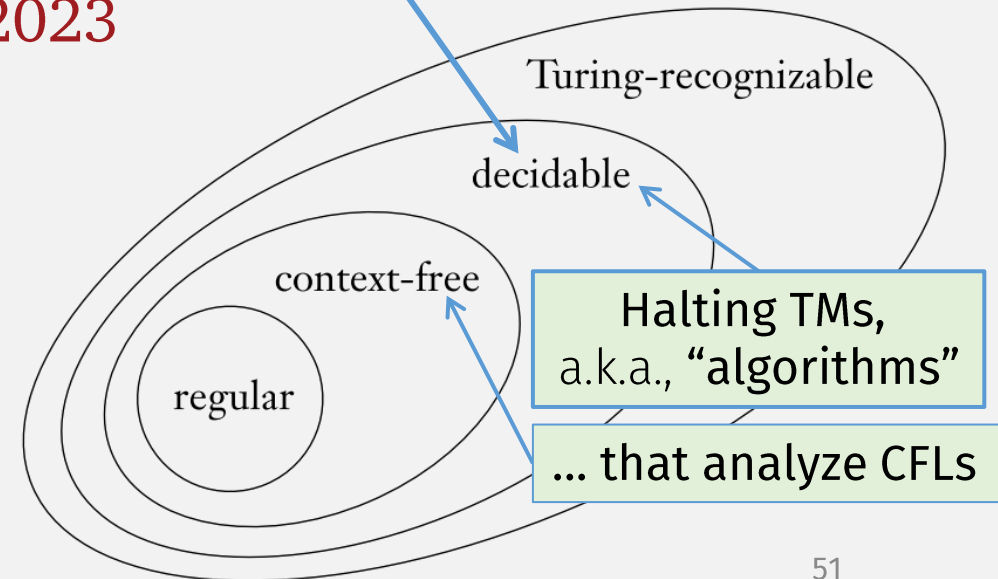


UMB CS 420

Decidability for CFLs

Wednesday, April 5, 2023



Announcements

- HW 7 in
 - ~~Due Tues April 4 11:59pm EST~~
- HW 8 out
 - Due Tues April 11 11:59pm EST

Quiz Preview

- Which of the following rules are valid for a grammar in **Chomsky Normal Form**?

Last Time: Decider Turing Machines

- 2 classes of Turing Machines
 - **Recognizers** (all TMs): may loop forever
 - TM that **loops** on an input does **not accept** that input
 - **Deciders** (subset of TMs) (**algorithms**) always halt
 - Must **accept** or **reject**
- **Decider definitions must include a termination argument:**
 - Explains (informally) why every step in the TM halts
 - (Pay special attention to loops)

Last Time: Decidable Languages About DFAs

- $A_{\text{DFA}} = \{\langle B, w \rangle \mid B \text{ is a DFA that accepts input string } w\}$
 - Decider TM: implements B DFA's extended δ fn
- $A_{\text{NFA}} = \{\langle B, w \rangle \mid B \text{ is an NFA that accepts input string } w\}$
 - Decider TM: uses **NFA**→**DFA** algorithm + A_{DFA} decider
- $A_{\text{REGEX}} = \{\langle R, w \rangle \mid R \text{ is a regular expression that generates string } w\}$
 - Decider TM: uses **Regex**→**NFA** algorithm + A_{NFA} decider

Remember:
TMs ~ programs
Creating TM ~ programming
Previous theorems ~ library

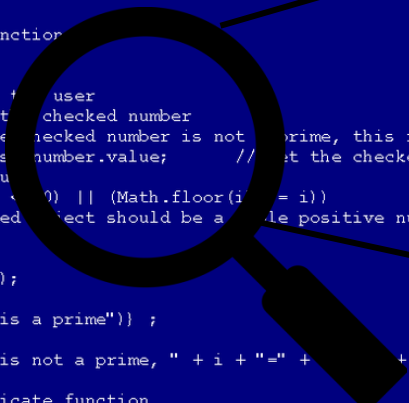
Flashback: Why Study Algorithms About Computing

To predict what programs will do
(without running them!)

Not possible for all programs! But ...

```
function check(n)
{ // check if the number n is a prime
  var factor; // if the checked number is not a prime, this is its first factor
  var c;
  factor = 0;
  // try to divide the checked number by all numbers till its square root
  for (c=2; (c <= Math.sqrt(n)); c++)
  {
    if (n%c == 0) // is n divisible by c?
      { factor = c; break }
  }
  return (factor);
} // end of check function

function communicate()
{ // communicate with the user
  var i; // i is the checked number
  var factor; // if the checked number is not a prime, this is its first factor
  i = document.primes.number.value; // get the checked number
  // is it a valid input?
  if ((isNaN(i)) || (i <= 0) || (Math.floor(i) != i))
  { alert ("The checked object should be a whole positive number"); }
  else
  {
    factor = check (i);
    if (factor == 0)
      { alert (i + " is a prime"); }
    else
      { alert (i + " is not a prime, " + i + "=" + factor + "X" + i/factor) }
  }
} // end of communicate function
```



RANSOMWARE ATTACK



???

Predicting What Some Programs Will Do ...

What if: look at simpler computation models
... like DFAs and regular languages!

