Regular language and context-free grammar

Lin Chen

Email: Lin.Chen@ttu.edu

Grader: zulfi.khan@ttu.edu

TEXAS TECH
UNIVERSITY

Context-free grammar

- Is CFG more general than REG?
- CFG is not a subset of REG, as $\{0^n 1^n : n \geq 0\}$ is in CFG
- $L_{REG} \subseteq L_{CFG}$

$L_{REG} \subseteq L_{CFG}$

- Given any regular expression r, we can create a CFG $G = (V, \Sigma, R, S)$ such that $L[G] = L[r]$
  - We prove it inductively

Regular expression

- The regular expressions of Σ^* are all strings over $\Sigma \cup \{(,), \emptyset, +, \star\}$ that can be obtained through the following operations:
  - $\emptyset$ and every member of Σ is a regular expression
  - If α and β are regular expressions, then so is $(\alpha\beta)$
  - if α and β are regular expressions, then so is $(\alpha + \beta)$
  - if α is a regular expression, then so is α^*
  - Nothing else is a regular expression

$L_{REG} \subseteq L_{CFG}$

- Base case

  - $r = a, a \in \Sigma$

- CFG : $S \rightarrow a$

$L_{REG} \subseteq L_{CFG}$

- Base case

  - $r = e$

- CFG : $S \rightarrow e$

$L_{REG} \subseteq L_{CFG}$

- Base case

  - $r = \emptyset$

- CGF: $S \rightarrow SS$ (no derivation to terminals)

$L_{REG} \subseteq L_{CFG}$

- Recursive case

  \- $r = (r_1 r_2)$

- CFG :

  \- suppose we have G_1, G_2 such that $L(G_i) = L(r_i)$ for $i = 1,2$

$L_{REG} \subseteq L_{CFG}$

- Recursive case
  - $r = (r_1 r_2)$

- CFG :
  - suppose we have G_1, G_2 such that $L(G_i) = L(r_i)$ for $i = 1{,}2$
  - Let S_1, S_2 be the start symbols of G_1, G_2
  - G = all rules from G_1, G_2, plus new start symbol S, and new rule:
    $S \rightarrow S_1 S_2$

$L_{REG} \subseteq L_{CFG}$

- Recursive case

  - $r = (r_1 + r_2)$

- CFG:

  - suppose we have G_1, G_2 such that $L(G_i) = L(r_i)$ for $i = 1,2$

$L_{REG} \subseteq L_{CFG}$

- Recursive case

  - $r = (r_1 + r_2)$

- CFG :

  - suppose we have G_1, G_2 such that $L(G_i) = L(r_i)$ for $i = 1,2$

  - Let S_1, S_2 be the start symbols of G_1, G_2

  - G = all rules from G_1, G_2, plus new start symbol S, and new rule:

    $S \rightarrow S_1 | S_2$

$L_{REG} \subseteq L_{CFG}$

- Recursive case
  - $r = (r_1^*)$

- CFG:
  - suppose we have G_1 such that $L(G_1) = L(r_1)$

$L_{REG} \subseteq L_{CFG}$

- Recursive case

 - $r = (r_1^*)$

- CFG :

