Automata and Grammars (BIE-AAG)

6. Properties of regular languages

Jan Holub

Department of Theoretical Computer Science
Faculty of Information Technology
Czech Technical University in Prague

Pumping lemma – problem statement

Suppose that language $L = \{0^n 1^n : n \geq 1\}$ were regular.

In such case language L is accepted by a finite automaton $\mathcal{A}$ with m states.

Pumping lemma – problem statement

Let an automaton $\mathcal{A}$ read a sequence 0^m. On the way it will travel as follows:

ε	p_0
0	p_1
00	p_2
$\ldots$	$\ldots$
0^m	p_m

i.e., $\exists i < j : p_i = p_j$ (Pigeonhole Principle). Label this state by q.

Pumping lemma – problem resolution

Suppose then that after reading a sequence 1^i the automaton goes from state q to state r.
It holds that:

- If state r is a final state, then the automaton accepts string 0^j1^i, which is unwanted.
- If state r is not a final state, then the automaton does not accept string 0^i1^i, which is unwanted as well.

By contradiction, language $L = \{0^n1^n : n \geq 1\}$ cannot be regular. □

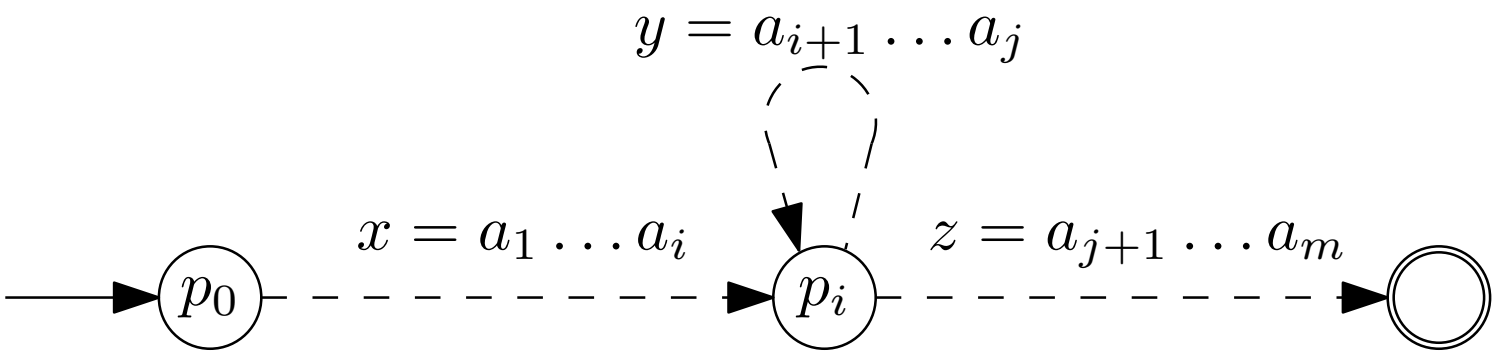
Pumping lemma formally

Pumping lemma

Let L be a regular language. Then for language L there exists a constant $p \geq 1$ such that for every sentence $w \in L$ it holds that:
If $|w| \geq p$, then w has form $w = xyz$ such that:

- $y \neq \varepsilon$ (i.e., $|y| \geq 1$),
- $|xy| \leq p$,
- $\forall k \geq 0$ it holds that $xy^k z \in L$.

Pumping lemma informally



Using PL for proving that L is not regular

Proof that language $L = \{0^n 1^n : n \geq 1\}$ is not regular

Let us assume that L is regular. Then by Pumping lemma it must hold that there exists a constant $p \geq 1$ such that for every sentence $w \in L$ it holds that:

If $|w| \geq p$, then w has form $w = xyz$ such that:

- $y \neq \varepsilon$ (i.e., $|y| \geq 1$),
- $|xy| \leq p$,
- $\forall k \geq 0$ it holds that $xy^k z \in L$.

Sentence $w = 0^p 1^p \in L$ is obviously longer than p and therefore it must meet the PL's requirements. To prove L is not regular, we try all possible partitionings of w into xyz. We must prove that PL conditions hold for none of them.

Using PL for proving that L is not regular

Proof that language $L = \{0^n 1^n : n \geq 1\}$ is not regular (cont.)

The partitioning either does not fullfils one the first two conditions, or

- xy is a non-empty sequence of zeroes
- y is non-empty sequence of zeroes
- z contains all the ones.