Thm: E_{DFA} is a decidable language

$$E_{\text{DFA}} = \{ \langle A \rangle \mid A \text{ is a DFA and } L(A) = \emptyset \}$$

E_{DFA} is a language ... of DFA descriptions,
i.e., $(Q, \Sigma, \delta, q_0, F)$...

... where the language of each DFA ...
must be $\{\}$, i.e., DFA accepts no strings

Is there a decider that
accepts/rejects DFA descriptions ...

... by predicting something
about the DFA's language
(by analyzing its description)

The key question we are studying:
Can we **determine** something about the
runtime computation of a program,
by analyzing *only* its source code?

Analogy
DFA's description : a program's source code ::
DFA's language : a program's runtime computation

Important: don't confuse the different languages here!

Thm: E_{DFA} is a decidable language

$$E_{\text{DFA}} = \{ \langle A \rangle \mid A \text{ is a DFA and } L(A) = \emptyset \}$$

Decider:

$T =$ “On input $\langle A \rangle$, where A is a DFA:

1. Mark the start state of A .
2. Repeat until no new states get marked:
3. Mark any state that has a transition coming into it from any state that is already marked.
4. If no accept state is marked, *accept*; otherwise, *reject*.”

If loop marks at least 1 state on each iteration, then it eventually terminates because there are finite states; else loop terminates

i.e., this is a “reachability” algorithm ...

Termination argument?

... check if accept states are “reachable” from start state

Note: TM T does not “run” the DFA!

... it computes something about the DFA’s language (runtime computation) by analyzing its description (source code)

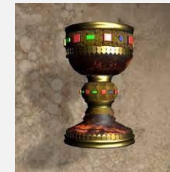
Thm: EQ_{DFA} is a decidable language

$$EQ_{\text{DFA}} = \{\langle A, B \rangle \mid A \text{ and } B \text{ are DFAs and } L(A) = L(B)\}$$

i.e., Can we compute whether
two DFAs are “equivalent”?



Replacing “**DFA**” with “**program**” =
A “**holy grail**” of computer science!



Thm: EQ_{DFA} is a decidable language

$$EQ_{\text{DFA}} = \{\langle A, B \rangle \mid A \text{ and } B \text{ are DFAs and } L(A) = L(B)\}$$

A Naïve Attempt (assume alphabet $\{\mathbf{a}\}$):

1. Simulate:

- A with input \mathbf{a} , and
- B with input \mathbf{a}
- **Reject** if results are different, else ...

2. Simulate :

- A with input \mathbf{aa} , and
- B with input \mathbf{aa}
- **Reject** if results are different, else ...

• ...

This might not terminate!
(Hence it's not a decider)

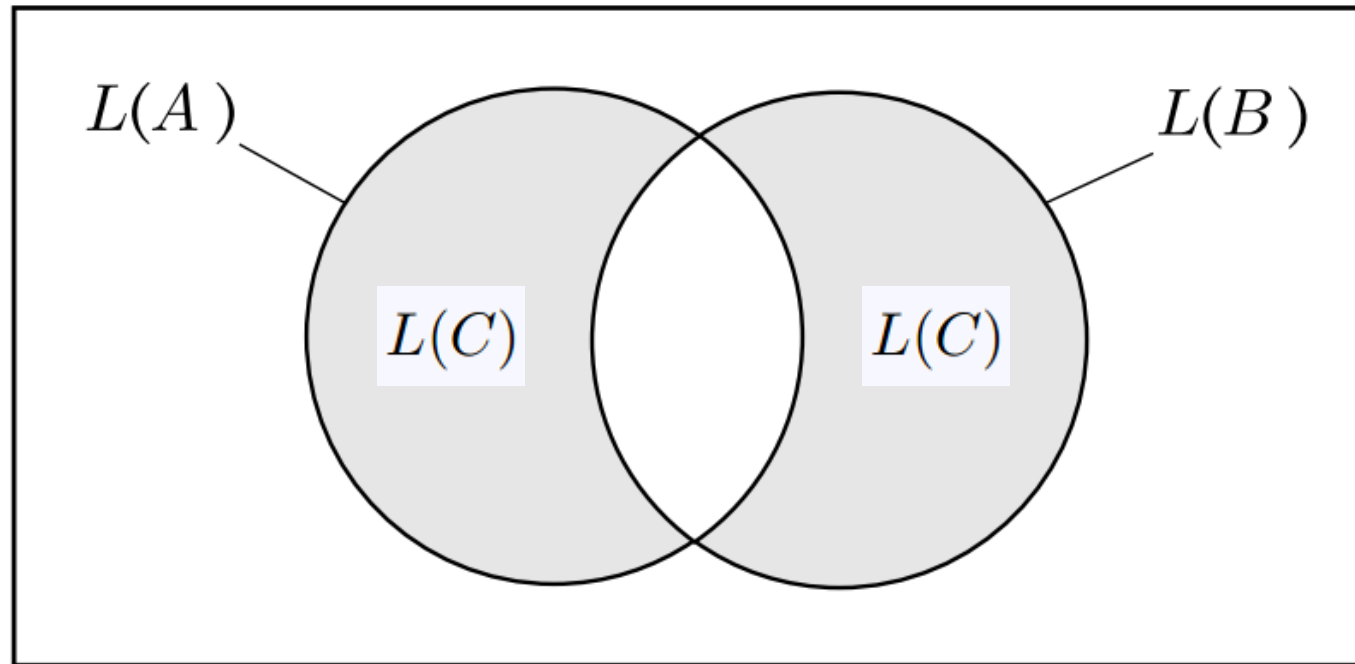
Can we compute this without
running the DFAs?

