# Part II Chapter 5

## The Pumping Lemma and Closure properties for Context-free Languages

## The pumping Lemma for CFLs

- **Issue: Is there any language not representable by CFGs ?**

**Ans: yes! Ex: $\{a^n b^n c^n | n > 0\}$. But how to show it ?**

- **For regular languages:**
  - **we use the pumping lemma that utilizes the "finite-state" property of finite automata to show the non-regularity of a language.**
- **For CFLs:**
  - **can we have analogous result for CFLs ?**
  - **==> Yes! But this time uses the property of parse tree instead of the machine (i.e., PDAs ) recognizing them.**

## Minimum height of parse trees for an input string

- Definition: Given a (parse) tree T,
- h(T) $=_{def}$ the height of T, is defined to be the distance of the longest path from the root to its leaves.
  - Ex: a single node tree has height 0,
  - $h(T_1) = m$ and $h(T_2) = n$ ==> $h( (\text{root } T_1\ T_2) ) = \max(m,n) + 1$.

- Lemma 5.1:

  G: a CFG in Chomsky Normal Form ;

  D = $A \rightarrow^*_G \omega$ a derivation whose parse tree $T_D$ has height n, where $A \in N$ and $\omega \in \Sigma^*$. Then

  $|\omega| \le 2^{n-1}$. [i.e, height = n (or ≤ n) => width ≤ $2^{n-1}$. ]

Note: since G is in cnf, every node of $T_D$ has at most two children, hence $T_D$ is a binary tree.

Pf: By ind. on the height n.

## Shallow trees cannot have many leaves

- Basis: $n = 1$ (not 0 since $A \neq \omega$ )

  Then $D : A \text{-->}_G a$ (or $S \text{-->}_G \varepsilon$). ==> $h(T_D) = 1$ and $|a| \leq 2^{1-1}$ .

Inductive case: $n = k + 1 > 1$. Then $\exists$ B, C, $D_1$, $D_2$ s.t.

$D : A \text{-->}_G BC \text{-->}^*_G \omega$ and $D_1$: $B \text{-->}^*_G \omega_1$ , $D_2$: $C \text{-->}^*_G \omega_2$ s.t.

$\omega = \omega_1\omega_2$ and $T_D = (A\ T_{D1}\ T_{D2})$ and $\max(h(T_{D1}), h(T_{D2})) = k$.

By ind. hyp., $|\omega_1| \leq 2^{h(T_{D1})-1} \leq 2^{k-1}$ and $|\omega_2| \leq 2^{h(T_{D2})-1} \leq 2^{k-1}$

Hence $\omega = |\omega_1| + |\omega_2| \leq (2^{k-1} + 2^{k-1}) = 2^{n-1}$. QED

Lemma 5.2: G: a CFG in cnf;

$S \text{-->}^*_G w$ in $\Sigma^*$: a derivation with parse tree T.

If $|w| \geq 2^n$ ==> $h(T) \geq n + 1$.

Pf: Assume $h(T) \leq n$

==> $|w| \leq 2^{n-1} < 2^n$ --- by lemma 5.1

==> a contradiction !! QED

## The pumping lemma for CFLs

- **Theorem: 5.3: L : a CFL. Then $\exists$ k > 0 s.t. for all member z of L of length $\geq$ k, there must exist a decomposition of z into uvwxy (i.e., z = uvwxy) s.t.**
  **(1). $|vwx| \leq k$,**
  **(2). $|v| + |x| > 0$ and**
  **(3). $uv^iwx^iy \in L$ for any $i \geq 0$.**

- **Formal rephrase of Theorem 5.3: $(L \in CFL) \Rightarrow$**

  **$\exists k>0\ \forall z \in L\ ( |z| \geq k \Rightarrow$**

  **$\exists u \exists v \exists w \exists x \exists y\ (( z = uvxyz) \wedge (1) \wedge (2) \wedge (3))\quad ))$.**

# Contrapositive form of the pumping lamma

- Contrapositive form of Theorem 5.3:
  -  (Recall that ~q => ~ p is the contrapositive of p => q)
  -  Let $Q =_{def} \exists k>0 \forall z \in L$ ( $|z| \geq k$ =>
    $\exists u \exists v \exists w \exists x \exists y$ (( $z = uvxyz$) /\ (1) /\ (2) /\ (3)) )).