 - suppose we have G_1 such that $L(G_1) = L(r_1)$

  - Let S_1 be the start symbols of G_1

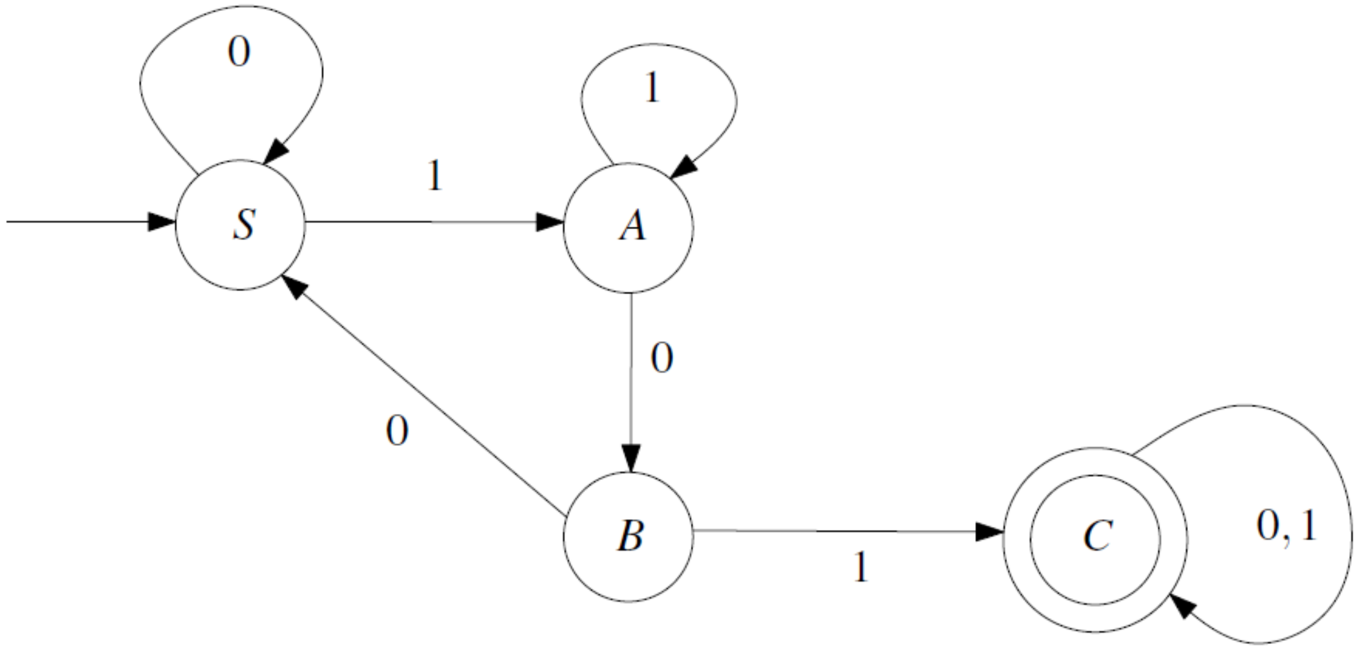
  - G = all rules from G_1, plus new start symbol S, and new rule:

   $S \rightarrow S_1 S | e$

$L_{REG} \subseteq L_{CFG}$

- Recall $L_{REG} = L_{DFA}$, we can also transform a DFA directly to CFG

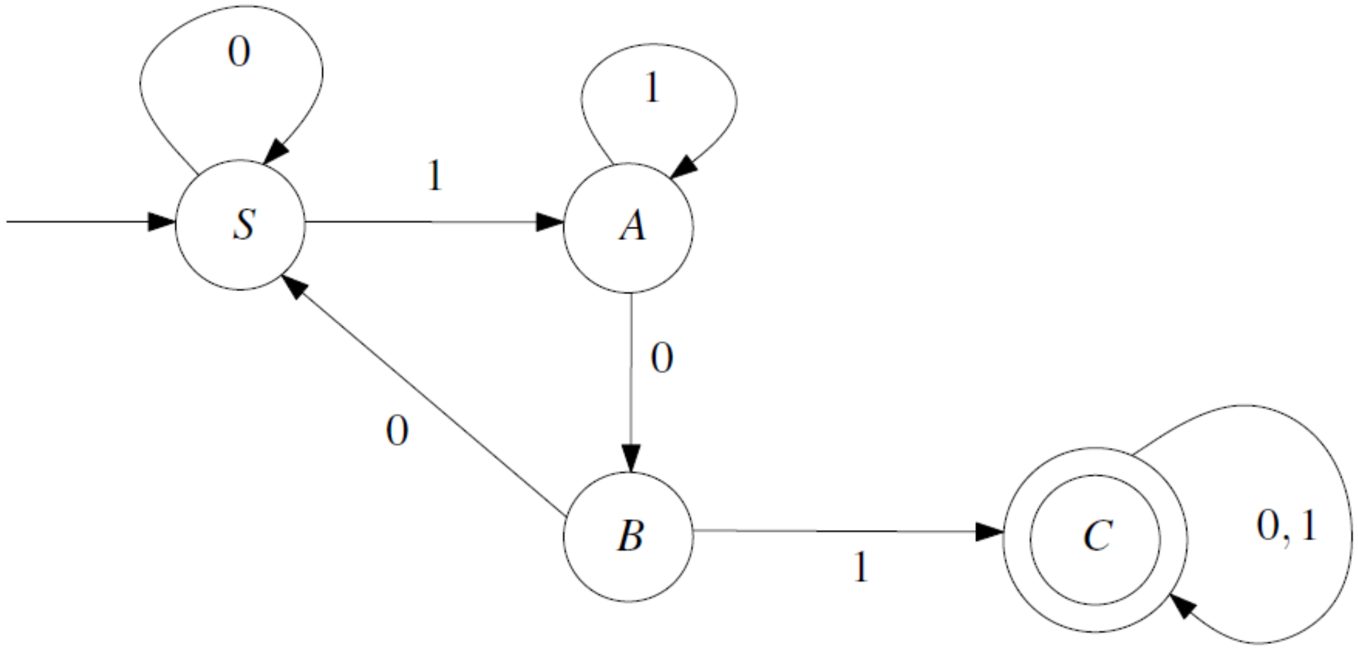
$$L = \{w \in \{0,1\}^* : \ 101 \text{ is a substring of } w\}.$$



$L_{REG} \subseteq L_{CFG}$

- Recall $L_{REG} = L_{DFA}$, we can also transform a DFA directly to CFG

$$L = \{w \in \{0,1\}^* : \ 101 \text{ is a substring of } w\}.$$



$$\begin{aligned} S &\rightarrow 0S|1A \\ A &\rightarrow 0B|1A \\ B &\rightarrow 0S|1C \\ C &\rightarrow 0C|1C|\epsilon \end{aligned}$$

$L_{REG} \subseteq L_{CFG}$

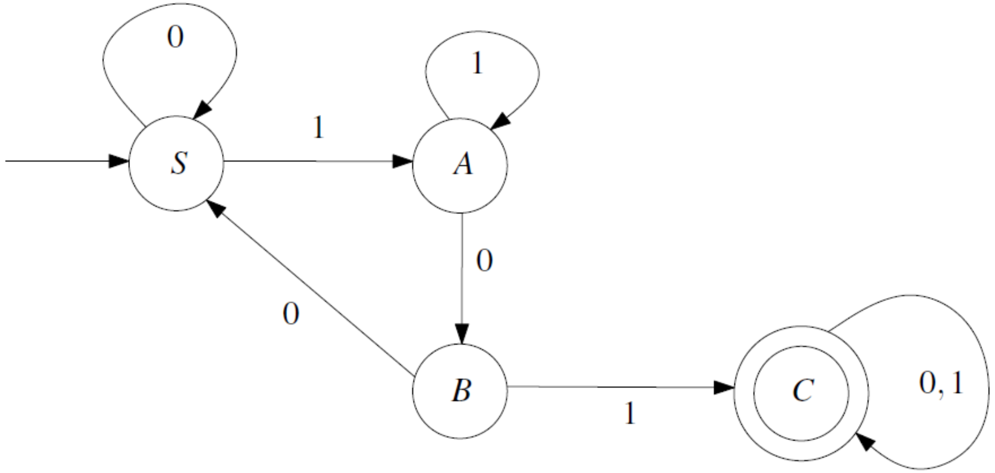
- Recall $L_{REG} = L_{DFA}$, we can also transform a DFA directly to CFG

$$L = \{w \in \{0,1\}^* : \ 101 \text{ is a substring of } w\}.$$

  - Example: watch the string 010011011

$$S, S, A, B, S, A, A, B, C, C.$$

$$\begin{aligned} S &\rightarrow 0S|1A \\ A &\rightarrow 0B|1A \\ B &\rightarrow 0S|1C \\ C &\rightarrow 0C|1C|\epsilon \end{aligned}$$



