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

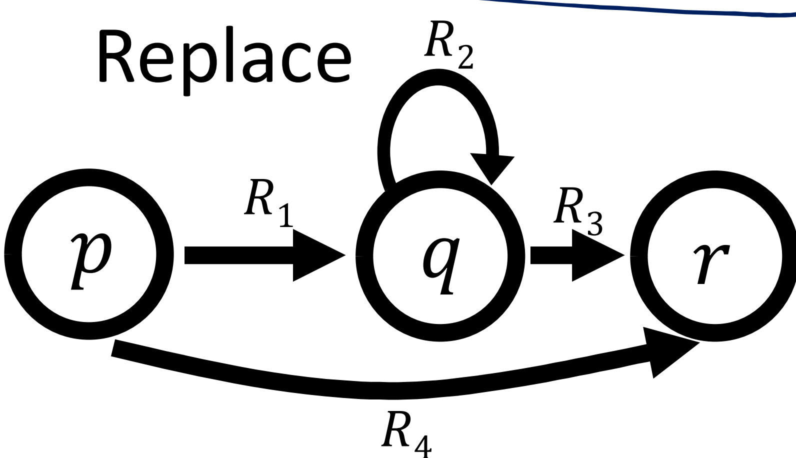
Theorem 2: Every NFA has an equivalent regular expression

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:

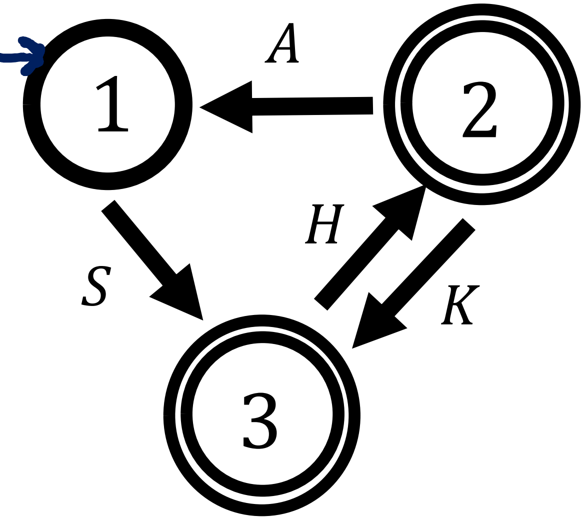
   For each pair of states p, r: p, r might be the same state

   Replace

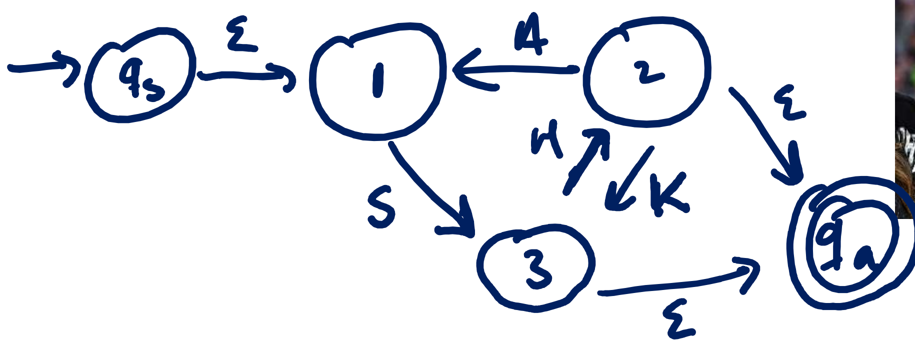


with

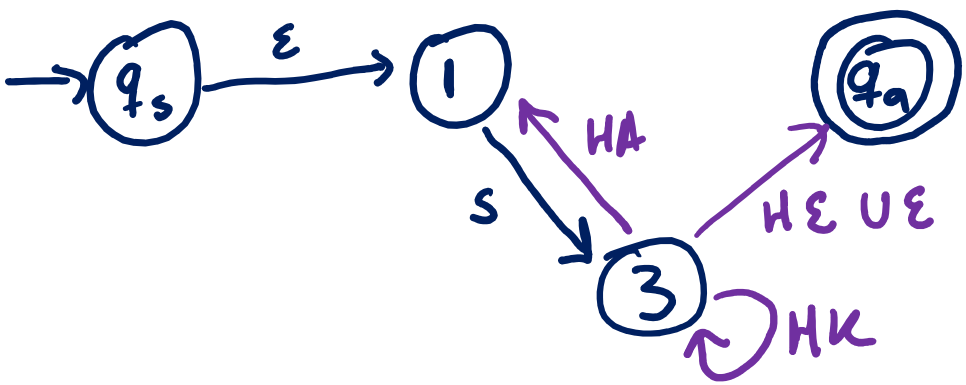
3. Output regex labeling transition from start to accept state