  Then ~ $Q = \forall k>0\ \exists z \in L$ ( $|z| \geq k$ /\
  $\forall u \forall v \forall w \forall x \forall y$ ( ($z = uvxyz$)/\(1)/\(2)) =>~(3)) )).

  = $\forall k>0 \exists z \in L$ ( $|z| \geq k$ /\
  $\forall u \forall v \forall w \forall x \forall y$ ( ($z = uvxyz$)/\(1)/\(2)) =>
  $\exists i \geq 0\ uv^iwx^iy \notin L$ ))

  = $\forall k>0\ \exists z \in L$ ( $|z| \geq k$ /\
  $\forall uvwxy=z$ ( (1)/\(2) => $\exists i \geq 0\ uv^iwx^iy \notin L$) ).

i.e., for all k > 0 there exists a member z of L with length ≥ k s.t. for any decomposition of z into uvwxy s.t. (1) /\ (2) hold, then there must exist i ≥ 0 s.t. $uv^iwx^iy \notin L$.

The contrapotive form of Theorem 5.3 : Given a language L, If ~Q then L is not context free.

## Game-theoretical form of the pumping lamma:

| ~ Q: | Game-theoretical argument: (to show ~Q true) |
|---|---|
| $\forall k>0$ | 1. D picks any $k > 0$ |
| $\exists z\in L$ $\|z\| \geq k$ /\ | 2. Y pick a $z\in L$ with length $\geq k$ |
| $\forall uvwxy=z$ (1)/\(2) => | 3. D decompose z into uvwxy with $\|vwx\| \leq k$ /\ $\|v\| + \|x\| > 0$ |
| $\exists i$ ($i \geq 0$ /\ $uv^iwx^iy \notin L$) . | 4. Y pick a number $i \geq 0$ |
| | 5. Y win iff ($uv^iwx^iy \notin L$ or D fails to pick k or decompose z at step1&3) |

Notes:

0. If Y has a strategy according to which he always win the game, then ~Q is true, otherwise ~Q is false.
1. To show that "$\exists x$ P" is true, it is Your responsibility to give a witness c s.t. P is indeed true for that individual c. if Your opponent, who always tries to win you, cannot show that P(c) is false then You wins.
2. On the contrary, to show that "$\forall x$ P" is true, for any value c given by your opponent, who always tries to win you and hence would never give you value that is true for P provided he knows some value is false for P, You must show that P(c) is true.

## The set of prime numbers is not context-free

Ex5.1:PRIME $=_{def}$ $\{a^k \mid k$ is a prime number $\}$ is not context-free.

Pf: The following is a winning strategy for Y:

1. Suppose D picks k > 0 // for any k picked by D
2. Y picks $z = a^p$ where p is any prime number >k+2 (note p>3) ( obviously $z \in$ PRIME and $|z| \geq k$ ).
3. Suppose D decompose z into $a^u\underline{a^v a^w a^x}a^y$ with $v + x > 0$ /\ $v + w + x \leq k$
4. Y pick $i = u + w + y$ // $= p-(v+x) > k+2 - k = 2$

Now $a^u a^{vi} a^w a^{xi} a^y = a^{u+w+y} a^{(v+x)i} = a^i a^{(v+x)i} = a^{(v+x+1)i}$. Since i>2 and $v+x+1 \geq 2$, $a^{(v+x+1)i} \notin$ PRIME.

==> Y win. Since Y always win the game no matter what k is chosen and how z is decomposed at step 1&3, by the game-theoretical argument, PRIME is not context-free. QED

## Additional example

**Ex 5.2: Let A = $\{a^n b^n c^n \mid n > 0\}$ is not context-free.**

**Pf: Consider the following strategy of Y in the game:**

1. D picks $k > 0$
2. Y pick $z = a^k b^k c^k$ // obviously $z \in A$ and $|z| \geq k$
3. Suppose D decompose z into uvwxy with
   $|vx| > 0 \wedge |vwx| \leq k$
4. Y pick i = 0 ==> who wins ?

case1: vwx = $a^J$ (or $b^J$ or $c^J$) where J = |vwx|

==> in $\alpha = uv^0wx^0y$, $\#a(\alpha) < \#b(\alpha) = \#c(\alpha)$ ==> $uv^0wx^0y \notin A$

The other two cases ($b^J$ or $c^J$)are similar.

case2: vwx = $a^I b^J$ (or $b^I c^J$) with I + J = |vwx|.

