Harvard CS 121 and CSCI E-207
Lecture 7: Non-Regular Languages

Harry Lewis

September 23, 2010

- **Reading:** Sipser, §1.3.

Examples of Regular Languages

- $\{w \in \{a, b\}^* : |w|$ even & every 3rd symbol is an $a\}$
- $\{w \in \{a, b\}^* :$ There are not 7 a's or 7 b's in a row$\}$
- $\{w \in \{a, b\}^* : w$ has both an even number of a's and an even number of b's$\}$
- $\{w : w$ is written using using the ASCII character set and every substring delimited by spaces, punctuation marks, or the beginning or end of the string is in the American Heritage Dictionary$\}$

Questions about regular languages

Given X = a regular expression, DFA, or NFA, how could you tell if:

- $x \in L(X)$, where x is some string?
- $L(X) = \emptyset$?
- $x \in L(X)$ but $x \notin L(Y)$?
- $L(X) = L(Y)$, where Y is another RE/FA?
- $L(X)$ is infinite?
- There are infinitely many strings that belong to both $L(X)$ and $L(Y)$?

Goal: Existence of Non-Regular Languages

Intuition:

- Every regular language can be described by a finite string (namely a regular expression).
- To specify an arbitrary language requires an infinite amount of information.
  - For example, an infinite sequence of bits would suffice:
  - Σ^* has a lexicographic ordering, and the i'th bit of an infinite sequence specifying a language would say whether or not the i'th string is in the language.

$\Rightarrow$ Some language must not be regular.

How to formalize?

Countability

- A set S is finite if there is a bijection $\{1, \ldots, n\} \leftrightarrow S$ for some $n \geq 0$.

- Countably infinite if there is a bijection $f : \mathcal{N} \leftrightarrow S$

  This means that S can be "enumerated," i.e. listed as $\{s_0, s_1, s_2, \ldots\}$ where $s_i = f(i)$ for $i = 0, 1, 2, 3, \ldots$

  So $\mathcal{N}$ itself is countably infinite

  So is $\mathcal{Z}$ (integers) since $\mathcal{Z} = \{0, -1, 1, -2, 2, \ldots\}$

  Q: What is f?

- Countable if S is finite or countably infinite

- Uncountable if it is not countable

More Countable Sets

- $\mathcal{N} \times \mathcal{N}$ (why?)
- The set of rational numbers (why?)
- Σ^* for any alphabet Σ (why?)
- The set of all regular expressions (why?)
- The set of all finite automata over alphabet Σ (why?)

Facts about Infinite Sets

- **Proposition:** The union of 2 countably infinite sets is countably infinite.

  If $A = \{a_0, a_1, \ldots\}$, $B = \{b_0, b_1, \ldots\}$

  Then $A \cup B = C = \{c_0, c_1, \ldots\}$

  where $c_i = \begin{cases} a_{i/2} & \text{if } i \text{ is even} \\ b_{(i-1)/2} & \text{if } i \text{ is odd} \end{cases}$

  Q: If we are being fussy, there is a small problem with this argument. What is it?

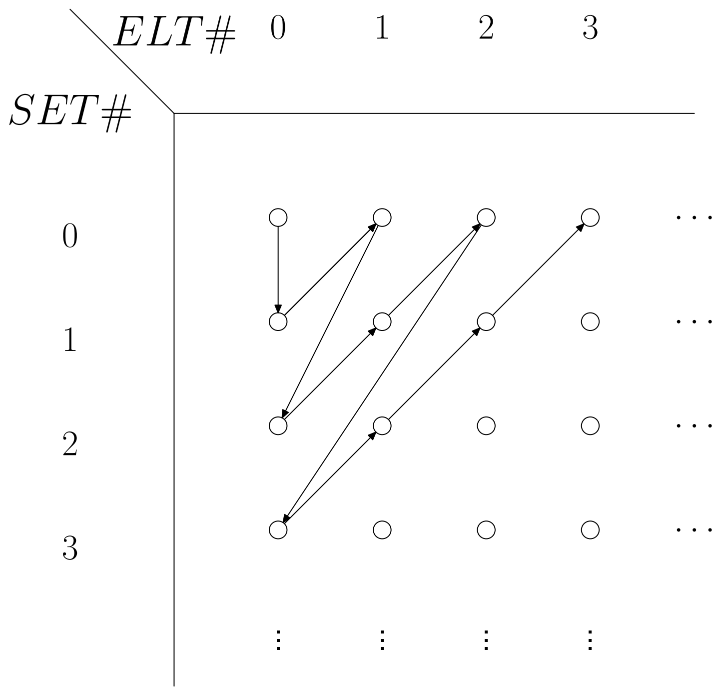
- **Proposition:** If there is a function $f : \mathcal{N} \to S$ that is onto S then S is countable.