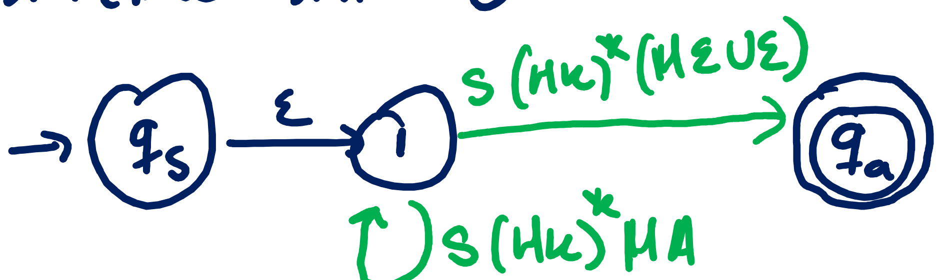
1. Convert to GNFA



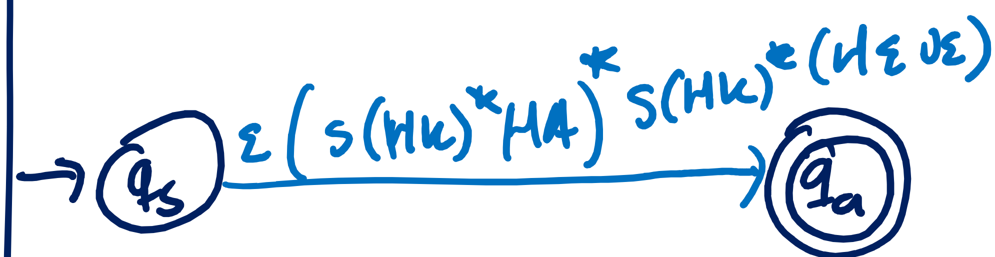
2. Remove state 2



3. Remove state 3



4. Remove state 1



Final regex:

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

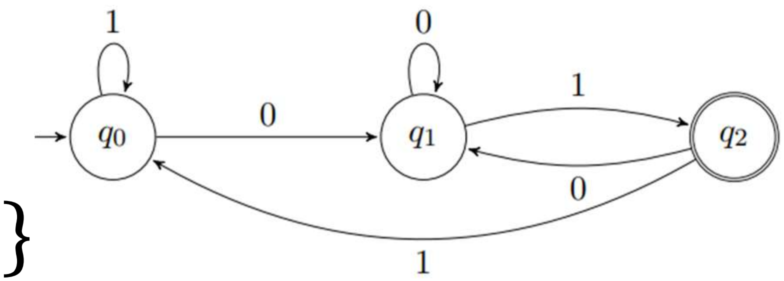
$$= (S(HK)^*HA)^*S(HK)^*(H \cup \varepsilon)$$

Limits of Finite Automata

Motivating Questions

- We've seen several techniques for showing languages are regular
  - Build DFA
  - Build NFA
  - Build Regular Expression
  - Use 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 $x = \varepsilon, y = 0, w = 01$

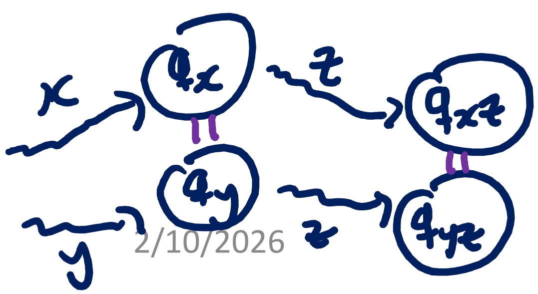
Let q_x = state M reaches after reading $x = \varepsilon$
q_y = " " $y = 0$
q_w = " " $w = 01$

WTS: q_x, q_y, q_z are all distinct states

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

Claim 2: $q_x \neq q_y$

Let $z = 1$, let q_{xz} = state M reaches when reading $xz = \varepsilon 1 = 1$ reject state
q_{yz} " " $yz = 01$ accept state



Suppose for contradiction that $q_x = q_y$
$\Rightarrow q_{xz} = q_{yz}$ contradiction, since can't be both accept and reject ✗