Thm: EQ_{DFA} is a decidable language

$$EQ_{\text{DFA}} = \{\langle A, B \rangle \mid A \text{ and } B \text{ are DFAs and } L(A) = L(B)\}$$

Trick: Use Symmetric Difference

Symmetric Difference



$$L(C) = \left(L(A) \cap \overline{L(B)} \right) \cup \left(\overline{L(A)} \cap L(B) \right)$$

$$L(C) = \emptyset \text{ iff } L(A) = L(B)$$

Thm: EQ_{DFA} is a decidable language

$$EQ_{\text{DFA}} = \{\langle A, B \rangle \mid A \text{ and } B \text{ are DFAs and } L(A) = L(B)\}$$

NOTE: This only works because: **negation**, i.e., set complement, and **intersection** is **closed** for regular languages

Construct **decider** using 2 parts:

1. **Symmetric Difference algo**: $L(C) = (L(A) \cap \overline{L(B)}) \cup (\overline{L(A)} \cap L(B))$
 - Construct C = Union, intersection, negation of machines A and B
2. **Decider T** (from “library”) **for**: $E_{\text{DFA}} = \{\langle A \rangle \mid A \text{ is a DFA and } L(A) = \emptyset\}$
 - Because $L(C) = \emptyset$ iff $L(A) = L(B)$

F = “On input $\langle A, B \rangle$, where A and B are DFAs:

1. Construct DFA C as described.
2. Run TM T deciding E_{DFA} on input $\langle C \rangle$.
3. If T accepts, *accept*. If T rejects, *reject*.”

Termination argument?

Predicting What Some Programs Will Do ...

microsoft.com/en-us/research/project/slam/

SLAM is a project for checking that software satisfies critical behavioral properties of the interfaces it uses and to aid software engineers in designing interfaces and software that ensure reliable and correct functioning. Static Driver Verifier is a tool in the Windows Driver Development Kit that uses the SLAM verification engine.

"Things like even software verification, this has been the Holy Grail of computer science for many decades but now in some very key areas, for example, driver verification we're building tools that can do actual proof about the software and how it works in order to guarantee the reliability." **Bill Gates, April 18, 2002.** [Keynote address at WinHec 2002](#)



Static Driver Verifier Research Platform README

Overview of Static Driver Verifier Research Platform

Static Driver Verifier (SDV) is a compile-time static verification Research Platform (SDVRP) is an extension to SDV that allows

- Support additional frameworks (or APIs) and write custom
- Experiment with the **model checking** step.

Model checking

From Wikipedia, the free encyclopedia

In **computer science**, **model checking** or **property checking** is a method for checking whether a **finite-state model** of a system meets a given **specification** (also known as **correctness**). This is typically

Its "language"

Summary: Algorithms About Regular Langs

- $A_{\text{DFA}} = \{\langle B, w \rangle \mid B \text{ is a DFA that accepts input string } w\}$
 - **Decider:** Simulates DFA by implementing extended δ function

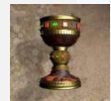
- $A_{\text{NFA}} = \{\langle B, w \rangle \mid B \text{ is an NFA that accepts input string } w\}$
 - **Decider:** Uses NFA \rightarrow DFA decider + A_{DFA} decider

- $A_{\text{REX}} = \{\langle R, w \rangle \mid R \text{ is a regular expression that generates string } w\}$
 - **Decider:** Uses RegExpr \rightarrow NFA decider + A_{NFA} decider

- $E_{\text{DFA}} = \{\langle A \rangle \mid A \text{ is a DFA and } L(A) = \emptyset\}$
 - **Decider:** Reachability algorithm

Lang of the DFA

- $EQ_{\text{DFA}} = \{\langle A, B \rangle \mid A \text{ and } B \text{ are DFAs and } L(A) = L(B)\}$



- **Decider:** Uses complement and intersection closure construction + E_{DFA} decider

Remember:
TMs ~ programs
Creating TM ~ programming
Previous theorems ~ library

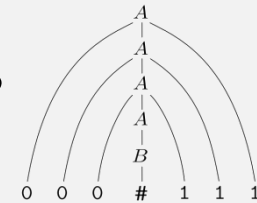
Next: Algorithms (Decider TMs) for CFLs?

- What can we predict about CFGs or PDAs?