But then xy^0z (that is, we remove y from $w = xyz$) does not belong to L (because the number of zeroes in xy^0z is surely less than the number of ones)!

Therefore L is not regular. □

Using PL for proving that L is not regular

Proof that language $L = \{1^m : m \text{ is prime}\}$ is not regular

Let us assume that L is regular. Then by the Pumping lemma it must hold that there exists a constant $p \geq 1$ such that ... (see Pumping lemma).

Let us assume sentence $w = 1^m$ for prime $m \geq p + 2$.

Let $w = xyz$ be any partitioning satisfying $|xy| \leq p$ and $1 \leq |y|$. Let us consider a "pumped" sentence $w_1 = xy^{m-|y|}z$. We show that w_1 is not in L, which contradicts Pumping lemma.

Using PL for proving that L is not regular

Proof that language $L = \{1^m : m \text{ is prime}\}$ is not regular (cont.)

Let us consider the length of the sentence $w_1 = xy^{m-|y|}z$. It holds that
$|xy^{m-|y|}z| =$
$|xz| + (m - |y|) * |y| =$
$(m - |y|) + (m - |y|) * |y| =$
$(m - |y|) * (1 + |y|)$.

$|w_1|$ would be a prime only if either $(m - |y|) = 1$ or $(1 + |y|) = 1$.

- $(1 + |y|) \neq 1$, because $|y| \geq 1$.
- $m \geq p + 2$, $|y| \leq |xy| \leq p$, therefore $m - |y| \geq p + 2 - p = 2$.

For an arbitrary partitioning of $w = xyz \in L$ by the first two conditions of PL it holds that $xy^{m-|y|}z \notin L$.

Therefore, L **is not regular**. □

Using PL for proving that L is not regular

Example of a language that satisfies PL but it is not regular.

- $L = \{u : u = a^+b^ic^i \vee u = b^ic^j, 0 \leq i, 0 \leq j\}$

A comprehension question

How big is the constant p in Pumping lemma for a *finite* language?

(Note: Every finite regular language is regular, therefore Pumping lemma conditions must hold for it.)

Myhill-Nerode theorem: motivation

- It characterizes fundamental relationships between finite automata over alphabet Σ and certain equivalence relations over strings in Σ^*,
- it describes some of the necessary and sufficient conditions for a language to be regular (used often for proving that a language is not regular),
- it provides a formal basis for an elegant proof of existence of a unique (with respect to isomorphism) minimal DFA for a given regular language.

Equivalences

Definition (Equivalence – recall from BIE-ZDM)
Equivalence $\sim$ is a binary relation that is *reflexive*, *symmetric*, and *transitive*.

Definition (Equivalence class – recall from BIE-ZDM)
Equivalence class of an element a (denoted as $[a]$) in set X is the subset of all elements in X that are equivalent to a.

Definition (Quotient set)
The set of all equivalence classes in X is called the *quotient set* of X by $\sim$ and is denoted by $X/\sim$.

Definition (Index of equivalence)
Index of equivalence $\sim$ is the number of equivalence classes in the quotient set $\Sigma/\sim$. If there are infinitely many equivalence classes, the index is defined to be ∞.

Right congruence and prefix equivalence

Definition (Right congruence)
Let Σ be an alphabet and $\sim$ is an equivalence on Σ^*. Equivalence $\sim$ is a *right congruence*, if for every $u, v, w \in \Sigma^*$ it holds that:

$$u \sim v \Rightarrow uw \sim vw$$

Definition (Prefix equivalence)
Let L be an arbitrary (not necessarily regular) language over alphabet Σ. We define *prefix equivalence* for L, a relation $\sim_L$ on set Σ^* like this:

$$u \sim_L v \Leftrightarrow \forall w \in \Sigma^* : uw \in L \Leftrightarrow vw \in L$$

Observation
For each language L its prefix equivalence is its right congruence.

Myhill-Nerode theorem

Myhill-Nerode theorem

Let L be a language over Σ. Then the following statements are equivalent:

1. L is a language accepted by a finite automaton.
2. L is a union of certain equivalence classes of the quotient set of Σ^* by the right congruence on Σ^* with a finite index.
3. Relation $\sim_L$ has a finite index.

Proof

We prove the following implications:

- $1 \Rightarrow 2$.
- $2 \Rightarrow 3$.
- $3 \Rightarrow 1$.

Myhill-Nerode theorem: $1 \Rightarrow 2$

If L is accepted by a DFA, then L is a union of some classes of the quotient set by the right congruence on Σ^* with a finite index.

