Last week, we showed the equivalence of DFA, NFA and REX is equivalent to

\[
\text{DFA } \approx \text{ NFA } \approx \text{ REX}
\]
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.

- $xyz \in A$, 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.

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

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

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

So what's $p$?

$p := \text{len(longest\_string)} + 1$

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 $\longrightarrow$ $A$ is not regular.

Out of the 3 examples we saw last week, the last one is actually regular.

$L_1$ \{0^n1^n | n ≥ 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/REG (you pick).

Non-regular languages are not closed under most operations:

if $L_1$ and $L_2$ are regular, then so are

$L_1 \cup L_2$
$L_1 \cdot L_2$
$L_1^*$

if $L_1$ and $L_2$ are not regular, then

$L_1 \cup L_2$
$L_1 \cdot L_2$
$L_1^*$ may or may not be regular!

$(L_1)^R$ is not regular.

Non RL are closed under complement.
This week is all about

Context-Free Languages

a superset of Regular Languages

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.*

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 $L = \{0^n1^n | n \geq 0\}$, 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.*

- It seems that with a stack at our fingertips we can even recognize **palindromes**! Yoo-hoo!
  - Palindromes are generated by the grammar $S \rightarrow \varepsilon | aSa | bSb$.
  - Let’s simplify for the moment and look at $S \rightarrow \# | aSa | bSb$. 
From CFG’s to Stack Machines

- The CFG $S \rightarrow # | aSa | bSb$ 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 $\varepsilon$ 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 $M = (Q, \Sigma, \Gamma, \delta, q_0, F)$:
  - $Q, \Sigma, \Gamma, q_0$, and $F$ are defined as for an FA.
  - $\Gamma$ is the stack alphabet.
  - $\delta$ 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 $y$.

$$\delta : Q \times \Sigma \times \Gamma \rightarrow P(Q \times \Gamma)$$
PDA Exercises

• Draw the PDA \{a^ib^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
  – \( x \rightarrow 0x \mid 1y \)
  – \( y \rightarrow 0x \mid 1z \)
  – \( z \rightarrow 0x \mid 1z \mid \epsilon \)

• 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.

• 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 $\Sigma$
  - A transition $q_i \rightarrow a \rightarrow q_j$ becomes the production $q_i \rightarrow aq_j$
  - For each transition, $q_i \rightarrow aq_j$ where $q_j$ is an accept state, add $q_i \rightarrow a$ to the grammar

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

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

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:
  - $S \rightarrow \epsilon$ (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 $\rightarrow$ 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 $\rightarrow$ CNF: Example

\[ S \rightarrow \varepsilon | a | b | aS | bSb \]

1. No start variable on right hand side

\[ S' \rightarrow S \]
\[ S \rightarrow \varepsilon | a | b | aS | bSb \]

2. Only start state is allowed to have $\varepsilon$

\[ S' \rightarrow S | \varepsilon \]
\[ S \rightarrow \varepsilon | a | b | aS | bSb | aa | bb \]

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

\[ S' \rightarrow \varepsilon | a | b | aS | bSb | aa | bb \]
\[ S \rightarrow a | b | aS | bSb | aa | bb \]

CFG $\rightarrow$ CNF: Example continued

\[ S' \rightarrow \varepsilon | a | b | aS | bSb | aa | bb \]
\[ S \rightarrow a | b | aS | bSb | aa | bb \]

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

\[ S' \rightarrow \varepsilon | a | b | aS | bSb | aa | bb | AB | CD | AA | CC \]
\[ S \rightarrow a | b | aS | bSb | aa | bb | AB | CD | AA | CC \]
\[ A \rightarrow a \]
\[ B \rightarrow SA \]
\[ C \rightarrow b \]
\[ D \rightarrow SC \]

CFG $\rightarrow$ PDA

- CFG's can be converted into PDA's.

- In "NFA $\rightarrow$ RE" 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 $\rightarrow$ 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 $\rightarrow$ GPDA $\rightarrow$ PDA: Example

- $S \rightarrow aTb \mid b$
- $T \rightarrow Ta \mid \varepsilon$

PDA $\rightarrow$ 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: $\text{PDA} \approx \text{CFG}$.

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 $\Sigma = \{a,b,c\}$ consider:
    - $S \rightarrow \varepsilon \mid \text{ASBC}$
    - $A \rightarrow a$
    - $bB \rightarrow bb$
    - $CB \rightarrow BC$
    - $bC \rightarrow bc$
    - $cC \rightarrow cc$

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

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

CFG $\rightarrow$ PDA: Now you try!

- Convert the grammar $S \rightarrow \varepsilon \mid a \mid bSa \mid bSb$
Are all languages context-free?

- Design a CFG (or PDA) for the following languages:
  - $L = \{ w \in \{0,1,2\}^* \mid \text{there are } k \text{ 0's, } k \text{ 1's, and } k \text{ 2's for } k \geq 0 \}$
  - $L = \{ w \in \{0,1,2\}^* \mid \text{with } |0| = |1| \text{ or } |0| = |2| \text{ or } |1| = |2| \}$
  - $L = \{ 0^k1^k2^k \mid k \geq 0 \}$

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 can write:
    - $w = uvxyz$
    - $|vy| \geq 1$ (pumpable areas are non-empty)
    - $|vxy| \leq p$ (pumping inside length-$p$ portion)
    - $uv^ixyz \in L$ 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

- $L = \{1^n0^n1^n0^n \mid n \text{ is non-negative} \}$
- Let’s try $w = 1^p0^p1^p0^p$. Clearly $w \in L$ and $|w| \geq p$.
- With $|vxy| \leq p$, there are only three places where the “sliding window” $vxy$ could be:

  - I 1...10...10...0
  - II
  - III

- 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!

- $L = \{ x=y+z \mid x, y, \text{ and } z \text{ are binary bit-strings satisfying the equation} \}$
- 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 $\varepsilon$ 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. $01001 \rightarrow 10110$)?