Thm: A_{CFG} is a decidable language

$$A_{CFG} = \{ \langle G, w \rangle \mid G \text{ is a CFG that generates string } w \}$$

- This is a very practically important problem ...
- ... equivalent to:
 - **Algorithm** to parse “program” w for a programming language with grammar G ?
- A Decider for this problem could ... ?
 - Try every possible derivation of G , and check if it's equal to w ?
 - But this might never halt
 - E.g., what if there are rules like: $S \rightarrow 0S$ or $S \rightarrow S$
 - This TM would be a recognizer but not a decider

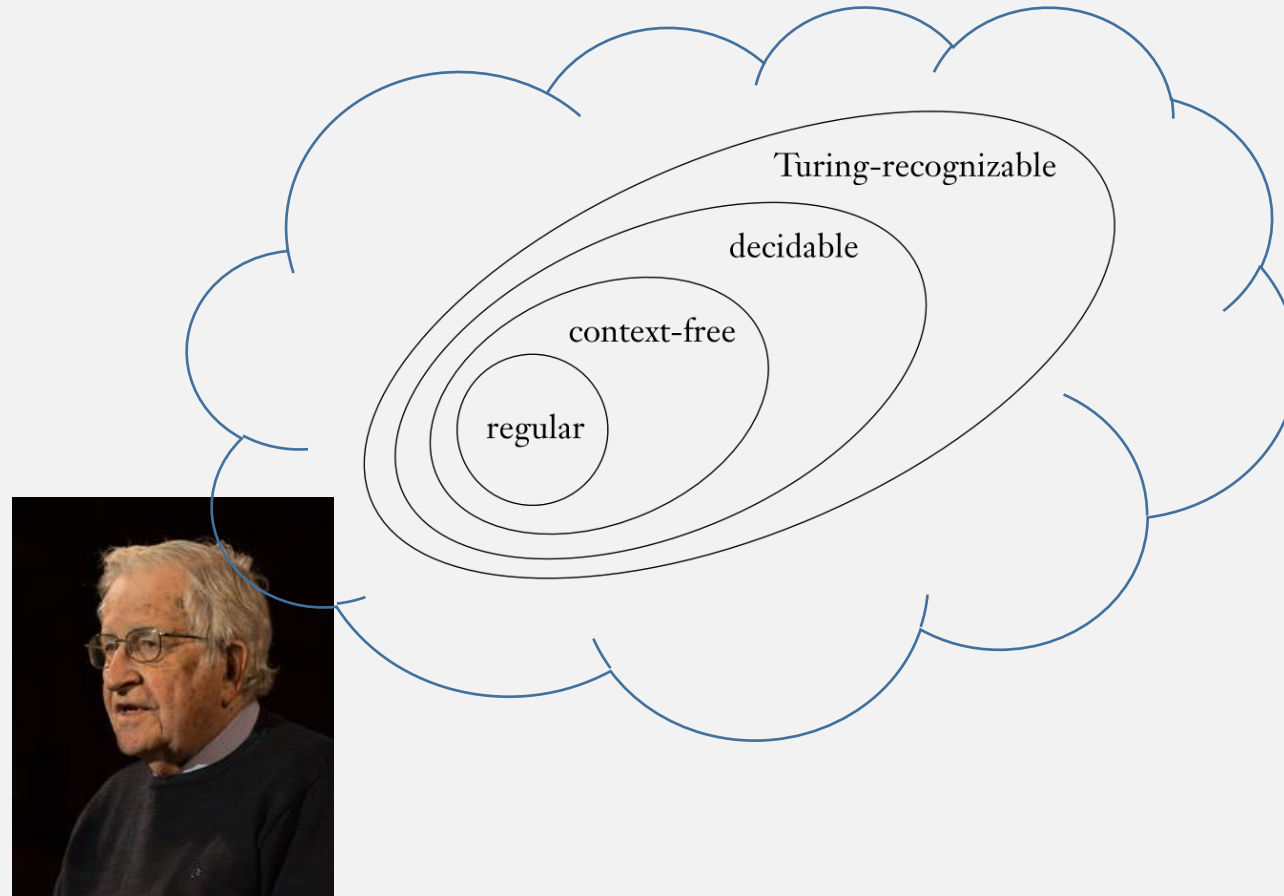


Idea: can the TM stop checking after some length?

- I.e., Is there upper bound on the number of derivation steps?

Chomsky Normal Form

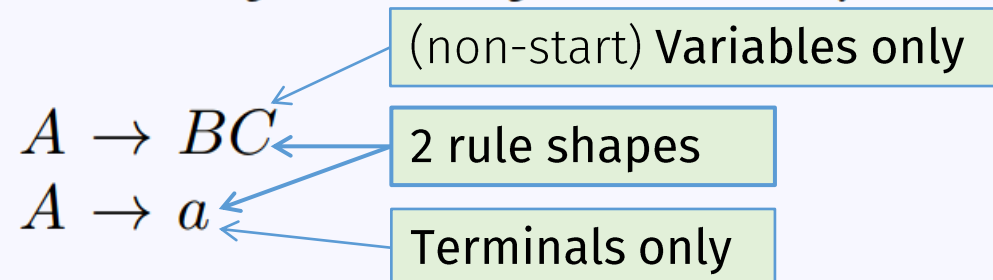
Noam Chomsky



He came up with this hierarchy of languages

Chomsky Normal Form

A context-free grammar is in *Chomsky normal form* if every rule is of the form



where a is any terminal and A , B , and C are any variables—except that B and C may not be the start variable. In addition, we permit the rule $S \rightarrow \epsilon$, where S is the start variable.

