{*}{\Rightarrow} w_1 S \overset{*}{\Rightarrow} w_1 w_2 = w'$

so $S \overset{*}{\Rightarrow} w'$

We have shown that $L(G) = L'$

Note that $L'$ is not regular as

$L' \cap a^* b^* = \{a^n b^n \mid n \geq 0\}$

**Theorem** If $L$ is a regular language then $L = L(G)$ for some context-free grammar $G$

P: let $M=(Q,\Sigma,\delta,q_0,F)$ be a DFA with $L(M)=L$

$Q=\{q_0,q_1,\ldots,q_k\} \longrightarrow$ Variables in $G=(V,\Sigma,R,S)$

$V=\{X_0,X_1,\ldots,X_k\}$

Rules in $G$: If $q_i \xrightarrow{a} q_j$ then $X_i \to aX_j \in R$

if $q_\ell \in F$ then $X_\ell \to \varepsilon \in R$

$X_0$ is starting symbol for $G$

Suppose $w\in L$, $w=a_1a_2\cdots a_n$: $q_0 \xrightarrow{a_1} q_{i_1} \xrightarrow{a_2} q_{i_2} \to \cdots \xrightarrow{a_n} q_{i_n}$

Then $X_0 \Rightarrow a_1X_{i_1} \Rightarrow a_1a_2X_{i_2} \Rightarrow \cdots \Rightarrow a_1a_2\cdots a_nX_{i_n} \Rightarrow a_1a_2\cdots a_n = w$

Suppose $w'\in L(G)$ $X_0 \Rightarrow b_1X_{j_1} \Rightarrow b_1b_2X_{j_2} \Rightarrow \cdots \Rightarrow w'X_{j_p} \Rightarrow w'$ $p=|w'|$

So $q_0 \xrightarrow{b_1} q_{j_1} \xrightarrow{b_2} q_{j_2} \to \cdots \xrightarrow{b_p} q_{j_p}$ so $w'\in L(M)$

Fact (we do not prove it)

Every context-free language over an alphabet $\Sigma$ with $|\Sigma| = 1$ is also regular

So for alphabets $\Sigma$ with $|\Sigma| = 1$

$L$ is regular $\Leftrightarrow$ $L$ is context-free

This allows us to conclude e.g. that $L = \{a^p \mid p \text{ is a prime}\}$

is not a context-free language

## Question 1: checking membership of a regular language

Given a DFA (NFA) $M$ and $w \in \Sigma^*$; Does $w \in L(M)$?

This is easy to check in linear time for a DFA: just 'eat' $w$ one character a a time and check if we reached an accepting state after reading $w$.

When $M$ is an NFA we must first convert $M$ into an equivalent $M'$ DFA and then do as above to check if $w \in L(M') = L(M)$

This may take exponential time as $|Q(M')| \in O(2^{|Q(M)|})$

## Question 2: checking membership of a context-free language

Given a context-free grammar $G$ and $w \in \Sigma^*$ is $w \in L(G)$?

Not so easy:

Example $S \to SA \mid \varepsilon$, $A \to a \mid \varepsilon$

$S \to SA \to \cdots \to SA^k \to A^k \to A^{k-1} \to \cdots \to AA \to aA \to aa$

We have no bound on the length of a derivation so we cannot check all possible derivations

**Solution:** consider Chomsky CFG's:

All rules of the form $A \to BC$ or $A \to a$ plus possibly $S \to \varepsilon$

Let $G$ be a CFG on Chomsky form and let $w \in \Sigma^*$ with $|w| = n$

Then every derivation of a string of length $n$ has exactly $2n-1$ steps:

$$S \to AB \xrightarrow[\text{steps}]{n-2} X_1 X_2 \cdots X_n$$

$$\xrightarrow[\text{steps}]{n} a_1 a_2 \cdots a_n = w$$

Algorithm for checking whether $w \in L(G)$ when $G$ is a CFG in Chomsky form:

- let $k = |w|$ and try all possible derivations of length $2k-1$
- Can be done more efficiently, but not important here

# Theorem 2.9

Every context-free language is generated by some CFG in Chomsky normal form

P: G in Chomsky form => G is a CFG

other direction: suppose G is a CFG. We will convert G to a Chomsky grammar G' in 4 steps

