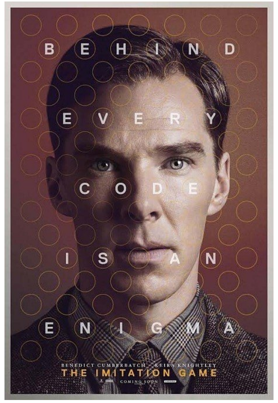
Automata & languages

A primer on the Theory of Computation



Laurent Vanbever

nsg.ee.ethz.ch

ETH Zürich (D-ITET)

October 8 2020

Part 4 out of 5

Last week, we showed the
equivalence of **DFA**, **NFA** and **REX**

is equivalent to

DFA ≍ NFA
)(
REX

We also started to look at
non-regular languages

Pumping lemma

If A is a regular language, then
there exist a number p s.t.

Any string in A whose length is at least p
can be divided into three pieces xyz s.t.

- $xy^iz \in A$, for each $i \geq 0$ and
- $|y| > 0$ and
- $|xy| \leq p$

Wait...

What happens if A is a finite language?!

Pumping lemma

If A is a regular language, then
there exist a number p s.t.

Any string in A whose length is at least p
can be divided into three pieces xyz s.t.

- $xy^iz \in A$, for each $i \geq 0$ and
- $|y| > 0$ and
- $|xy| \leq p$

Pumping lemma

If **A is a regular language**, then
there exist a number p s.t.

As we saw two weeks ago,
all finite languages are regular...

So what's ***p****?*

p := len(longest_string) + 1
makes the lemma hold vacuously

To prove that a language A is not regular:

1 Assume that A is regular

2 Since A is regular, it must have a pumping length p

3 Find one string s in A whose length is at least p

4 For any split $s=xyz$,
Show that you cannot satisfy all three conditions

5 Conclude that s cannot be pumped

To prove that a language A is not regular:

1 Assume that A is regular

2 Since A is regular, it must have a pumping length p

3 Find one string s in A whose length is at least p

4 For any split $s=xyz$,
Show that you cannot satisfy all three conditions

5 Conclude that **s cannot be pumped** $\longrightarrow$ **A is not regular**

Out of the 3 examples we saw last week

the last one is actually regular

L_1 $\{0^n1^n \mid n \geq 0\}$

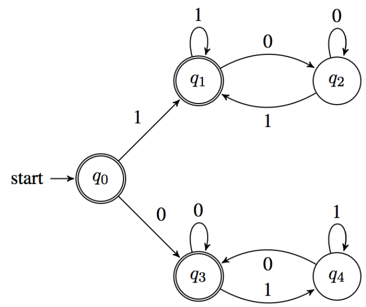
L_2 {w | w has an equal number of 0s and 1s}

L_3 {w | w has an equal number of occurrences of 01 and 10}

how do you show that? You provide a DFA/NFA/REX (you pick)

L_3 {w | w has an equal number of occurrences of 01 and 10}

101 is in L3, not 1010



Key observation

Any binary string beginning and ending with the same digit has an equal number of occurrences of the substrings 01 and 10

This week is all about

Context-Free Languages

a superset of Regular Languages

CFG's: Proving Correctness (Alternative proof)

- The recursive nature of CFG's means that they are especially amenable to correctness proofs.
- For example let's consider again our grammar

  $G = (S \rightarrow \varepsilon \mid ab \mid ba \mid aSb \mid bSa \mid SS)$
- We claim that $L(G) = L = \{ x \in \{a,b\}^* \mid n_a(x) = n_b(x) \}$, where $n_a(x)$ is the number of a's in x, and $n_b(x)$ is the number of b's.
- *Proof*: To prove that L = L(G) is to show both inclusions:
  - *i.* $L \subseteq L(G)$: Every string in L can be generated by G.
  - *ii.* $L \supseteq L(G)$: G only generate strings of L.

