Chomsky normal form EXAMPLE

G: $S \to YZ \mid ZZZ$

$X \to \varepsilon$

$Y \to XY$

$Z \to \varepsilon \mid aZ$

① Eliminate useless symbols

ⓐ Eliminate symbols that cannot generate a terminal string

G': $S \to ZZZ$ since Y is not generating

$X \to \varepsilon$

$Z \to \varepsilon \mid aZ$

ⓑ Eliminate symbols that are not reachable (i.e., X)

G'': $S \to ZZZ$

$Z \to \varepsilon \mid aZ$

② Eliminate ε productions ($Z \to \varepsilon$, $S \overset{*}{\Rightarrow} \varepsilon$)

G'': $S \to ZZZ \mid ZZ \mid Z$ Note $L(G'') = L(G) \setminus \{\varepsilon\}$

$Z \to aZ \mid a$

③ Eliminate unit productions ($S \to Z$)

G''': $S \to ZZZ \mid ZZ \mid aZ \mid a$

$Z \to aZ \mid a$

④ Convert to CNF by splitting larger productions

G'''': $S \to BZ \mid ZZ \mid AZ \mid a$

$B \to ZZ$

$Z \to AZ \mid a$

$A \to a$

This is the CNF-form grammar that we obtain mechanically.

Looking closer, it can be simplified to

$S \to a \mid SS$

with $L(S) = L(a^+)$, compared to $L(G) = L(a^*) = L(a^+ + \varepsilon)$.

Decision problems.

① Emptiness:

Given: CFL L in form of a CFG G

Question: Is $L(G) = \emptyset$?

Idea: If $L(G) = \emptyset$ then S is not generating
$\Rightarrow S$ is useless

Easy to check?

② Membership

Given CFL L, string ω

Q: Is $\omega \in L$?

Idea (Much harder than for regular languages) Assume $L = L(G)$ for CFG G

1. If $\omega = \varepsilon$, check if S is nullable
2. Construct a Chomsky normal form grammar G' for $L \setminus \{\varepsilon\}$.
3. Hw #4, problem 1
   If $|\omega| = n$ then $S' \overset{*}{\Rightarrow} \omega$ in exactly $2n-1$ steps.
4. Try all possible derivations of length $2n-1$
   $\leq |P'|^{2n-1}$ possibilities, $|P'|$ = # productions in G'.

Better approach:

Dynamic Programming $V_{ij} = \{X \in V \mid X \overset{*}{\Rightarrow} \omega_i \omega_{i+1} \dots \omega_j\}$

③ Equality Given CFGs G_1, G_2

Question: Is $L(G_1) = L(G_2)$?

Claim: There is no decision procedure? (show later)

Remark:

1. Easy to test equality for regular languages
   $L_1 = L_2 \iff (L_1 \cap \overline{L_2}) \cup (\overline{L_1} \cap L_2) = \emptyset$.
2. [Sénizergues, 2001]:
   Many other undecidable properties of CFGs, but
   algorithm for deciding if $L(M) = L(M')$
   for deterministic PDAs M, M' 170 pages, Gödel prize
   Deciding if $L(M) \subseteq L(M')$ still impossible. [GGG66]
   or $L = M$ also OKAY [Ginsburg, Greibach 1966a]
   (L: CFL, M: reg)

Dynamic programming alg. to decide membership in a CFL

Take a CFG in CN form.

Goal: For each substring $x(i\,j) := x_i x_{i+1} \cdots x_j$

compute $V_{ij} := \{ X \in V \mid X \overset{*}{\Rightarrow} x(i\,j) \}$

Induction on $j - i$:

Base case $j = i$: Easy,

Induction:

$X \overset{*}{\Rightarrow} x(i\,j)$

must start with $X \Rightarrow BC$

where for some k, $i \le k < j$,

$B \overset{*}{\Rightarrow} x(i\,k)$, $C \overset{*}{\Rightarrow} x(k+1, j)$ ✓

$O(n)$ time where $n = |x|$

$\Rightarrow O(n^3)$ time total, if CFG size is constant.

(then also conversion to CNF is $O(1)$)

Pumping Lemma for Context-Free Languages: [H-M-U 7.2]

Question: Of $L_1 = \{a^n b^n \mid n \geq 1\}$
$L_2 = \{a^n b^{n+m} c^m \mid n, m \geq 1\}$
$L_3 = \{a^n b^{2n} c^n \mid n \geq 1\}$,
which are CFLs?

Answer: L_1, L_2 are CFLs, L_3 is not a CFL.
Intuitively, because the number of b's depends on its context, both to its left and right.

Pumping Lemma
Let L be a CFL.
Then there exists a constant N such that
For all $z \in L$ with $|z| > N$
can write $z = uvwxy$ with
a) $|vwx| \leq N$
b) $|vx| > 0$
c) $\forall j \geq 0$, $uv^j wx^j y \in L$

Pumping Lemma for regular langs
Let L be regular.
Then $\exists N$
$\forall z \in L$, $|z| > N$
$z = uvw$ with
a) $|uv| \leq N$
b) $|v| > 0$
c) $\forall j \geq 0$, $uv^j w \in L$

Application: Claim: $L_3 = \{a^n b^{2n} c^n \mid n \geq 1\}$ is not CFL.
Proof:
① Assume L_3 is CFL and apply P.L.
② Get constant $N > 0$
③ Choose $z = a^N b^{2N} c^N \in L_3$
④ Get u, v, w, x, y such that
$z = uvwxy$
and $|vwx| \leq N$, $|vx| \geq 1$
⑤ We choose $j = 0$ and claim
$z' = uv^0 wx^0 y = uwy \notin L_3$. (contradiction)

Why? Observe: $|vwx| \leq N$ implies that either vwx has no c's or it has no a's
Case I [vwx has no c's]:
$|z'| = |z| - |vx| < |z| = 4N$
but $n_c(z') = N$, so more than $\frac{1}{4}$ of its symbols are c's.
Every string in L_3 has exactly $\frac{1}{4}$ c's, so $z' \notin L_3$. ✓
Case II [vwx has no a's]: By a similar argument, $z' \notin L_3$.

