

# Chomsky Normal Form for Context-Free Grammars

Deepak D'Souza

Department of Computer Science and Automation  
Indian Institute of Science, Bangalore.

9 October 2018

# Outline

- 1 CNF
- 2 Converting to CNF
- 3 Correctness

# Chomsky Normal Form

A Context-Free Grammar  $G$  is in **Chomsky Normal Form** if all productions are of the form

$$\begin{aligned} X &\rightarrow YZ \text{ or} \\ X &\rightarrow a \end{aligned}$$

Its a “normal form” in the sense that

## CNF

Every CFG  $G$  can be converted to a CFG  $G'$  in Chomsky Normal Form, with  $L(G') = L(G) - \{\epsilon\}$ .

# Example

CFG  $G_4$

$$S \rightarrow (S) \mid SS \mid \epsilon.$$

“Equivalent” grammar in CNF:

CFG  $G'_4$  in CNF

$$\begin{aligned} S &\rightarrow LX \mid SS \mid LR \\ X &\rightarrow SR \\ L &\rightarrow ( \\ R &\rightarrow ) \end{aligned}$$

## Why is CNF useful?

- Gives us a way to solve the parsing problem for a CFG: Given CFG  $G$  and  $w \in A^*$ , does  $w \in L(G)$ ?

## Why is CNF useful?

- Gives us a way to solve the parsing problem for a CFG: Given CFG  $G$  and  $w \in A^*$ , does  $w \in L(G)$ ?
  - If  $G$  is in CNF, then the length of derivation of  $w$  (if one exists) can be bounded by  $2|w|$ .

## Why is CNF useful?

- Gives us a way to solve the parsing problem for a CFG: Given CFG  $G$  and  $w \in A^*$ , does  $w \in L(G)$ ?
  - If  $G$  is in CNF, then the length of derivation of  $w$  (if one exists) can be bounded by  $2|w|$ .
- Makes proofs of properties of CFG's simpler.

# Procedure to convert a CFG to CNF

- Main problem is “unit” productions of the form  $A \rightarrow B$  and  $\epsilon$ -productions of the form  $B \rightarrow \epsilon$ .
- Once these productions are eliminated, converting to CNF is easy.



## Procedure to remove unit and $\epsilon$ -productions

Given a CFG  $G = (N, A, S, P)$ .

- Repeatedly add productions according to the steps below till no more new productions can be added.
  - 1 If  $A \rightarrow \alpha B \beta$  and  $B \rightarrow \epsilon$  then add the production  $A \rightarrow \alpha \beta$ .
  - 2 If  $A \rightarrow B$  and  $B \rightarrow \gamma$  then add the production  $A \rightarrow \gamma$ .
- Let resulting grammar be  $G' = (N, A, S, P')$ .
- Let  $G''$  be grammar  $(N, A, S, P'')$ , where  $P''$  is obtained from  $P'$  by dropping unit- and  $\epsilon$ -productions.
- Return  $G''$ .

# Example

Apply procedure to the grammar below:

CFG  $G_4$

$$S \rightarrow (S) \mid SS \mid \epsilon.$$

# Correctness claims

- Algorithm terminates.

# Correctness claims

- Algorithm terminates.
  - Notice that each new production added has a RHS that is a subsequence of RHS of an original production in  $P$ .

# Correctness claims

- Algorithm terminates.
  - Notice that each new production added has a RHS that is a subsequence of RHS of an original production in  $P$ .
- $G'$  generates the same language as  $G$ .
  - Let  $G'_i$  be grammar obtained after  $i$ -th step, with  $G'_0 = G$ .
  - Then clearly  $L(G'_{i+1}) = L(G'_i)$ .

# Correctness of $G''$

## Claim

$$L(G'') = L(G') - \{\epsilon\}.$$

## Subclaim

Let  $w \in L(G')$  with  $w \neq \epsilon$ . Then any **minimal-length** derivation of  $w$  in  $G'$  does not use unit or  $\epsilon$ -productions.

