BU CS 332 – Theory of Computation

https://forms.gle/QHsstTW4vZS2CgDV7

Lecture 7:

- Distinguishing sets
- Non-regular languages

Reading:

"Myhill-Nerode" note

Alexander Poremba & Mark Bun

February 10, 2026

Last Time

- Regular expressions characterize the regular languages
  - Every regex can be converted to an NFA recognizing its language
  - Every NFA can be converted to a regex generating its language

NFA → Regex

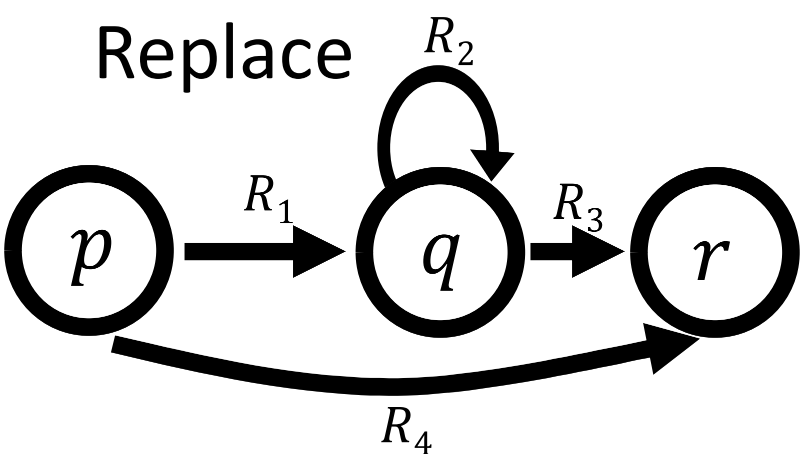
Theorem 2: Every NFA has an equivalent regular expression

Given NFA N:

1. Make N into a GNFA by adding new start and accept state
2. For each non-start and non-accept state q:

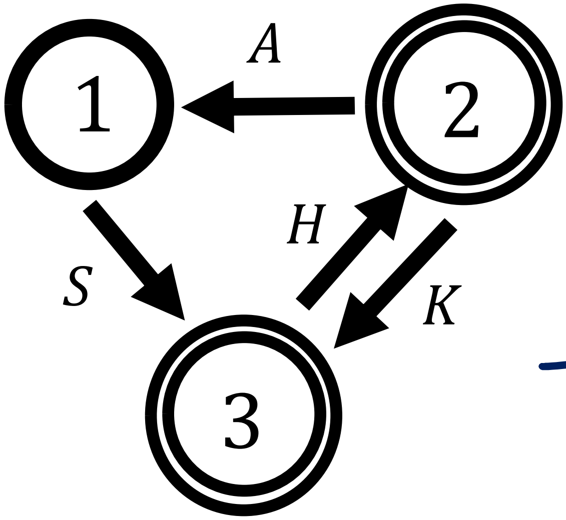
   For each pair of states p, r: Possible for $p = r$