==> $uv^0wx^0y$ decreases only occurrences of (a or b) or (b or c) but not c (or a) ==> $uv^2wx^2y \notin A$

In all cases $uv^2wx^2y \notin A$ So Y always win and A $\notin$ CFL. QED

# Proof of the pumping lemma

pf: Let $G = (N,S,P,S)$ be any CFG in cnf s.t. $L= L(G)$.

Suppose $|N| = n$ and let $k = 2^n$.

Now for any $z \in L(G)$ if $|z| \geq k$, by Lem 5.2, $\exists$ a parse tree T for z with $h(T) = m \geq n+1$. Now let

$$P = X_0 X_1 \ldots X_m$$

be any longest path from the root of T to a leaf of T.

Hence 1. $X_0 = S$ is the start symbol

2. $X_0, X_1, \ldots X_{m-1}$ are nonterminal symbols and

3. $X_m$ is a terminal symbol.

Since $X_0 X_1 \ldots X_{m-1}$ has $m > n$ nodes, by the pigeon-hole principle, there must exist $i \neq j$ s.t. $X_i = X_j$

Now let $l < m-1$ be the largest number s.t. $X_{l+1}, \ldots X_{m-1}$ consist of distinct symbols and $X_l = X_J$ for some $l<J< m$.

Let $X_l = X_J = A$.

## Proof of the pumping lemma (cont'd)

Let $T_I$ be the subtree of T with root $X_I$ and
$T_J$ the subtree of T with root $X_J$

Let yield($T_J$) = w (hence $X_J \rightarrow^+_G w$ or $A \rightarrow^+_G w$ --- (1) )

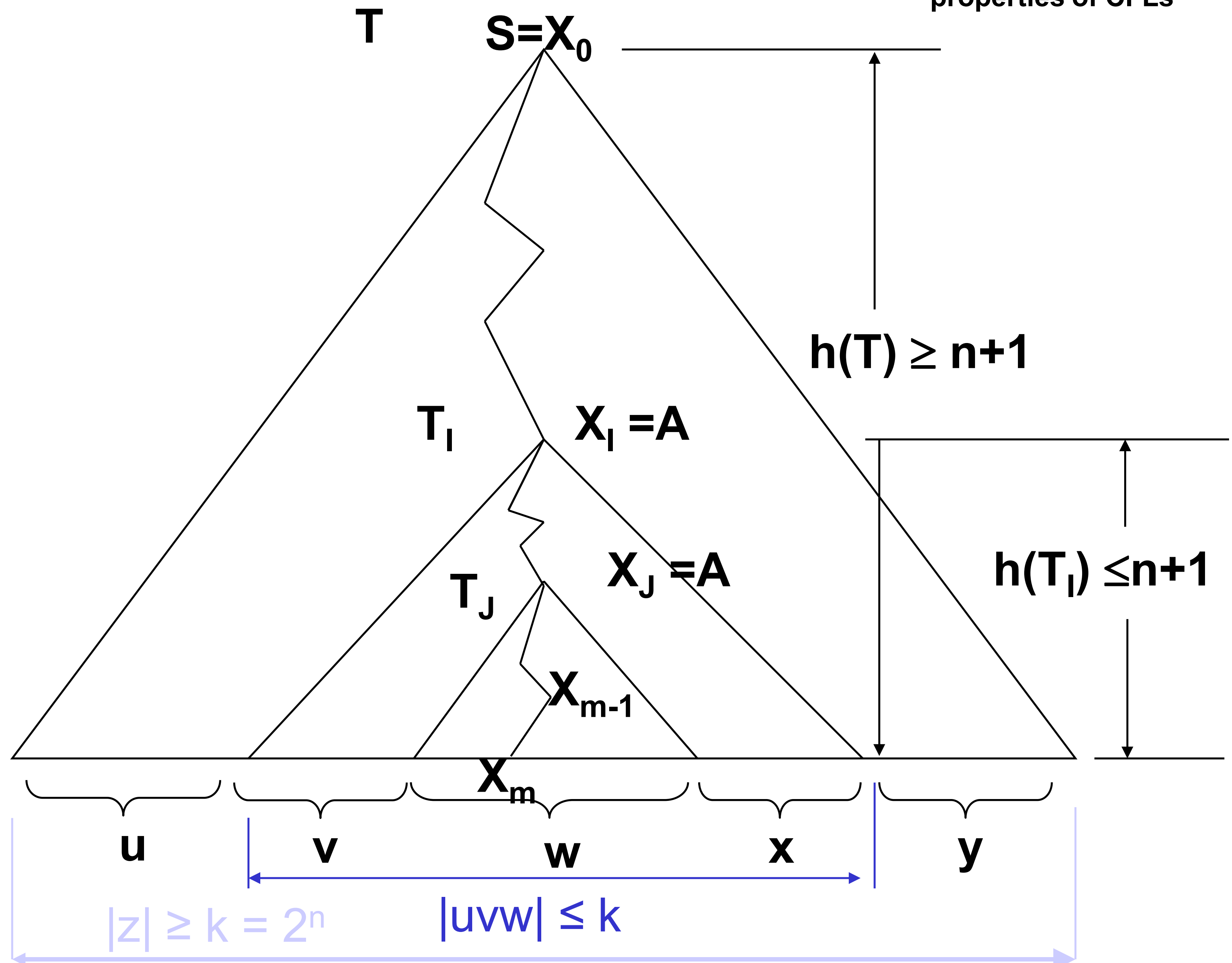
Since $T_J$ is a subtree of $T_I$,
$X_I \rightarrow^+_G v X_J x$ for some v,x in $\Sigma^*$. hence $A \rightarrow^+_G vAx$ --- (2)