Chomsky Normal Form Example

Makes the string long enough

Convert variables to terminals

- $S \rightarrow AB$
- $A \rightarrow AB$
- $A \rightarrow a$
- $B \rightarrow b$

- To generate string of length: 2
 - Use S rule: 1 time; Use A or B rules: 2 times
 - $S \Rightarrow AB \Rightarrow aB \Rightarrow ab$
 - Derivation total steps: $1 + 2 = 3$
- To generate string of length: 3
 - Use S rule: 1 time; A rule: 1 time; A or B rules: 3 times
 - $S \Rightarrow AB \Rightarrow AAB \Rightarrow aAB \Rightarrow aaB \Rightarrow aab$
 - Derivation total steps: $1 + 1 + 3 = 5$
- To generate string of length: 4
 - Use S rule: 1 time ; A rule: 2 times; A or B rules: 4 times
 - $S \Rightarrow AB \Rightarrow AAB \Rightarrow AAAB \Rightarrow aAAB \Rightarrow aaAB \Rightarrow aaaB \Rightarrow aaab$
 - Derivation total steps: $3 + 4 = 7$
- ...

2 rule shapes

- ✓ $A \rightarrow BC$
- $A \rightarrow a$

A context-free grammar is in *Chomsky normal form* if every rule is of the form

where a is any terminal and A , B , and C are any variables—except that B and C may not be the start variable. In addition, we permit the rule $S \rightarrow \epsilon$, where S is the start variable.

Chomsky Normal Form: Number of Steps

To generate a string of length n :

$n - 1$ steps: to generate n variables

Makes the string long enough

+ n steps: to turn each variable into a terminal

Convert string to terminals

Total: $2n - 1$ steps

(A *finite* number of steps!)

Chomsky normal form

$A \rightarrow BC$ Use $n-1$ times

$A \rightarrow a$ Use n times

Thm: A_{CFG} is a decidable language

$$A_{CFG} = \{ \langle G, w \rangle \mid G \text{ is a CFG that generates string } w \}$$

Proof: create the decider:

$S =$ “On input $\langle G, w \rangle$, where G is a CFG and w is a string:

1. Convert G to an equivalent grammar in Chomsky normal form.
2. List all derivations with $2n - 1$ steps, where n is the length of w ; except if $n = 0$, then instead list all derivations with one step.
3. If any of these derivations generate w , *accept*; if not, *reject*.”

We first need to prove this is true for all CFGs!

Step 1: Conversion to Chomsky Normal Form is an algorithm ...

Step 2:

Step 3:

Termination argument?

Thm: Every CFG has a Chomsky Normal Form

Proof: Create algorithm to convert any CFG into Chomsky Normal Form

Chomsky normal form

1. Add new start variable S_0 that does not appear on any RHS
 - I.e., add rule $S_0 \rightarrow S$, where S is old start var

$A \rightarrow BC$

$A \rightarrow a$

$$\begin{aligned} S &\rightarrow ASA \mid aB \\ A &\rightarrow B \mid S \\ B &\rightarrow b \mid \epsilon \end{aligned}$$


$S_0 \rightarrow S$

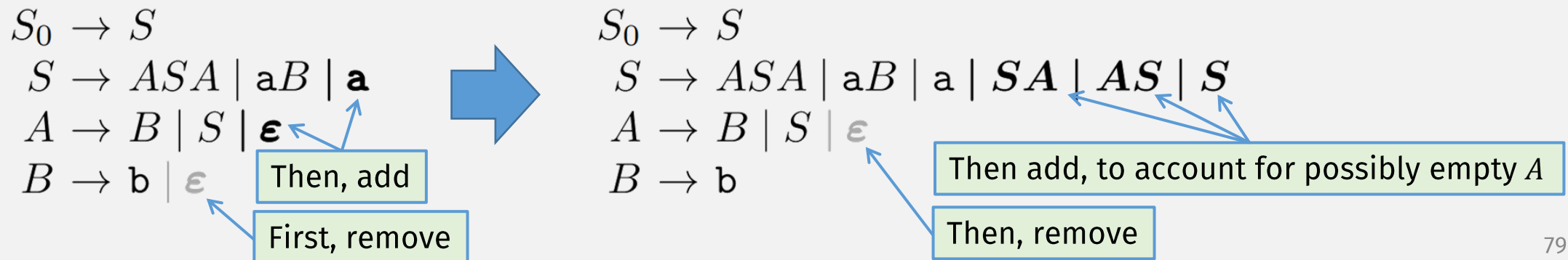
$$\begin{aligned} S &\rightarrow ASA \mid aB \\ A &\rightarrow B \mid S \\ B &\rightarrow b \mid \epsilon \end{aligned}$$

Thm: Every CFG has a Chomsky Normal Form

Chomsky normal form

1. Add new start variable S_0 that does not appear on any RHS
 - I.e., add rule $S_0 \rightarrow S$, where S is old start var
2. Remove all “empty” rules of the form $A \rightarrow \epsilon$
 - A must not be the start variable
 - Then for every rule with A on RHS, add new rule with A deleted
 - E.g., if $R \rightarrow uAv$ is a rule, add $R \rightarrow uv$
 - Must cover all combinations if A appears more than once in a RHS
 - E.g., if $R \rightarrow uAvAw$ is a rule, add 3 rules: $R \rightarrow uvAw$, $R \rightarrow uAvw$, $R \rightarrow uvw$