   Replace

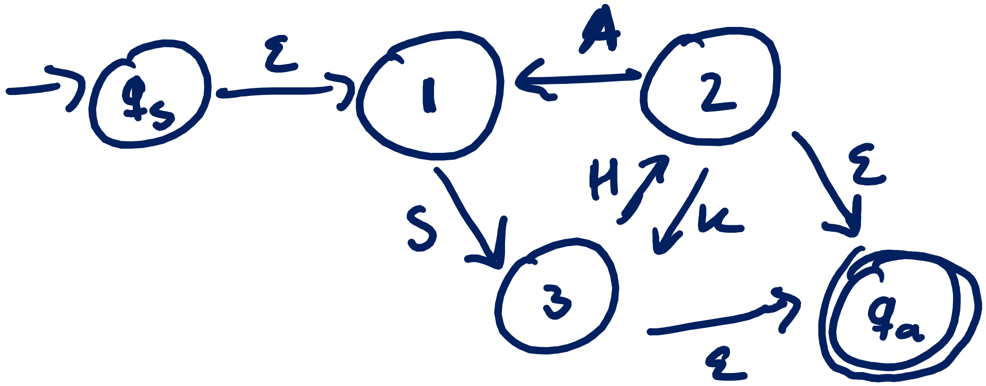


   with

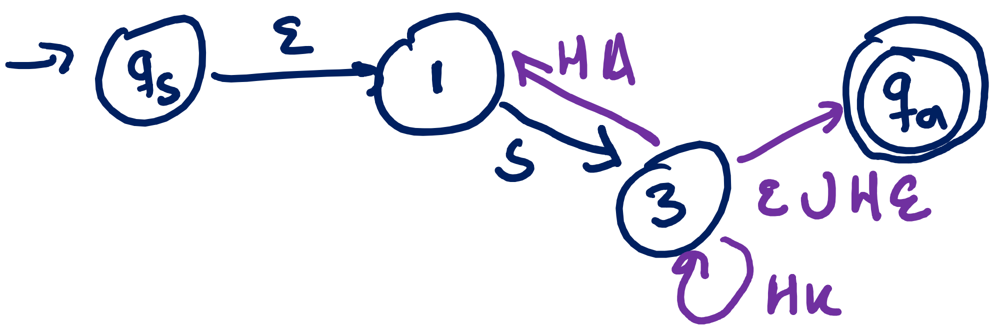
3. Output regex labeling transition from start to accept state



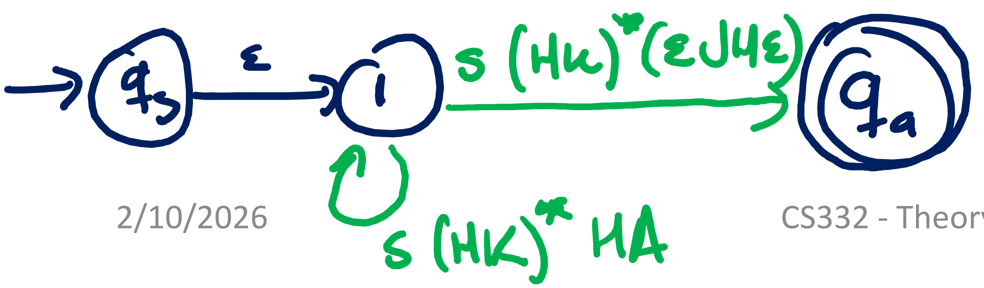
1. Convert to GNFA



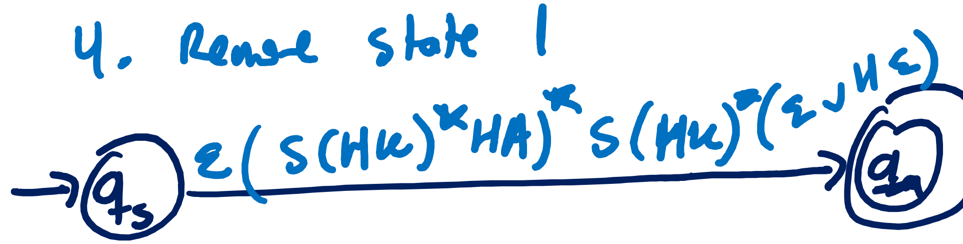
2. Remove state 2



3. Remove state 3



4. Remove state 1



Final regex:

$$\varepsilon\,(S(HK)^*HA)^*\,S(HK)^*(\varepsilon \cup H\varepsilon)$$

$$= (S(HK)^*HA)^*\,S(HK)^*(\varepsilon \cup H)$$

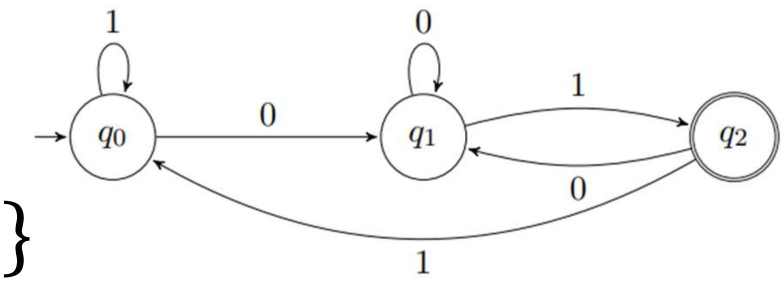
Limits of Finite Automata

Motivating Questions

- We've seen several techniques for showing languages are regular
  - Construct DFA
  - Construct NFA
  - Construct regex
  - Closure properties
- How can we tell if we've found the smallest DFA recognizing a given language?
- Which languages are not regular? How can we prove so?