CFG's: Proving Correctness

- The recursive nature of CFG's means that they are especially amenable to correctness proofs.
- For example let's consider again our grammar
  $G = (S \to \varepsilon \mid ab \mid ba \mid aSb \mid bSa \mid SS)$
- We claim that L(G) = L = { $x \in \{a,b\}^* \mid n_a(x) = n_b(x)$ },
  where $n_a(x)$ is the number of *a*'s in *x*, and $n_b(x)$ is the number of *b*'s.
- *Proof*: To prove that L = L(G) is to show both inclusions:
  i. $L \subseteq L(G)$: Every string in L can be generated by *G*.
  ii. $L \supseteq L(G)$: *G* only generate strings of L.

Part *ii.* is easy (see why?), so we'll concentrate on part *i*.

Proving $L \subseteq L(G)$

- $L \subseteq L(G)$: Show that every string *x* with the same number of *a*'s as *b*'s is generated by *G*. Prove by induction on the length $n = |x|$.
- Base case: The empty string is derived by $S \to \varepsilon$
- Inductive hypothesis:
  Assume that *G* generates all strings of equal number of *a*'s and *b*'s of (even) length up to *n*.

  Consider any string of length *n+2*. There are essentially 4 possibilities:
  1. *awb*
  2. *bwa*
  3. *awa*
  4. *bwb*

Proving $L \subseteq L(G)$

- Inductive hypothesis:

  Consider any string of length *n*+2. There are essentially 4 possibilities:
  1. *awb*
  2. *bwa*
  3. *awa*
  4. *bwb*

Given $S \Rightarrow^* w$, *awb* and *bwa* are generated from *w* using the rules $S \to aSb$ and $S \to bSa$ (induction hypothesis)

Proving $L \subseteq L(G)$

- Inductive hypothesis:

  Now, consider a string like *awa*. For it to be in *L* requires that *w* isn't in *L* as *w* needs to have 2 more *b*'s than *a*'s.

  – Split *awa* as follows: $_0a_1 \ldots \ _{-1}a_0$
    where the subscripts after a prefix *v* of *awa* denotes $n_a(v) - n_b(v)$

  – Think of this as counting starting from *0*.
    Each *a* adds *1*. Each *b* subtracts *1*. At the end, we should be at *0.*

Somewhere along the string (in *w*), the counter crosses 0 (more b's)

Proving $L \subseteq L(G)$

- Inductive hypothesis:

  Somewhere along the string (in *w*), the counter crosses 0:

  $$\overbrace{{}_0a_1 \ldots \quad {}_{-1}x_0}^{u}\ \underbrace{y \ldots \quad {}_{-1}a_0}_{v} \quad \text{with } x, y \in \{a, b\}$$

  – *u* and *v* have an equal number of *a's* and *b*'s and are shorter than *n*.
  – Given $S \Rightarrow^* u$ and $S \Rightarrow^* v$, the rule $S \rightarrow SS$ generates *awa* = *uv* (induction hypothesis)
  – The same argument applies for strings like *bwb*

Push-Down Automata (PDA)

- Finite automata where the machine interpretation of regular languages.
- Push-Down Automaton are the machine interpretation for grammars.
- The problem of finite automata was that they couldn't handle languages that needed some sort of unbounded memory… something that could be implemented easily by a single (unbounded) integer register!
- Example: To recognize the language L = $\{0^n1^n \mid n \geq 0\}$, all you need is to count how many 0's you have seen so far…
- Push-Down Automata allow even more than a register: a full stack!

Recursive Algorithms and Stacks

- A stack allows the following basic operations
  – Push, pushing a new element on the top of the stack.
  – Pop, removing the top element from the stack (if there is one).
  – Peek, checking the top element without removing it.
- General Principle in Programming:
  Any recursive algorithm can be turned into a non-recursive one using a stack and a while-loop which exits only when stack is empty.

Recursive Algorithms and Stacks

- A stack allows the following basic operations
  – Push, pushing a new element on the top of the stack.
  – Pop, removing the top element from the stack (if there is one).
  – Peek, checking the top element without removing it.
- General Principle in Programming:
  Any recursive algorithm can be turned into a non-recursive one using a stack and a while-loop which exits only when stack is empty.
- It seems that with a stack at our fingertips we can even recognize palindromes! Yoo-hoo!
  – Palindromes are generated by the grammar S → ε | aSa | bSb.
  – Let's simplify for the moment and look at S → # | aSa | bSb.