Also note that since G is in cnf form it is impossible that
v = x = $\varepsilon$. (o/w $X_I \rightarrow^+ X_I$ implies existence of unit rule or $\varepsilon$-rule.

Since $T_I$ is a subtree of T,
S= $X_0$ -->$^*_G$ u $X_I$ y = u A y for some u,y in $\Sigma^*$.
-->$^*_G$ u $v^i$ A $x^i$ y ---- apply (2) i times
-->$^*_G$ u $v^i$ w $x^i$ y ---- apply (1).

Hence u $v^i$ w $x^i$ y $\in$ L for any $i \geq 0$.

Also note that since $X_I$ …. $X_m$ is the longest path in subtree $T_I$ and has length $\leq$ n+1, h($T_I$) = length of its longest path $\leq$ n+1.
==> (by lem 5.1) |vwx| = |yield($T_I$)| $\leq 2^{h(T_I)-1} = 2^n$ = k. QED


T
S=X0
h(T) ≥ n+1
TI
XI =A
TJ
XJ =A
h(TI) ≤n+1
Xm-1
Xm
u
v
w
x
y
|uvw| ≤ k
|z| ≥ k = 2n

## Example:

Ex5.3: $B = \{a^i b^j a^i b^j \mid i,j > 0\}$ is not context free.

Pf: Assume B is context-free.

Then by the pumping lemma, $\exists$ k > 0 s.t. $\forall z \in B$ of length $\geq k$,

$\exists$ uvxyz = z s.t. $|vwx| \leq k \wedge |vx| > 0 \wedge uv^iwx^iy \in B$ for any $i \geq 0$.

Now for any given k > 0, let $z = a^k b^k a^k b^k$ ---( ** ).

Let z = *uvwxy* be any decomposition with $|vwx| \leq k \wedge |vx| > 0$.