# Proof of Subclaim

## Subclaim

Let  $w \in L(G')$  with  $w \neq \epsilon$ . Then any **minimal-length** derivation of  $w$  in  $G'$  does not use unit or  $\epsilon$ -productions.

Consider a derivation of  $w$  in  $G'$  which uses a production  $B \rightarrow \epsilon$ . It must be of the form

$$S \xRightarrow{1} \alpha X \beta \xRightarrow{1} \alpha \gamma B \delta \beta \xRightarrow{m} \alpha' \gamma' B \delta' \beta' \xRightarrow{1} \alpha' \gamma' \delta' \beta' \xRightarrow{n} w.$$

# Proof of Subclaim

## Subclaim

Let  $w \in L(G')$  with  $w \neq \epsilon$ . Then any **minimal-length** derivation of  $w$  in  $G'$  does not use unit or  $\epsilon$ -productions.

Consider a derivation of  $w$  in  $G'$  which uses a production  $B \rightarrow \epsilon$ . It must be of the form

$$\begin{array}{l} S \xRightarrow{1} \alpha X \beta \xRightarrow{1} \alpha \gamma B \delta \beta \xRightarrow{m} \alpha' \gamma' B \delta' \beta' \xRightarrow{1} \alpha' \gamma' \delta' \beta' \xRightarrow{n} w. \\ S \xRightarrow{1} \alpha X \beta \xRightarrow{1} \alpha \gamma \delta \beta \xRightarrow{m} \alpha' \gamma' \delta' \beta' \xRightarrow{n} w. \end{array}$$

Now consider a derivation of  $w$  in  $G'$  which uses a production  $A \rightarrow B$ . It must be of the form

$$S \xRightarrow{1} \alpha A \beta \xRightarrow{m} \alpha' A \beta' \xRightarrow{1} \alpha' B \beta' \xRightarrow{n} \alpha'' B \beta'' \xRightarrow{1} \alpha'' \gamma \beta'' \xRightarrow{p} w.$$



# Proof of Subclaim

## Subclaim

Let  $w \in L(G')$  with  $w \neq \epsilon$ . Then any **minimal-length** derivation of  $w$  in  $G'$  does not use unit or  $\epsilon$ -productions.

Consider a derivation of  $w$  in  $G'$  which uses a production  $B \rightarrow \epsilon$ . It must be of the form

$$\begin{array}{l} S \xRightarrow{1} \alpha X \beta \xRightarrow{1} \alpha \gamma B \delta \beta \xRightarrow{m} \alpha' \gamma' B \delta' \beta' \xRightarrow{1} \alpha' \gamma' \delta' \beta' \xRightarrow{n} w. \\ S \xRightarrow{1} \alpha X \beta \xRightarrow{1} \alpha \gamma \delta \beta \xRightarrow{m} \alpha' \gamma' \delta' \beta' \xRightarrow{n} w. \end{array}$$

Now consider a derivation of  $w$  in  $G'$  which uses a production  $A \rightarrow B$ . It must be of the form

$$\begin{array}{l} S \xRightarrow{1} \alpha A \beta \xRightarrow{m} \alpha' A \beta' \xRightarrow{1} \alpha' B \beta' \xRightarrow{n} \alpha'' B \beta'' \xRightarrow{1} \alpha'' \gamma \beta'' \xRightarrow{p} w. \\ S \xRightarrow{1} \alpha A \beta \xRightarrow{m} \alpha' A \beta' \xRightarrow{1} \alpha' \gamma \beta' \xRightarrow{n} \alpha'' \gamma \beta'' \xRightarrow{p} w. \end{array}$$

Where did we use  $w \neq \epsilon$ ?

## Where did we use $w \neq \epsilon$ ?

The only time we *cannot* guarantee that the non-terminal  $B$  would have been introduced in the derivation, is when the production  $(B \rightarrow \epsilon)$  is  $S \rightarrow \epsilon$ .