$A \rightarrow BC$
 $A \rightarrow a$



Thm: Every CFG has a Chomsky Normal Form

Chomsky normal form

$A \rightarrow BC$

$A \rightarrow a$

1. Add new start variable S_0 that does not appear on any RHS
 - I.e., add rule $S_0 \rightarrow S$, where S is old start var
2. Remove all “empty” rules of the form $A \rightarrow \varepsilon$
 - A must not be the start variable
 - Then for every rule with A on RHS, add new rule with A deleted
 - E.g., if $R \rightarrow uAv$ is a rule, add $R \rightarrow uv$
 - Must cover all combinations if A appears more than once in a RHS
 - E.g., if $R \rightarrow uAvAw$ is a rule, add 3 rules: $R \rightarrow uvAw$, $R \rightarrow uAvw$, $R \rightarrow uvw$
3. Remove all “unit” rules of the form $A \rightarrow B$
 - Then, for every rule $B \rightarrow u$, add rule $A \rightarrow u$

$S_0 \rightarrow S$
 $S \rightarrow ASA \mid aB \mid a \mid SA \mid AS \mid S$
 $A \rightarrow B \mid S$
 $B \rightarrow b$

Remove, no add
(same variable)

$S_0 \rightarrow S \mid ASA \mid aB \mid a \mid SA \mid AS$
 $S \rightarrow ASA \mid aB \mid a \mid SA \mid AS$
 $A \rightarrow B \mid S$
 $B \rightarrow b$

Remove, then add S RHSs to S_0

$S_0 \rightarrow ASA \mid aB \mid a \mid SA \mid AS$
 $S \rightarrow ASA \mid aB \mid a \mid SA \mid AS$
 $A \rightarrow S \mid b \mid ASA \mid aB \mid a \mid SA \mid AS$
 $B \rightarrow b$

Remove, then add S RHSs to A

Termination argument of this algorithm?

Thm: Every CFG has a Chomsky Normal Form

Chomsky normal form

$$A \rightarrow BC$$

$$A \rightarrow a$$

1. Add new start variable S_0 that does not appear on any RHS

- i.e., add rule $S_0 \rightarrow S$, where S is old start var

2. Remove all “empty” rules of the form $A \rightarrow \varepsilon$

- A must not be the start variable
- Then for every rule with A on RHS, add new rule with A deleted
 - E.g., if $R \rightarrow uAv$ is a rule, add $R \rightarrow uv$
- Must cover all combinations if A appears more than once in a RHS
 - E.g., if $R \rightarrow uAvAw$ is a rule, add 3 rules: $R \rightarrow uvAw$, $R \rightarrow uAvw$, $R \rightarrow uvw$

$$\begin{aligned} S_0 &\rightarrow \boxed{ASA} \mid \boxed{aB} \mid a \mid SA \mid AS \\ S &\rightarrow ASA \mid aB \mid a \mid SA \mid AS \\ A &\rightarrow b \mid ASA \mid aB \mid a \mid SA \mid AS \\ B &\rightarrow b \end{aligned}$$



3. Remove all “unit” rules of the form $A \rightarrow B$

- Then, for every rule $B \rightarrow u$, add rule $A \rightarrow u$

4. Split up rules with RHS longer than length 2

- E.g., $A \rightarrow wxyz$ becomes $A \rightarrow wB$, $B \rightarrow xC$, $C \rightarrow yz$

5. Replace all terminals on RHS with new rule

- E.g., for above, add $W \rightarrow w$, $X \rightarrow x$, $Y \rightarrow y$, $Z \rightarrow z$

$$\begin{aligned} S_0 &\rightarrow \boxed{AA_1} \mid \boxed{UB} \mid a \mid SA \mid AS \\ S &\rightarrow AA_1 \mid UB \mid a \mid SA \mid AS \\ A &\rightarrow b \mid AA_1 \mid UB \mid a \mid SA \mid AS \\ A_1 &\rightarrow \boxed{SA} \\ \boxed{U} &\rightarrow a \\ B &\rightarrow b \end{aligned}$$

Thm: A_{CFG} is a decidable language

$$A_{\text{CFG}} = \{ \langle G, w \rangle \mid G \text{ is a CFG that generates string } w \}$$

Proof: create the decider:

$S =$ “On input $\langle G, w \rangle$, where G is a CFG and w is a string:

We first
need to
prove this is
true for all
CFGs!



1. Convert G to an equivalent grammar in Chomsky normal form.
2. List all derivations with $2n - 1$ steps, where n is the length of w ; except if $n = 0$, then instead list all derivations with one step.
3. If any of these derivations generate w , *accept*; if not, *reject*.”

Termination argument:

Step 1: any CFG has only a finite # rules

Step 2: $2n-1 =$ finite # of derivations to check

Step 3: checking finite number of derivations

Thm: E_{CFG} is a decidable language.

$$E_{\text{CFG}} = \{ \langle G \rangle \mid G \text{ is a CFG and } L(G) = \emptyset \}$$

Recall:

$$E_{\text{DFA}} = \{ \langle A \rangle \mid A \text{ is a DFA and } L(A) = \emptyset \}$$

$T =$ “On input $\langle A \rangle$, where A is a DFA:

1. Mark the start state of A .
2. Repeat until no new states get marked:
3. Mark any state that has a transition coming into it from any state that is already marked.
4. If no accept state is marked, *accept*; otherwise, *reject*.”