$L_{REG} \subseteq L_{CFG}$

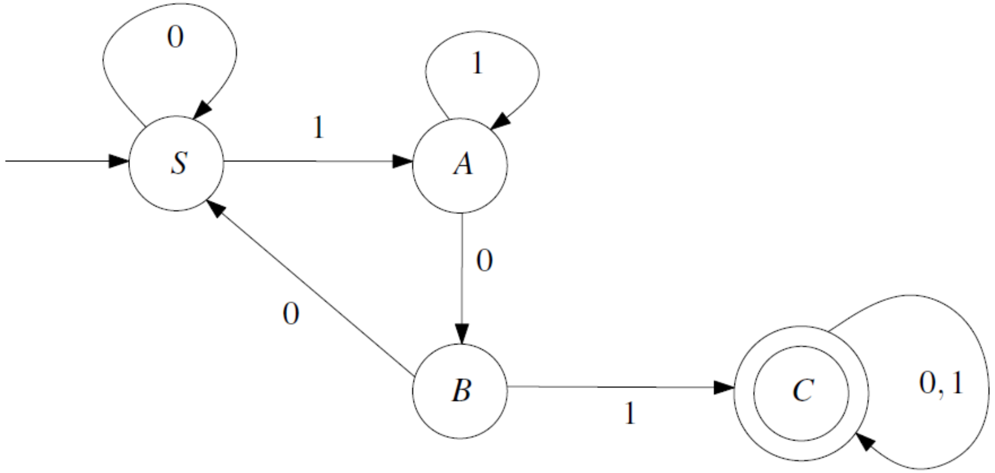
- Recall $L_{REG} = L_{DFA}$, we can also transform a DFA directly to CFG

$$L = \{w \in \{0,1\}^* : \text{ 101 is a substring of } w\}.$$

  - Example: watch the string **010011011**

$$S, S, A, B, S, A, A, B, C, C.$$

$$\begin{aligned} S &\rightarrow 0S|1A \\ A &\rightarrow 0B|1A \\ B &\rightarrow 0S|1C \\ C &\rightarrow 0C|1C|\epsilon \end{aligned}$$



$$\begin{aligned} S &\Rightarrow 0S \\ &\Rightarrow 01A \\ &\Rightarrow 010B \\ &\Rightarrow 0100S \\ &\Rightarrow 01001A \\ &\Rightarrow 010011A \\ &\Rightarrow 0100110B \\ &\Rightarrow 01001101C \\ &\Rightarrow 010011011C \\ &\Rightarrow 010011011. \end{aligned}$$

$L_{REG} \subseteq L_{CFG}$

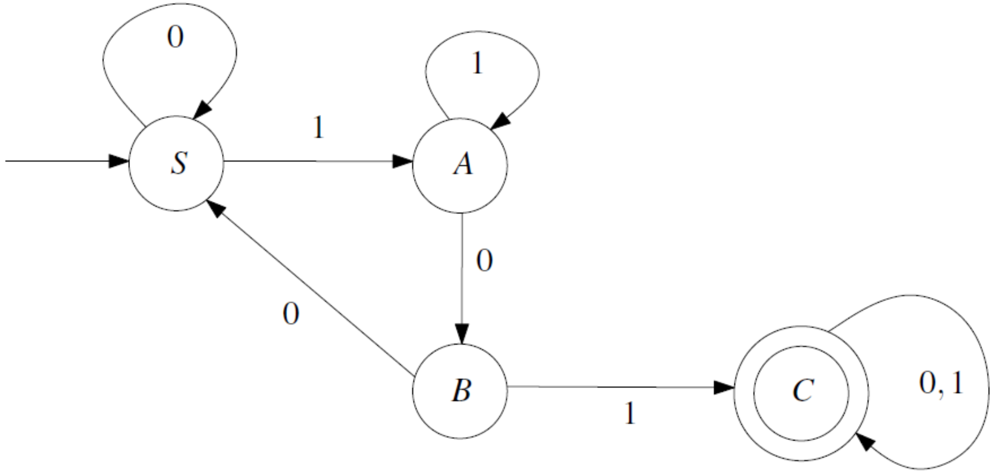
- Recall $L_{REG} = L_{DFA}$, we can also transform a DFA directly to CFG