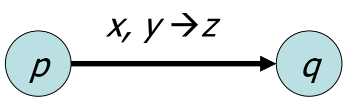
From CFG's to Stack Machines

- The CFG S → # | aSa | bSb describes palindromes containing exactly 1 #.
- Question: Using a stack, how can we recognize such strings?

PDA's à la Sipser

- To aid analysis, theoretical stack machines restrict the allowable operations. Each text-book author has his/her own version.
- Sipser's machines are especially simple:
  - Push/Pop rolled into a single operation: replace top stack symbol.
  - In particular, replacing top by ε is a pop.
- No intrinsic way to test for empty stack.
  - Instead often push a special symbol ("$") as the very first operation!
- Epsilon's used to increase functionality
  - result in default nondeterministic machines.

Sipser's PDA Version



If at state p and next input is x and top stack is y, then go to state q and replace y by z on stack.

- $x = \varepsilon$: ignore input, don't read
- $y = \varepsilon$: ignore top of stack and push z
- $z = \varepsilon$: pop y

In addition, push "$" initially to detect the empty stack.

PDA: Formal Definition

- Definition: A pushdown automaton (PDA) is a 6-tuple $M = (Q, \Sigma, \Gamma, \delta, q_0, F)$:
  - Q, Σ, and q_0, and F are defined as for an FA.
  - Γ is the stack alphabet.
  - δ is as follows:
    Given a state p, an input symbol x and a stack symbol y, $\delta(p,x,y)$ returns all (q,z) where q is a target state and z a stack replacement for y.

$$\delta : Q \times \Sigma_\varepsilon \times \Gamma_\varepsilon \to P(Q \times \Gamma_\varepsilon)$$

PDA Exercises

- Draw the PDA $\{a^i b^j c^k \mid i,j,k \geq 0 \text{ and } i=j \text{ or } i=k\}$

- Draw the PDA for $L = \{x \in \{a,e\}^* \mid n_a(x) = 2n_e(x)\}$

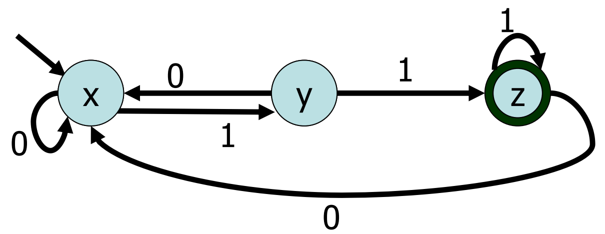
Model Robustness

- The class of regular languages was quite robust
  - Allows multiple ways for defining languages (automaton vs. regexp)
  - Slight perturbations of model do not change result (non-determinism)

- The class of context free languages is also robust: you can use either PDA's or CFG's to describe the languages in the class.

- However, it is less robust than regular languages when it comes to slight perturbations of the model:
  - Smaller classes
    - Right-linear grammars
    - Deterministic PDA's
  - Larger classes
    - Context Sensitive Grammars

Right Linear Grammars vs. Regular Languages



- The DFA above can be simulated by the grammar
  - $x \rightarrow 0x \mid 1y$
  - $y \rightarrow 0x \mid 1z$
  - $z \rightarrow 0x \mid 1z \mid \varepsilon$

- Definition: A right-linear grammar is a CFG such that every production is of the form $A \rightarrow uB$, or $A \rightarrow u$ where u is a terminal string, and A,B are variables.

Right Linear Grammars vs. Regular Languages

- Theorem: If $M = (Q, \Sigma, \delta, q_0, F)$ is an NFA then there is a right-linear grammar $G(M)$ which generates the same language as M.

- *Proof*:
  - Variables are the states: $V = Q$
  - Start symbol is start state: $S = q_0$
  - Same alphabet of terminals Σ
  - A transition $q_1 \rightarrow a \rightarrow q_2$ becomes the production $q_1 \rightarrow aq_2$
  - For each transition, $q_1 \rightarrow aq_2$ where q_2 is an accept state, add $q_1 \rightarrow a$ to the grammar

- Homework: Show that the reverse holds. Right-linear grammar can be converted to a FSA. This implies that RL ≈ Right-linear CFL.