Corollary: Unlike regular languages, CFLs are not necessarily closed under intersection.
$L_3 = L_2 \cap \{a^n b^m c^n \mid m \geq 0, n \geq 1\}$.

Proof of the CFL Pumping Lemma:

Given CFL L

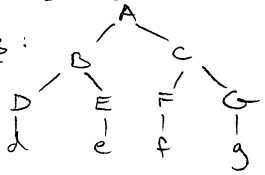
Let $G = (V, T, P, S)$ be a Chomsky normal form grammar for $L \setminus \{\varepsilon\}$.

Claim: If in a parse tree for G, all root-leaf paths have length $\leq \ell$, then the yield of the tree has $\leq 2^{\ell-1}$ terminals.

Proof: By induction in ℓ. Intuitively obvious since the tree is binary.

Worst cases: $\ell = 1$: $2^{\ell-1} = 1$

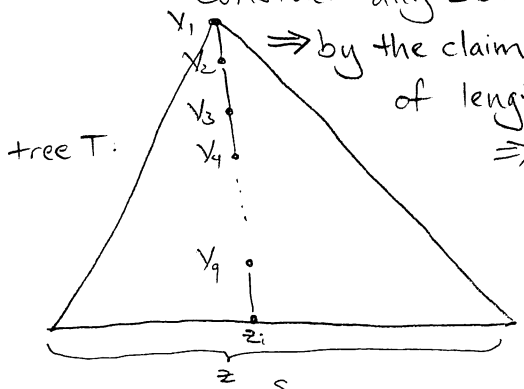
$\ell = 2$: 

$\ell = 3$: 

✓

Now define $k = |V| = \#$ variables, $N = 2^k$.

Consider any $z \in L(G)$ with $|z| > N$.

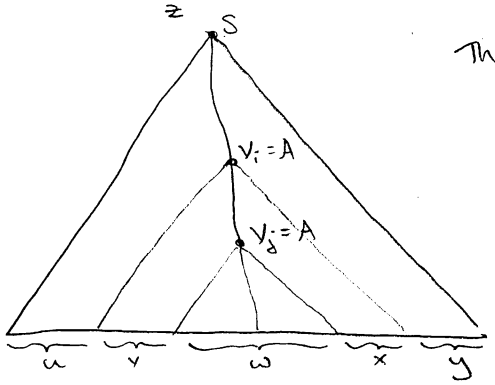


tree T.

$\Rightarrow$ by the claim, any parse tree T for z has a root-leaf path Q of length $q \geq k+1$.

$\Rightarrow$ By the pigeon-hole principle, some variable repeats along the path Q. In fact the last $k+1$ variables on Q must contain a repetition.

Let i, j be such that $q-k \leq i < j \leq q$, $V_i = V_j = A$.



Therefore $z = uvwxy$, where

$S \overset{*}{\Rightarrow} uAy \overset{*}{\Rightarrow} uvAxy \overset{*}{\Rightarrow} uvwxy$

so $A \overset{*}{\Rightarrow} vAx$, $A \overset{*}{\Rightarrow} w$

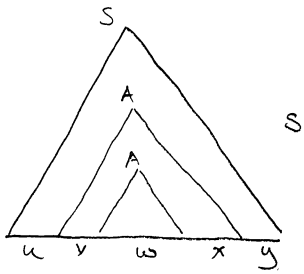
Hence $A \overset{*}{\Rightarrow} vAx \overset{*}{\Rightarrow} v^2Ax^2$

$\overset{*}{\Rightarrow} \cdots \overset{*}{\Rightarrow} v^jAx^j \overset{*}{\Rightarrow} v^jwx^j$

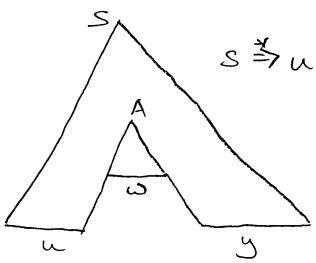
and so $\forall j \geq 0$,

$S \overset{*}{\Rightarrow} uAy \overset{*}{\Rightarrow} uv^jwx^jy$,

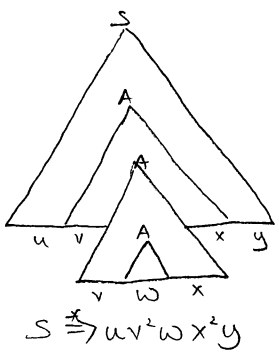
implying $uv^jwx^jy \in L(G)$. ✓



$S \overset{*}{\Rightarrow} uvwxy = z$



$S \overset{*}{\Rightarrow} uwy$



$S \overset{*}{\Rightarrow} uv^2wx^2y$

Also: $|vx| > 0$ since CNF G has no ε or unit productions.

Finally: $|vwx| \leq N = 2^k$, since $V_i = A$ has height $\leq k+1$ and so by the claim its yield vwx has length at most $2^{(k+1)-1} = 2^k$ terminals.

□