$$L = \{w \in \{0,1\}^* : \ 101 \text{ is a substring of } w\}.$$

  - Example: watch the string 10011

$$S, A, B, S, A, A,$$

$$\begin{aligned} S &\rightarrow 0S|1A \\ A &\rightarrow 0B|1A \\ B &\rightarrow 0S|1C \\ C &\rightarrow 0C|1C|\epsilon \end{aligned}$$



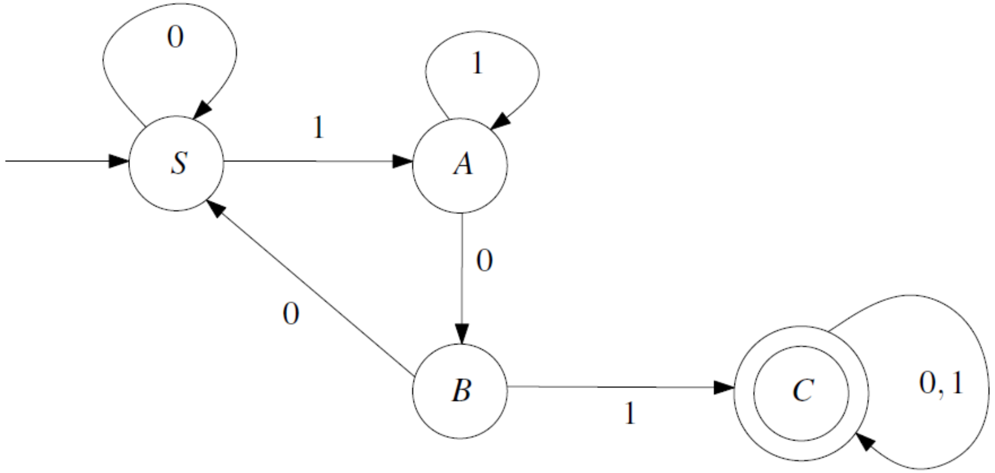
$L_{REG} \subseteq L_{CFG}$

- Recall $L_{REG} = L_{DFA}$, we can also transform a DFA directly to CFG

$$L = \{w \in \{0,1\}^* : \text{ 101 is a substring of } w\}.$$

  - Example: watch the string **10011**

$$S, A, B, S, A, A,$$



$$\begin{aligned} S &\Rightarrow 1A \\ &\Rightarrow 10B \\ &\Rightarrow 100S \\ &\Rightarrow 1001A \\ &\Rightarrow 10011A. \end{aligned}$$

$$\begin{aligned} S &\rightarrow 0S|1A \\ A &\rightarrow 0B|1A \\ B &\rightarrow 0S|1C \\ C &\rightarrow 0C|1C|\epsilon \end{aligned}$$

Chomsky normal form

- A context-free grammar $G = (V, \Sigma, R, S)$ is said to be in Chomsky normal form, if every rule in R has one of the following three forms
  - 1. $A \rightarrow BC$, where A, B, C nonterminals, $B \neq S$, and $C \neq S$.
  - 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal.
  - 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Theorem: For every CFG L, there exists a CFG in Chomsky normal form whose language is L.

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Theorem: For every CFG L, there exists a CFG in Chomsky normal form whose language is L.
- We can always modify a given CFG into Chomsky normal form in 5 steps.

- 1. $A \to BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \to a$, where A is a nonterminal and a is a terminal
- 3. $S \to e$, where S is the start variable.