Right Linear Grammars vs. Regular Languages

- Theorem: If $M = (Q, \Sigma, \delta, q_0, F)$ is an NFA then there is a right-linear grammar $G(M)$ which generates the same language as M.

- *Proof*:
  - Variables are the states: $V = Q$
  - Start symbol is start state: $S = q_0$
  - Same alphabet of terminals Σ
  - A transition $q_1 \rightarrow a \rightarrow q_2$ becomes the production $q_1 \rightarrow aq_2$
  - For each transition, $q_1 \rightarrow aq_2$ where q_2 is an accept state, add $q_1 \rightarrow a$ to the grammar

- Homework: Show that the reverse holds. Right-linear grammar can be converted to a FSA. This implies that RL ≈ Right-linear CFL.

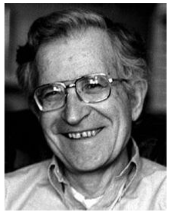
- Question: Can every CFG be converted into a right-linear grammar?

Chomsky Normal Form

- Chomsky came up with an especially simple type of context free grammars which is able to capture all context free languages, the Chomsky normal form (CNF).

- Chomsky's grammatical form is particularly useful when one wants to prove certain facts about context free languages. This is because assuming a much more restrictive kind of grammar can often make it easier to prove that the generated language has whatever property you are interested in.

- Noam Chomsky, linguist at MIT, creator of the Chomsky hierarchy, a classification of formal languages. Chomsky is also widely known for his left-wing political views and his criticism of the foreign policy of U.S. government.



Chomsky Normal Form

- Definition: A CFG is said to be in Chomsky Normal Form if every rule in the grammar has one of the following forms:

  - $S \rightarrow \varepsilon$ (ε for epsilon's sake only)
  - $A \rightarrow BC$ (dyadic variable productions)
  - $A \rightarrow a$ (unit terminal productions)

  where S is the start variable, A,B,C are variables and a is a terminal.

- Thus epsilons may only appear on the right hand side of the start symbol and other rights are either 2 variables or a single terminal.

CFG → CNF

- Converting a general grammar into Chomsky Normal Form works in four steps:

1. Ensure that the start variable doesn't appear on the right hand side of any rule.
2. Remove all epsilon productions, except from start variable.
3. Remove unit variable productions of the form $A \rightarrow B$ where A and B are variables.
4. Add variables and dyadic variable rules to replace any longer non-dyadic or non-variable productions

CFG → CNF: Example

$S \to \varepsilon|a|b|aSa|bSb$

1. No start variable on right hand side

$S' \to S$
$S \to \varepsilon|a|b|aSa|bSb$

2. Only start state is allowed to have ε

$S' \to S|\varepsilon$
$S \to \cancel{\varepsilon}|a|b|aSa|bSb|aa|bb$

3. Remove unit variable productions of the form $A \to B$.

$S' \to \cancel{S}|\varepsilon|a|b|aSa|bSb|aa|bb$
$S \to a|b|aSa|bSb|aa|bb$

CFG → CNF: Example continued

$S' \to \cancel{S}|\varepsilon|a|b|aSa|bSb|aa|bb$
$S \to a|b|aSa|bSb|aa|bb$

4. Add variables and dyadic variable rules to replace any longer productions.

$S' \to \varepsilon|a|b|\cancel{aSa|bSb|aa|bb}|AB|CD|AA|CC$
$S \to a|b|\cancel{aSa|bSb|aa|bb}|AB|CD|AA|CC$
$A \to a$
$B \to SA$
$C \to b$
$D \to SC$

CFG → PDA

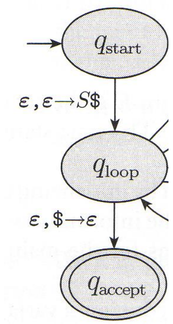
- CFG's can be converted into PDA's.
- In "NFA → REX" it was useful to consider GNFA's as a middle stage. Similarly, it's useful to consider Generalized PDA's here.
- A Generalized PDA (GPDA) is like a PDA, except it allows the top stack symbol to be replaced by a whole string, not just a single character or the empty string. It is easy to convert a GPDA's back to PDA's by changing each compound push into a sequence of simple pushes.