Let us introduce a general transition function $\hat{\delta}$ for DFA $M = (Q, \Sigma, \delta, q_0, F)$.
$\hat{\delta} : Q \times \Sigma^* \to Q$ such that
$\forall q_1, q_2 \in Q, w \in \Sigma^* : \hat{\delta}(q_1, w) = q_2 \Leftrightarrow (q_1, w) \vdash_M^* (q_2, \varepsilon)$.

Myhill-Nerode theorem: $1 \Rightarrow 2$

Proof

For a given L accepted by DFA M we construct $\sim$ with the necessary properties:

- Let $M = (Q, \Sigma, \delta, q_0, F)$ be total DFA.
- We set $\sim$ as a binary relation on Σ^* such that $u \sim v \Leftrightarrow \hat{\delta}(q_0, u) = \hat{\delta}(q_0, v)$.
- We show that $\sim$ has the necessary properties:
  - $\sim$ is an equivalence: it is reflexive, transitive, and symmetric.
  - $\sim$ has a finite index: the equivalence classes correspond to the automaton states.
  - $\sim$ is a right congruence: Let $u \sim v$ and $a \in \Sigma$. Then $\hat{\delta}(q_0, ua) = \delta(\hat{\delta}(q_0, u), a) = \delta(\hat{\delta}(q_0, v), a) = \hat{\delta}(q_0, va)$ and therefore $ua \sim va$.
  - L is a union of some equivalence classes of $\Sigma^*/\sim$ – those that correspond to F.

Myhill-Nerode theorem: $2 \Rightarrow 3$

If there is a relation $\sim$ satisfying condition 2, then $\sim_L$ has a finite index.

Proof

- For all $u, v \in \Sigma^*$ such that $u \sim v$, it holds that $u \sim_L v$:
  - Let $u \sim v$. We will show that also $u \sim_L v$, that is $\forall w \in \Sigma^* : uw \in L \Leftrightarrow vw \in L$.
  - We know $uw \sim vw$ and since L is a union of some classes of the quotient set $\Sigma^*/\sim$, it also holds $uw \in L \Leftrightarrow vw \in L$.
- We therefore know that $\sim \subseteq \sim_L$ (that is, $\sim_L$ is the largest right congruence with the given properties).
- Every class of $\sim$ is contained in some class of $\sim_L$.
- Index of $\sim_L$ cannot be greater than index of $\sim$.
- $\sim$ has a finite index and therefore even $\sim_L$ has a finite index.

Myhill-Nerode theorem: $3 \Rightarrow 1$

If $\sim_L$ has a finite index, then L is accepted by some finite automaton.

Proof

- We create $M = (Q, \Sigma, \delta, q_0, F)$ accepting L:
  - $Q = \Sigma^*/\sim_L$ (states are the classes of the quotient set),
  - $\forall u \in \Sigma^*, a \in \Sigma : \delta([u], a) = [ua]$,
  - $q_0 = [\varepsilon]$,
  - $F = \{[x] \mid x \in L\}$.
- The shown construction is correct, that is $L = L(M)$:
  - We show by induction over the length of word v that $\forall v \in \Sigma^* : \hat{\delta}([\varepsilon], v) = [v]$.
  - $v \in L \Leftrightarrow [v] \in F \Leftrightarrow \hat{\delta}([\varepsilon], v) \in F$. □

Proof of irregularity using M-N

Prove that language $L = \{0^n 1^n : n \geq 1\}$ is not regular.

Proof

- No strings $\varepsilon, 0, 0^2, 0^3, \ldots$ are $\sim_L$-equivalent, because $0^i 1^i \in L$, but $0^j 1^i \notin L$ for $i \neq j$.
- $\sim_L$ therefore has infinitely many classes (or infinite index).
- According to Myhill-Nerode theorem L cannot be accepted by a finite automaton. □

M-N theorem and DFA minimality

Theorem (second version of M-N theorem)
Number of states of any minimal DFA accepting L is equal to index of $\sim_L$.
(Such a DFA exists if and only if the index is finite.)

Proof

- Every DFA (without unreachable states) defines a certain right congruence with a finite index and vice versa.
- If L is regular, $\sim_L$ is the greatest right congruence with a finite index such that L is the union of certain classes of the appropriate quotient set.
- Finite automaton that corresponds to $\sim_L$ (see proof of $3 \Rightarrow 1$ of M-N), is therefore a minimal DFA accepting L. □