Countable Unions of Countable Sets

- **Proposition:** The union of countably many countably infinite sets is countably infinite

Countable Unions of Countable Sets

- **Proposition:** The union of countably many countably infinite sets is countably infinite



$$(0,0),(0,1),(1,0),(0,2),(1,1),(2,0),\ldots$$

Each element is "reached" eventually in this ordering

- What assumption is implicit in this argument?

Are there uncountable sets? (Infinite but not countably infinite)

Theorem: $P(\mathcal{N})$ is uncountable

(The set of all sets of natural numbers)

Proof by contradiction:

(i.e. assume that $P(\mathcal{N})$ is countable and show that this results in a contradiction)

- Suppose that $P(\mathcal{N})$ were countable.
- Then there is an enumeration of all subsets of $\mathcal{N}$ say $P(\mathcal{N}) = \{S_0, S_1, \ldots\}$

Diagonalization

S_i \ $j =$	0	1	2	3	4	
S_0	Y	N	N	Y	N	$\cdots$
S_1	N	N	N	N	N	$\cdots$
S_2	Y	Y	N	Y	Y	$\cdots$
S_3	N	N	N	Y	N	$\cdots$
$\vdots$				D		

"Y" in row i, column j means $j \in S_i$

- Let $D = \{i \in \mathcal{N} : i \in S_i\}$ be the diagonal.
- $D = YNNY \ldots = \{0, 3, \ldots\}$
- Let $\overline{D} = \mathcal{N} - D$ be its complement.
- $\overline{D} = NYYN \ldots = \{1, 2, \ldots\}$
- **Claim:** $\overline{D}$ is omitted from the enumeration, contradicting the assumption that every set of natural numbers is one of the S_is.

  Pf: $\overline{D}$ is different from each row because they differ at the diagonal.

Cardinality of Languages

- An alphabet Σ is finite by definition

- **Proposition:** Σ^* is countably infinite

- So every language is either finite or countably infinite

- $P(\Sigma^*)$ is uncountable, being the set of subsets of a countable infinite set.

  i.e. There are uncountably many languages over any alphabet

  Q: Even if $|\Sigma| = 1$?

Existence of Non-regular Languages

Theorem: For every alphabet Σ, there exists a non-regular language over Σ.

Proof:

- There are only countably many regular expressions over Σ.

  $\Rightarrow$ There are only countably many regular languages over Σ.

- There are uncountably many languages over Σ.
- Thus at least one language must be non-regular.

$\Rightarrow$ In fact, "almost all" languages must be non-regular.

Q: Could we do this proof using DFAs instead?

Q: Can we get our hands on an *explicit* non-regular language?

Cardinality of Languages

- An alphabet Σ is finite by definition

- **Proposition:** Σ^* is countably infinite

- So every language is either finite or countably infinite

- $P(\Sigma^*)$ is uncountable, being the set of subsets of a countably infinite set.

  i.e. There are uncountably many languages over any alphabet

  Q: Even if $|\Sigma| = 1$?

Existence of Non-regular Languages

Theorem: For every alphabet Σ, there exists a non-regular language over Σ.

Proof:

- There are only countably many regular expressions over Σ.

  $\Rightarrow$ There are only countably many regular languages over Σ.
- There are uncountably many languages over Σ.
- Thus at least one language must be non-regular.

$\Rightarrow$ In fact, "almost all" languages must be non-regular.

Q: Could we do this proof using DFAs instead?

Q: Can we get our hands on an *explicit* non-regular language?

Goal: Explicit Non-Regular Languages

It **appears** that a language such as

$$\begin{aligned} L &= \{x \in \Sigma^* : |x| = 2^n \text{ for some } n \geq 0\} \\ &= \{a, b, aa, ab, ba, bb, aaaa, \ldots, bbbb, aaaaaaaa, \ldots\} \end{aligned}$$

can't be regular because the "gaps" in the set of possible lengths become arbitrarily large, and no DFA could keep track of them.

But this isn't a proof!

Approach:

1. Prove some general property P of all regular languages.
2. Show that L does not have P.

Pumping Lemma (Basic Version)

If L is regular, then there is a number p (the <u>pumping length</u>) such that
every string $s \in L$ of length at least p
can be divided into $s = xyz$, where $y \neq \varepsilon$ and
for every $n \geq 0$, $xy^n z \in L$.

$n = 1$	x	y	z	
$n = 0$	x	z		
$n = 2$	x	y	y	z
$\ldots$				

- Why is the part about p needed?

- Why is the part about $y \neq \varepsilon$ needed?