CFG → GPDA Recipe

1. Push the marker symbol \$ and the start symbol S on the stack.
2. Repeat the following steps forever
   a. If the top of the stack is the variable symbol *A*, nondeterministically select a rule of *A*, and substitute *A* by the string on the right-hand-side of the rule.
   b. If the top of the stack is a terminal symbol *a*, then read the next symbol from the input and compare it to *a*. If they match, continue. If they do not match reject this branch of the execution.
   c. If the top of the stack is the symbol *\$*, enter the accept state. (Note that if the input was not yet empty, the PDA will still reject this branch of the execution.)

CFG → GPDA → PDA: Example

- S → aTb | b
- T → Ta | ε



CFG → PDA: Now you try!

- Convert the grammar *S* → ε |*a* | *b* | *aSa* | *bSb*

PDA → CFG

- To convert PDA's to CFG's we'll need to simulate the stack inside the productions.
- Unfortunately, in contrast to our previous transitions, this is not quite as constructive. We will therefore only state the theorem.
- Theorem: For each push-down automation there is a context-free grammar which accepts the same language.
- Corollary: PDA ≈ CFG.

Context Sensitive Grammars

- An even more general form of grammars exists.
  In general, a non-context free grammar is one in which whole mixed variable/terminal substrings are replaced at a time.
  For example with Σ = {*a*,*b*,*c*} consider:

  $S \to \varepsilon \mid ASBC$ $\quad aB \to ab$
  $A \to a$ $\quad bB \to bb$
  $CB \to BC$ $\quad bC \to bc$
  $\quad cC \to cc$

  What language is generated by this non-context-free grammar?

- When length of LHS always ≤ length of RHS (plus some other minor restrictions), these general grammars are called context sensitive.

Are all languages context-free?

- Design a CFG (or PDA) for the following languages:
- L = { w ∈ {0,1,2}* | there are k 0's, k 1's, and k 2's for $k \geq 0$ }
- L = { w ∈ {0,1,2}* | with |0| = |1| or |0| = |2| or |1| = |2| }
- L = { $0^k 1^k 2^k$ | $k \geq 0$ }

Tandem Pumping

- Analogous to regular languages there is a pumping lemma for context free languages. The idea is that you can pump a context free language at two places (but not more).
- Theorem: Given a context free language L, there is a number p (tandem-pumping number) such that any string in L of length $\geq p$ is tandem-pumpable within a substring of length p. In particular, for all $w \in L$ with $|w| \geq p$ we we can write:
  - $w = uvxyz$
  - $|vy| \geq 1$ (pumpable areas are non-empty)
  - $|vxy| \leq p$ (pumping inside length-p portion)
  - $uv^ixy^iz \in L$ for all $i \geq 0$ (tandem-pump v and y)
- If there is no such p the language is not context-free.

Proving Non-Context Freeness: Example

- L ={$\mathbf{1^n 0^n 1^n 0^n}$ | n is non-negative }
- Let's try w = $1^p0^p1^p0^p$. Clearly w ∈ L and |w| ≥ p.
- With |vxy| ≤ p, there are only three places where the "sliding window" *vxy* could be:

  I III
  1...10...01...10...0
  II

- In all three cases, pumping up such a case would only change the number of 0s and 1s in that part and not in the other two parts; this violates the language definition.

Proving Non-Context Freeness: You try!

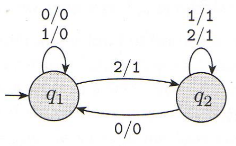
- L = { x=y+z | x, y, and z are binary bit-strings satisfying the equation }
- The hard part is to come up with a word which cannot be pumped, such as…

Transducers

- Definition: A finite state transducer (FST) is a type of finite automaton whose output is a string and not just accept or reject.

- Each transition of an FST is labeled with two symbols, one designating the input symbol for that transition (as for automata), and the other designating the output symbol.
  - We allow ε as output symbol if no symbol should be added to the string.

- The figure on the right shows an example of a FST operating on the input alphabet {0,1,2} and the output alphabet {0,1}



- Exercise: Can you design a transducer that produces the inverted bit-string of the input string (e.g. 01001 → 10110)?