1. Add a new start variable $S_0$
2. Eliminate $\varepsilon$-rules ($A \to \varepsilon$) except for new starting variable
3. Eliminate $A \to B$ rules
4. Convert long rules $A \to A_1 A_2 \dots A_k$, $k \geq 3$ to several shorter rules and convert $A \to cd$, $A \to cD$, $A \to Cd$ to proper format

ad 1. add $S_0$ and $S_0 \to S$

then $S_0$ does not appear on any righthand side of a rule

clearly $S \overset{*}{\underset{G}{\Rightarrow}} w \iff S_0 \overset{*}{\Rightarrow} w$

Fix an ordering of the variables in $V$

ad 2: removing $\varepsilon$-rules

process the transitions according to the ordering of variables given above

assume $A \to \varepsilon$ and $A$ is next in the order with an $\varepsilon$-transition

- For each rule $X \to \gamma$ in $R$ contains at least one occurrence of $A$ in $\gamma$
  replace each subset of occurrences of $A$ in $\gamma$ by $\varepsilon$

e.g. $X \to uAvAw$

$$\left.\begin{array}{l} X \to uvw \\ X \to uAvw \\ X \to uvAw \\ X \to uAvAw \end{array}\right\}$$

add them

- If $X \to A$ is a rule, then add $X \to \varepsilon$, unless the transition $X \to \varepsilon$ was already removed ($X$ before $A$ in order) ad 3 removing unit rules

Process variables according to the fixed order of the right hand side:

let $A \to B$ be a rule of $R$

add rules $A \to u$ for all $u$ s.t. $B \to u$ is in $R$, unless $A \to u$ is a unit rule $(A \to C)$ that we already processed

repeat until no more unit rules

Easy to see that we still have

$$S_0 \overset{*}{\Rightarrow} w \iff S \overset{*}{\underset{G}{\Rightarrow}} w$$

## ad 4. Eliminating long rules

a) let $A \to u_1 u_2 \cdots u_k$, $k \geq 3$ $u_i \in V \cup \Sigma$

be a rule of $R$

create $k-2$ new variables $A_1, A_2 \cdots A_{k-2}$

(private to this replacement)

replace $A \to u_1 u_2 \cdots u_k$ by

$$A \to u_1 A_1$$
$$A_1 \to u_2 A_2$$
$$\vdots$$
$$A_{k-2} \to u_{k-1} u_k$$

b) If $A \to u_1 u_2$ is in $R$ with

at least one of $u_1, u_2$ in $\Sigma$

replace such a $u_i$ by a new variable $U_i$ and

$U_i \to u_i$ e.g. $A \to bX$ is replaced by $A \to U_b X$

$U_b \to b$

# Example $S \to aSa \mid bSb \mid A$ and $A \to a \mid b \mid \varepsilon$

1. add $S_0$ $S_0 \to S$, $S \to aSa \mid bSb \mid A$, $A \to a \mid b \mid \varepsilon$

Fix order of variables $A < S < S_0$

2a remove $A \to \varepsilon$: $S_0 \to S$, $S \to aSa \mid bSb \mid A \mid \varepsilon$, $A \to a \mid b$

2b remove $S \to \varepsilon$: $S_0 \to S \mid \varepsilon$, $S \to aa \mid aSa \mid bb \mid bSb \mid A$, $A \to a \mid b$

3a remove $S \to A$: $S_0 \to S \mid \varepsilon$, $S \to aa \mid aSa \mid bb \mid bSb \mid a \mid b$, $A \to a, b$

3b remove $S_0 \to S$: $S_0 \to aa \mid aSa \mid bb \mid bSb \mid a \mid b \mid \varepsilon$

$S \to aa \mid aSa \mid bb \mid bSb \mid a \mid b$,

$A \to a \mid b$

4 eliminate long rules

add new variables $C$, $D$ and rules $C \to Sa$, $D \to Sb$

use them to break rules of length 3

Fix rules of length 2 with at least one non-variable

$S_0 \to aa \mid aC \mid bb \mid bD \mid a \mid b \mid \varepsilon$

$S \to aa \mid aC \mid bb \mid bD \mid a \mid b$

$A \to a, b$, $C \to Sa$, $D \to Sb$

use $A$, $C$ and $D$ and new variable $B$ with $B \to b$

$S_0 \to AA \mid AC \mid BB \mid BD \mid a \mid b \mid \varepsilon$

$S \to AA \mid AC \mid BB \mid BD \mid a \mid b$

$A \to a$

$B \to b$

$C \to SA$

$D \to SB$