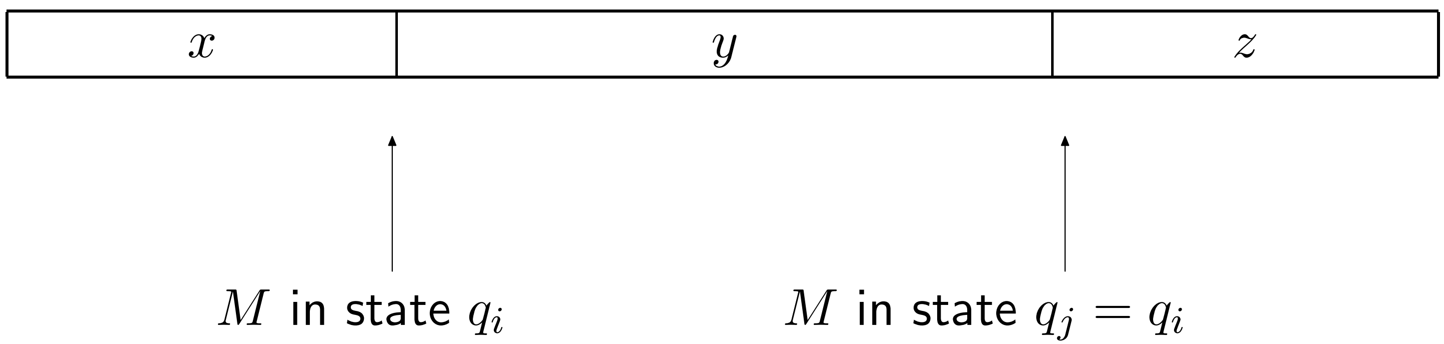
Proof of Pumping Lemma

(Another fooling argument)

- Since L is regular, there is a DFA M recognizing L.
- Let $p =$ # states in M.
- Suppose $s \in L$ has length $l \geq p$.
- M passes through a sequence of $l + 1 > p$ states while accepting s (including the first and last states): say, $q_0, \ldots, q_l$.
- Two of these states must be the same: say, $q_i = q_j$ where $i < j$

Pumping, continued

- Thus, we can break s into x, y, z where $y \neq \varepsilon$ (though x, z may equal ε):



- If more copies of y are inserted, M “can’t tell the difference,” i.e., the state entering y is the same as the state leaving it.

- So since $xyz \in L$, then $xy^n z \in L$ for all n.

Proof also shows (why?):

- We can take p = # states in smallest DFA recognizing L.

- Can guarantee division $s = xyz$ satisfies $|xy| \leq p$ (or $|yz| \leq p$).

Pumping Lemma Example

- Consider

$$L = \{x : x \text{ has an even \# of } a\text{'s and an odd \# of } b\text{'s}\}$$

- Since L is regular, pumping lemma holds.

  (i.e, every sufficiently long string s in L is "pumpable")

- For example, if $s = aab$, we can write $x = \varepsilon$, $y = aa$, and $z = b$.

Pumping the even a's, odd b's language

- Claim: L satisfies pumping lemma with pumping length $p = 4$.
- Proof:
- **Q:** Can the Pumping Lemma be used to prove that L is regular?

Use PL to Show Languages are NOT Regular

Claim: $L = \{a^n b^n : n \geq 0\} = \{\varepsilon, ab, aabb, aaabbb, \ldots\}$ is not regular.

Proof by contradiction:

- Suppose that L is regular.
- So L has some pumping length $p > 0$.
- Consider the string $s = a^p b^p$. Since $|s| = 2p > p$, we can write $s = xyz$ for some strings x, y, z as specified by lemma.
- Claim: No matter how s is partitioned into xyz with $y \neq \varepsilon$, we have $xy^2z \notin L$.
- This violates the conclusion of the pumping lemma, so our assumption that L is regular must have been false.

Strings of exponential lengths are a nonregular language

Claim: $L = \{w : |w| = 2^n$ for some $n \geq 0\}$ is not regular.

Proof:

“Regular Languages Can’t Do Unbounded Counting”

Claim: $L = \{w : w$ has the same number of a’s and b’s$\}$ is not regular.

Proof #1:

- Use pumping lemma on $s = a^p b^p$ with $|xy| \leq p$ condition.

“Regular Languages Can’t Do Unbounded Counting”

Claim: $L = \{w : w$ has the same number of a’s and b’s$\}$ is not regular.

Proof #1:

- Use pumping lemma on $s = a^p b^p$ with $|xy| \leq p$ condition.

Proof #2:

- If L were regular, then $L \cap a^* b^*$ would also be regular.

Reprise on Regular Languages

Which of the following are necessarily regular?

- A finite language
- A union of a finite number of regular languages
- $\{x : x \in L_1$ and $x \notin L_2\}$, L_1 and L_2 are both regular
- A cofinite language (a set is *cofinite* if its complement is finite)
- The reversal of a regular language
- A subset of a regular language