An Example

$A = \{ w \in \{0,1\}^* \mid w \text{ ends with } 01 \}$



Claim: *Every* DFA recognizing A needs at least 3 states

Proof: Let M be any DFA recognizing A. Consider running M on each of $x = \varepsilon, y = 0, w = 01$

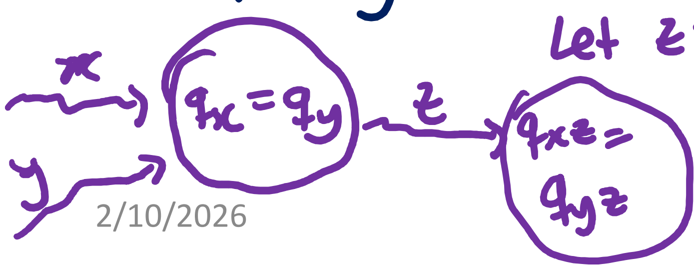
Let q_x = state M reaches on input $x = \varepsilon$
q_y = " " $y = 0$
q_w = " " $w = 01$

Claim: q_x, q_y, q_w are all distinct states

Claim 1: $q_w \neq q_x$ and $q_w \neq q_y$ because q_w is an accept state but q_x, q_y are reject states

Claim 2: $q_x \neq q_y$. Assume for contradiction that $q_x = q_y$

Let $z = 1$. Let q_{xz} = state M reaches on input $xz = \varepsilon 1 = 1 \notin A$
q_{yz} = " " $yz = 01 \in A$



$\Rightarrow q_{xz}$ is a reject state but q_{yz} is an accept state

But $q_x = q_y \Rightarrow q_{xz} = q_{yz}$ is a contradiction ⚡

$A = \{ w \in \{0,1\}^* \mid w \text{ ends with } 01 \}$

A General Technique

Definition: Strings x and y are **distinguishable** by L if there exists a "distinguishing extension" $z \in \Sigma^*$ such that exactly one of xz or yz is in L.

Ex. $x = \varepsilon, \quad y = 0$ are distinguishable by A because

z = 1 is a distinguishing extension:

$xz = \varepsilon 1 = 1 \notin A$

$yz = 01 \in A$

Definition: A set of strings S is **pairwise distinguishable** by L if every pair of distinct strings $x, y \in S$ is distinguishable by L.

Ex. $S = \{\varepsilon, 0, 01\}$

$x = \varepsilon \quad y = 0 \quad z = 1$

$x = \varepsilon \quad y = 01 \quad z = \varepsilon \Rightarrow xz = \varepsilon\varepsilon = \varepsilon \notin A$, $yz = 01 \in A$

$x = 0 \quad y = 01 \quad z = \varepsilon \Rightarrow xz = 0 \notin A$, $yz = 01 \in A$

A General Technique

Theorem: If S is pairwise distinguishable by L, then every DFA recognizing L must have at least $|S|$ states

Proof: Let M be a DFA with $< |S|$ states. WTS: M does not recognize L

<u>Claim:</u> There are distinct strings $x, y \in S$ such that M ends up in same state on x and y Why? Pigeonhole principle

Pigeons: strings in S Assign pigeon x to hole q_x = state M reaches when run on input x

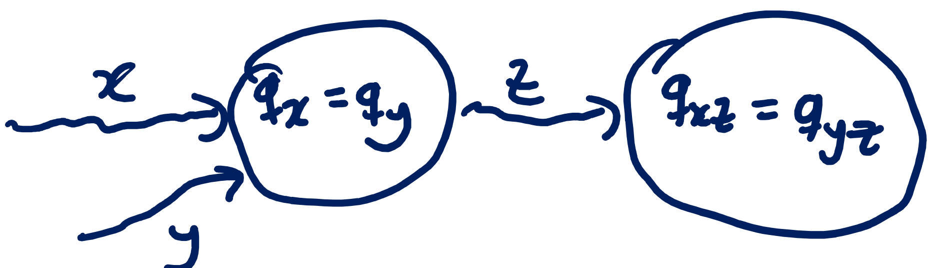
Holes: states of M

Let $x, y \in S$ be strings guaranteed by claim

Since S is pairwise distinguishable by L, $\exists$ a distinguishing extension z

i.e. exactly one of $xz \in L$ or $yz \in L$

WTS: M gives the wrong answer on either xz or yz



Let $q_x = q_y$ = state M reaches on x (or on y)