Chomsky normal form

- We can always modify a given CFG into Chomsky normal form in 5 steps.

  - Step 1: Eliminate the start variable from the right-hand side of the rules.

  - $G_1 = (V_1, \Sigma, R_1, S_1)$,
  - New start variable S1 is the start variable
  - $V_1 = V \cup \{S_1\}$,
  - $R_1 = R \cup \{S_1 \rightarrow S\}$.

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- We can always modify a given CFG into Chomsky normal form in 5 steps.
  - Step 2: Eliminate all rules of the form $A \to e$ for $A \neq S$

1. Remove $A \to e$
2. Patch the rules such that:

   a). $B \to A$, add the rule $B \to e$ unless this rule has already been deleted

   b). $B \to uAv$ (where u and v are strings that are not both empty), add the rule $B \to uv$;

   c). $B \to uAvAw$ (where u, v, w are strings), add the rules $B \to uvw$, $B \to uAvw$, and $B \to uvAw$; if $u = v = w = e$ and the rule $B \to e$ has already been deleted, then we do not add the rule $B \to e$;

   d). treat rules in which A occurs more than twice on the right-hand side in a similar fashion

- 1. $A \to BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \to a$, where A is a nonterminal and a is a terminal
- 3. $S \to e$, where S is the start variable.

Chomsky normal form

- We can always modify a given CFG into Chomsky normal form in 5 steps.

  - Step 3: Eliminate all rules of the form $A \to B$ for nonterminals A, B

    1. Remove $A \to B$
    2. Patch the rules such that:

       a). $B \to u$, add the rule $A \to u$ unless this rule has already been deleted

- 1. $A \to BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \to a$, where A is a nonterminal and a is a terminal
- 3. $S \to e$, where S is the start variable.

Chomsky normal form

- We can always modify a given CFG into Chomsky normal form in 5 steps.
  - Step 4: Eliminate all rules having more than 2 symbols on the right

1. Remove $A \rightarrow u_1 u_2 \cdots u_k$
2. Patch the rules such that:

$$
\begin{array}{lcl}
A & \rightarrow & u_1 A_1 \\
A_1 & \rightarrow & u_2 A_2 \\
A_2 & \rightarrow & u_3 A_3 \\
 & \vdots & \\
A_{k-3} & \rightarrow & u_{k-2} A_{k-2} \\
A_{k-2} & \rightarrow & u_{k-1} u_k
\end{array}
$$

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- We can always modify a given CFG into Chomsky normal form in 5 steps.

  - Step 5: Eliminate $A \to uv$ where u, v are not both nonterminals

1. Remove $A \to u_1 u_2$
2. Patch the rules such that:

   a). u_1 terminal, u_2 nonterminal, then add $A \to U_1 u_2$, $U_1 \to u_1$

   b). u_1 nonterminal, u_2 terminal, then add $A \to u_1 U_2$, $U_2 \to u_2$

   c). u_1, u_2 different terminals, then add $A \to U_1 U_2$, $U_1 \to u_1$, $U_2 \to u_2$

   d). u_1, u_2 same terminal, then add $A \to U_1 U_1$, $U_1 \to u_1$

- 1. $A \to BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \to a$, where A is a nonterminal and a is a terminal
- 3. $S \to e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0, 1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

$$A \rightarrow BAB|B|e$$

$$B \rightarrow 00|e$$

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0,1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

$$A \rightarrow BAB|B|e$$

$$B \rightarrow 00|e$$

- Step 1: Eliminate the start variable from the right-hand side of the rules.

$$S \rightarrow A$$
$$A \rightarrow BAB|B|e$$
$$B \rightarrow 00|e$$

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0,1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

$S \rightarrow A$

$A \rightarrow BAB|B|e$ After Step 1

$B \rightarrow 00|e$

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0,1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