“Reachability” (of accept state from start state) algorithm

Can we compute “reachability” for a CFG?

Thm: E_{CFG} is a decidable language.

$$E_{CFG} = \{ \langle G \rangle \mid G \text{ is a CFG and } L(G) = \emptyset \}$$

Proof: create **decider** that calculates reachability for grammar G

- Go backwards, start from **terminals**, to avoid getting stuck in looping rules

$R =$ “On input $\langle G \rangle$, where G is a CFG:

1. Mark all terminal symbols in G .
2. Repeat until no new variables get marked:
3. Mark any variable A where G has a rule $A \rightarrow U_1 U_2 \cdots U_k$ and each symbol U_1, \dots, U_k has already been marked.
4. If the start variable is not marked, *accept*; otherwise, *reject*.”

Loop marks 1 new variable on each iteration or stops: it eventually terminates because there are a finite # of variables

Termination argument?

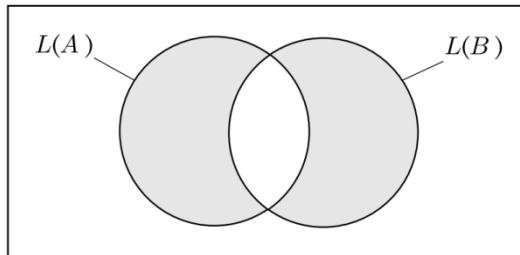
Thm: EQ_{CFG} is a decidable language?



$$EQ_{CFG} = \{\langle G, H \rangle \mid G \text{ and } H \text{ are CFGs and } L(G) = L(H)\}$$

Recall: $EQ_{DFA} = \{\langle A, B \rangle \mid A \text{ and } B \text{ are DFAs and } L(A) = L(B)\}$

- Used Symmetric Difference



$$L(C) = \emptyset \text{ iff } L(A) = L(B)$$

- where C = complement, union, intersection of machines A and B
- Can't do this for CFLs!
 - Intersection and complement are not closed for CFLs!!!

Intersection of CFLs is Not Closed!

Proof (by contradiction), Assume intersection is closed for CFLs

- Then intersection of these CFLs should be a CFL:

$$A = \{a^m b^n c^n \mid m, n \geq 0\}$$

$$B = \{a^n b^n c^m \mid m, n \geq 0\}$$

- But $A \cap B = \{a^n b^n c^n \mid n \geq 0\}$
- ... which is not a CFL! (So we have a contradiction)

Complement of a CFL is not Closed!

- Assume CFLs closed under complement, then:

if G_1 and G_2 context-free

$\overline{L(G_1)}$ and $\overline{L(G_2)}$ context-free From the assumption

$\overline{L(G_1) \cup L(G_1)}$ context-free Union of CFLs is closed


$\overline{\overline{L(G_1) \cup L(G_1)}}$ context-free From the assumption

$L(G_1) \cap L(G_2)$ context-free DeMorgan's Law!

But intersection is not closed for CFLS (prev slide)

Thm: EQ_{CFG} is a decidable language?

$$EQ_{CFG} = \{\langle G, H \rangle \mid G \text{ and } H \text{ are CFGs and } L(G) = L(H)\}$$

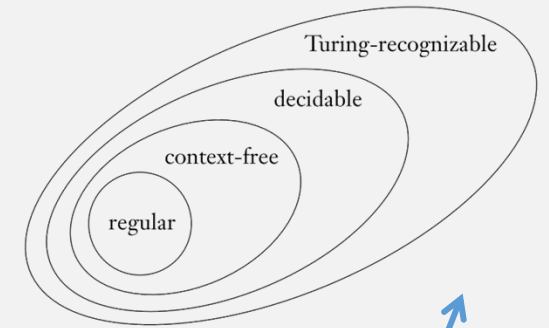
- No! 
 - There's no algorithm to decide whether two grammars are equivalent!
- It's not recognizable either! (Can't create any TM to do this!!!)
 - (details later)
- I.e., this is an impossible computation!