case1: $vwx = a^J$ (or $b^J$), $1 \leq J \leq k$

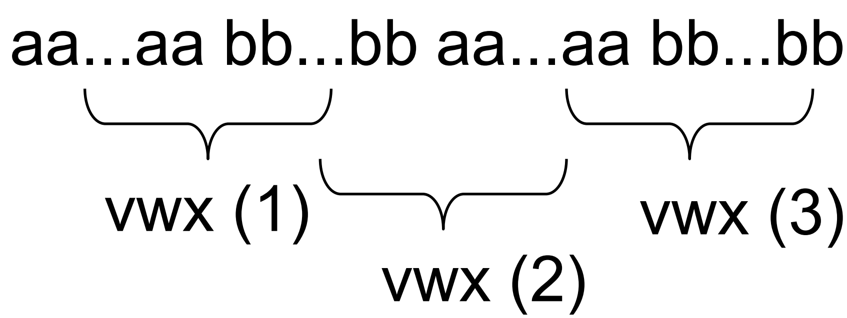
$\implies a^J < v^2wx^2 < a^{2J} \implies u\,v^2wx^2\,y \notin B$

case2: $vwx = a^J b^I$ (or $b^I a^J$), $1 \leq I + J \leq k$, $I > 0$, $J > 0$

==> For the string $uv^2wx^2y$, in all cases (1&2 &3, see next slide) only the first $a^kb^k$ or the last $a^kb^k$ or the middle $b^ka^k$ of $z = a^kb^ka^kb^k$ is increased $\implies u\,v^2wx^2\,y \notin B$

This shows that the statement (**) is not true for B.

Hence by the pumping lemma, B is not context free. QED


aa...aa bb...bb aa...aa bb...bb
vwx (1)
vwx (2)
vwx (3)

## Closure properties of CFLs

**Theorem 5.2: CFLs are closed under union, concatenation and Kleene's star operation.**

**Pf: Let $L_1 = L(G_1)$, $L_2 = L(G_2)$ : two CFLs generated by CFG $G_1$ and $G_2$, respectively, where $G_1 = (N_1, \Sigma_1, S_1, P_1)$ and $G_2 = (N_2, \Sigma_2, S_2, P_2)$.**

**==> Then**

1. $L_1 \cup L_2 = L(G')$ where $G' = (N_1 \cup N_2, \Sigma_1 \cup \Sigma_2, S', P')$ has rules:
    -  $P' = P_1 \cup P_2 \cup \{S' \rightarrow S_1; S' \rightarrow S_2\}$

2. $L_1 L_2 = L(G'')$ where $G'' = (N_1 \cup N_2, \Sigma_1 \cup \Sigma_2, S'', P'')$ has rules:
    -  $P'' = P_1 \cup P_2 \cup \{S'' \rightarrow S_1 S_2\}$

3. $L_1^* = L(G''')$ where $G''' = (N_1, \Sigma_1, S''', P''')$ has rules:
    -  $P''' = P_1 \cup \{S''' \rightarrow \varepsilon \mid S_1 S'''\}$

## Non-closure properties of CFLs

- are CFLs closed under complementation ?
  - i.e., L is context free => $\Sigma^*$ - L is context free ?
  - Ans : No.
- The set $L_1$ = {a,b}* - {ww | w $\in$ {a,b}*} is context free but its complement {ww | w $\in$ {a,b}*} is known to be not Context-free.
- Exercise: Design a CFG for $L_1$ .

Hint: x $\in$ $L_1$ iff

(1) |x| is odd or

(2) x = yazybz' or ybzyaz' for some y,z,z' $\in$ {a,b}*

with |z|=|z'|, which also means

x = yay'ubu' or yby'uau' for some y,y',u,u' $\in$ {a,b}*

with |y|=|y'| and |u|=|u'|.

## Non-closure properties of CFLs

- are CFLs closed under intersection ?
  - i.e., $L_1$ and $L_2$ context free => $L_1 \cap L_2$ is context free ?
  - Ans : No.

- Ex: Let $L_1 = \{a^i b^+ a^i b^+ \mid i > 0\}$ and
  - $L_2 = \{ a^+ b^j a^+ b^j \mid j > 0 \}$.
  - $L_1$ and $L_2$ are two CFLs.
  - But $L_1 \cap L_2 = B = \{ a^i b^j a^i b^j \mid i,j > 0 \}$ is not context free.

- CFL Language is not closed under intersection. But how about CFL and RL ?

Exercise: Let L be a CFL and R a Regular Language. Then $L \cap R$ is context free.

Hint: Let $M_1$ be a PDA accept L by final state and $M_2$ a FA accepting R, then the product machine $M_1XM_2$ can be used to accept $L \cap R$ by final state. The definition of the product PDA $M_1XM_2$ is similar to that of the product of two FAs.