$S \rightarrow A$
$A \rightarrow BAB|B|e$ After Step 1
$B \rightarrow 00|e$

- Step 2: Eliminate all rules of the form $A \rightarrow e$ for $A \neq S$

1. Remove $A \rightarrow e$
2. Patch the rule:

   $S \rightarrow A$, add $S \rightarrow e$

   $A \rightarrow BAB$, add $A \rightarrow BB$

3. Remove $B \rightarrow e$

   $A \rightarrow BAB$, add $A \rightarrow AB$, $A \rightarrow BA$

   $A \rightarrow B$, add $A \rightarrow e$, but is deleted already, do not add

   $A \rightarrow BB$, add $A \rightarrow B$, do not add $A \rightarrow e$

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0,1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

$S \rightarrow A|e$

$A \rightarrow BAB|B|BB|AB|BA$

$B \rightarrow 00$

After Step 2

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0,1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

$S \rightarrow A|e$

$A \rightarrow BAB|B|BB|AB|BA$

$B \rightarrow 00$

After Step 2

- Step 3: Eliminate all rules of the form $A \rightarrow B$ for nonterminals A, B

1. Remove $S \rightarrow A$

   Patch the rule:

   Add $S \rightarrow BAB|B|BB|AB|BA$
2. Remove $S \rightarrow B$

   Add $S \rightarrow 00$
3. Remove $A \rightarrow B$

   Add $A \rightarrow 00$

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0,1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

$S \rightarrow e|BAB|BB|AB|BA|00$

$A \rightarrow BAB|BB|AB|BA|00$

$B \rightarrow 00$

After Step 3

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0{,}1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

$S \rightarrow e|BAB|BB|AB|BA|00$
$A \rightarrow BAB|BB|AB|BA|00$
$B \rightarrow 00$

After Step 3

- Step 4: Eliminate all rules having more than 2 symbols on the right

1. Remove $S \rightarrow BAB$
   Add $S \rightarrow BA_1$ and $A_1 \rightarrow AB$
2. Remove $A \rightarrow BAB$
   Add $A \rightarrow BA_2$ and $A_2 \rightarrow AB$

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0{,}1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

$S \rightarrow e|BB|AB|BA|00|BA_1$
$A \rightarrow BB|AB|BA|00|BA_2$
$B \rightarrow 00$
$A_1 \rightarrow AB$
$A_2 \rightarrow AB$

After Step 4

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0,1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

  $S \rightarrow e|BB|AB|BA|00|BA_1$
  $A \rightarrow BB|AB|BA|00|BA_2$
  $B \rightarrow 00$
  $A_1 \rightarrow AB$
  $A_2 \rightarrow AB$

After Step 4

- Step 5: Eliminate $A \rightarrow uv$ where u, v are not both nonterminals

  1. Remove $S \rightarrow 00$
     Add $S \rightarrow A_3A_3$ and $A_3 \rightarrow 0$
  2. Remove $A \rightarrow 00$
     Add $A \rightarrow A_4A_4$ and $A_4 \rightarrow 0$
  3. Remove $B \rightarrow 00$
     Add $B \rightarrow A_5A_5$ and $A_5 \rightarrow 0$

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.

Chomsky normal form

- Example: $G = (V, \Sigma, R, A)$, where $V = \{A, B, 0,1\}, \Sigma = \{0, 1\}$, start variable A, Rules:

$$S \rightarrow e|BB|AB|BA|A_3A_3|BA_1$$
$$A \rightarrow BB|AB|BA|A_4A_4|BA_2$$
$$B \rightarrow A_5A_5$$
$$A_1 \rightarrow AB$$
$$A_2 \rightarrow AB$$
$$A_3 \rightarrow 0$$
$$A_4 \rightarrow 0$$
$$A_5 \rightarrow 0$$

After Step 5

- 1. $A \rightarrow BC$, where A, B, C are nonterminals, $B \neq S$, and $C \neq S$.
- 2. $A \rightarrow a$, where A is a nonterminal and a is a terminal
- 3. $S \rightarrow e$, where S is the start variable.