Summary Algorithms About CFLs

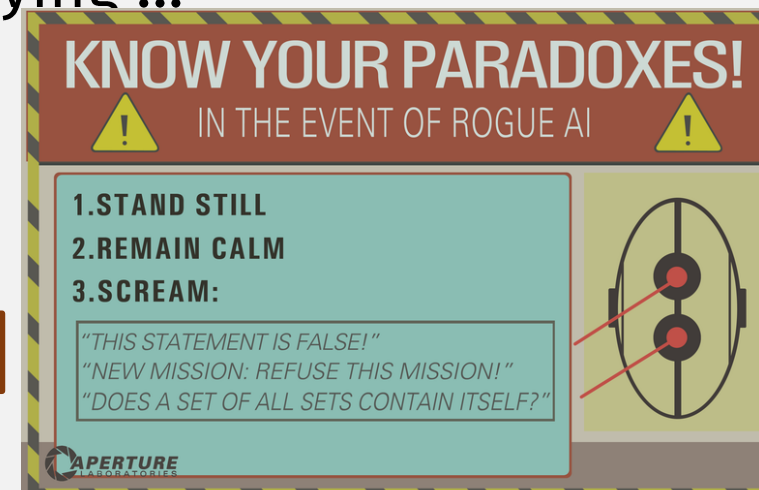
- $A_{CFG} = \{\langle G, w \rangle \mid G \text{ is a CFG that generates string } w\}$
 - **Decider:** Convert grammar to Chomsky Normal Form
 - Then check all possible derivations up to length $2|w| - 1$ steps
- $E_{CFG} = \{\langle G \rangle \mid G \text{ is a CFG and } L(G) = \emptyset\}$
 - **Decider:** Compute “reachability” of start variable from terminals
- $EQ_{CFG} = \{\langle G, H \rangle \mid G \text{ and } H \text{ are CFGs and } L(G) = L(H)\}$
 - We couldn't prove that this is decidable!
 - (So you can't use this theorem when creating another decider)

The Limits of Turing Machines?

- TMs represent all possible “computations”
 - I.e., any (Python, Java, ...) program you write is a TM
- But some things are **not** computable? I.e., some langs are out here ?
- To explore the limits of computation, we have been studying ...
... computation about other computation ...
 - Thought: Is there a decider (algorithm) to determine whether a TM is an decider?

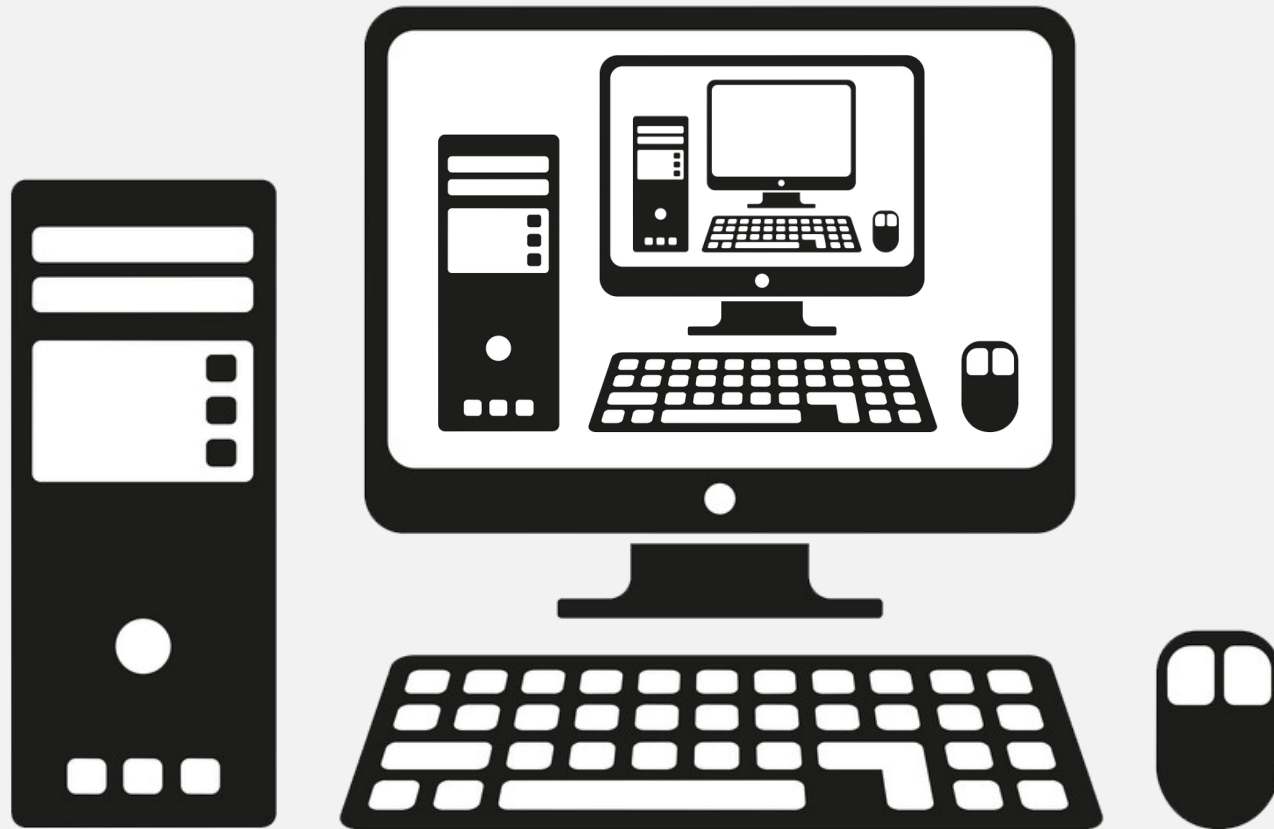


Hmmm, this doesn't feel right ...



Next time: Is A_{TM} decidable?

$$A_{\text{TM}} = \{ \langle M, w \rangle \mid M \text{ is a TM and } M \text{ accepts } w \}$$



Check-in Quiz 4/5

On gradescope