A General Technique

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

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. $x = \varepsilon, \quad y = 0$

$z = 1$ is a distinguishing extension because

$xz = \varepsilon 1 = 1 \notin L$

$yz = 01 \in L$

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

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

$x = \varepsilon \quad y = 01 \quad z = \varepsilon$ "

$x = 0 \quad y = 01 \quad z = \varepsilon$ "

A General Technique

$|S| \geq 2$

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

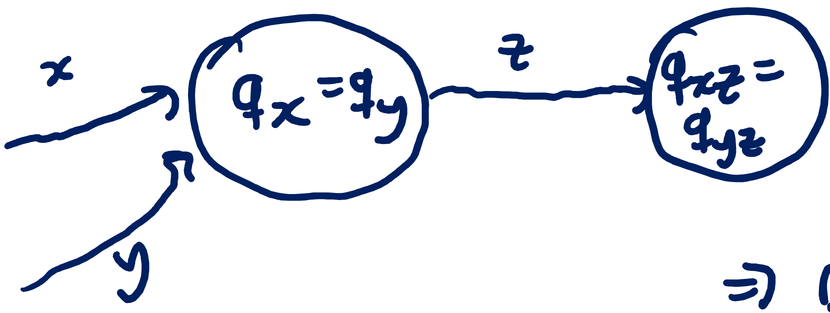
Claim: There are distinct strings $x, y \in S$ such that M ends up in same state on x and y why? Pigeon hole principle

Pigeons: Strings in S Map pigeon $x \in S$ to hole q_x (a state in M)

Holes: States in M by taking q_x = state M reaches when run on x

Fix strings x, y from claim, let q_x state reached on x ($= q_y$ state reached on y)

Let z be a distinguishing extension for x, y (guaranteed by S being pairwise distinguishable)



$q_{xz} = q_{yz}$ is either an accept state or a reject state (can't be both)

but exactly one of xz or $yz \in L$

$\Rightarrow$ DFA messes up on either xz or yz

Another Example

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

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

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

Imagine trying to construct a DFA for L:

If run on ε: be prepared to accept if I see exactly 2 chars

" 0: " 1 char

00: Accept

000: Reject

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

$xz = \varepsilon 0 = 0 \notin L \quad yz = 00 \in L$

$x = \varepsilon \quad y = 00 \quad z = \varepsilon$

$xz = \varepsilon \notin L \quad yz = 00\varepsilon = 00 \in L$

$x = \varepsilon \quad y = 000 \quad z = 00$

$xz = 00 \in L \quad yz = 00000 \notin L$

$x = 0 \quad y = 00 \quad z = \varepsilon$

$xz = 0 \notin L \quad yz = 00 \in L$

$x = 0 \quad y = 000 \quad z = 0$

$xz = 00 \in L \quad yz = 0000 \notin L$

$x = 00 \quad y = 000 \quad z = \varepsilon$

$xz = 00 \in L \quad yz = 000 \notin L$

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) $z = 00$

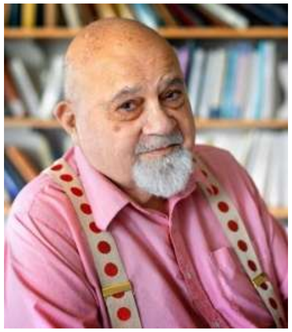
$xz = 000 \notin B$ $yz = 0000 \notin B$ ✗

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: If $\exists$ a DFA recognizing L with $< k$ states
Then there is no pairwise distinguishable set of size k

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

Contrapositive form: If L is recognizable by some DFA
Then L does not have an infinite pairwise dist. set

Proof that contrapositive of Corollary follows from Contrapositive of theorem:

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

$\Rightarrow$ L has no pairwise dist. set of size $\geq k+1$

$\Rightarrow$ L does not have an infinite pairwise distinguishable set

The Classic Example

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

Proof: We construct an infinite pairwise distinguishable set

$S = \{0^n \mid n \geq 0\}$

WTS: S is pairwise distinguishable by A

Let x, y be arbitrary distinct strings in S

can write $x = 0^m$ (for some $m \geq 0$)
$y = 0^n$ (for some $n \geq 0$, $n \neq m$)

Let $z = 1^n$. Then $xz = 0^m 1^n \notin A$ so z is a distinguishing extension for x, y.
$yz = 0^n 1^n \in A$