$\Rightarrow q_{xz} = q_{yz}$

Can't be both an accept and reject state

so M messes up on either xz or yz

Another Example

Claim: Every DFA recognizing B must have ≥ 4 states

$B = \{ w \in \{0,1\}^* \mid |w| = 2 \}$

Theorem: If S is pairwise distinguishable by L, then every DFA recognizing L must have at least $|S|$ states

$S = \{ \varepsilon, 0, 00, 000 \}$

Claim: S is pairwise distinguishable

Proof:

$x = \varepsilon \quad y = 0 \quad z = 0$

Then $xz = \varepsilon 0 = 0 \notin B$
$yz = 00 \in B$

$x = \varepsilon \quad y = 00 \quad z = \varepsilon \quad xz = \varepsilon \notin B \quad yz = 00 \in B$

$x = \varepsilon \quad y = 000 \quad z = 00 \quad xz = 00 \in B \quad yz = 00000 \notin B$

$x = 0 \quad y = 00 \quad z = \varepsilon \quad xz = 0 \notin B \quad yz = 00 \in B$

$x = 0 \quad y = 000 \quad z = 0 \quad xz = 00 \in B \quad yz = 0000 \notin B$

Intuition: A DFA recognizing B must:

On input ε: be prepared to accept iff it sees 2 more characters

On input 0: " " 1 more character

00: Accept

000: Reject

$x = 00 \quad y = 000 \quad z = \varepsilon$
$xz = 00 \in B$
$yz = 000 \notin B$

Distinguishing Extension

Which of the following is a distinguishing extension for $x = 0$ and $y = 00$ for language $B = \{ w \in \{0,1\}^* \mid |w| = 2 \}$?

✓ a) $z = \varepsilon$

$xz = 0\varepsilon = 0 \notin B$
$yz = 00\varepsilon = 00 \in B$

✓ b) $z = 0$

$xz = 00 \in B$
$yz = 000 \notin B$

✓ c) $z = 1$

$xz = 01 \in B$
$yz = 001 \notin B$

✗ d) $z = 00$

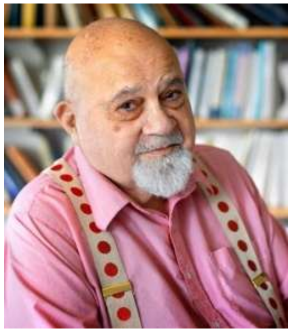
$xz = 0000 \notin B$
$yz = 0000 \notin B$

Historical Note

Converse to the distinguishing set method:

If L has **no** distinguishing set of size $> k$, then L is recognized by a DFA with k states

Myhill-Nerode Theorem (1958): L is recognized by a DFA with $\leq k$ states **if and only if** L does not have a distinguishing set of size $> k$



Non-Regularity

Theorem: If S is pairwise distinguishable by L, then every DFA recognizing L must have at least $|S|$ states

Contrapositive form: $\exists$ a DFA recognizing L w/ $< k$ states
$\Rightarrow$ there is no pairwise distinguishable set for L with $\geq k$ states

Corollary: If S is an **infinite** set that is pairwise distinguishable by L, then no DFA recognizes L

Contrapositive form: $\exists$ a DFA recognizing L
$\Rightarrow$ L does not have an infinite pairwise distinguishable set

Proof that contrapositive of Thm $\Rightarrow$ contrapositive of Corollary:

Let L be recognized by some DFA. Let $k = \#$ of states of this DFA.

By (contrapositive of Thm) there is no pairwise dist set for L of size $\geq k+1$)

$\Rightarrow$ L does not have an infinite pairwise dist set.

The Classic Example

$0^* 1^*$

$01 \in A$ $\quad 001 \notin A$

Theorem: $A = \{0^n 1^n \mid n \geq 0\}$ is not regular

Proof: We construct an infinite pairwise distinguishable set

$S = \{\varepsilon, 0, 00, 000, \ldots\} = \{0^n \mid n \geq 0\}$

Proof that S is pairwise distinguishable:

Let $x, y \in S$ be arbitrary distinct strings

Then we can write $x = 0^n$, $y = 0^m$ for $m \neq n$ (≥ 0)

claim: $z = 1^n$ is a distinguishing extension

$xz = 0^n 1^n \in A$

$yz = 0^m 1^n \notin A$

Since S is an infinite pairwise dist. set for A, A is not regular