5. Pumping Lemma and Ultimate Periodicity

Recall: We already have a characterization of the regular/non-regular languages via Myhill & Nerode.

Goal: Establish a necessary condition for regularity (not a characterization), that has two benefits
- ↳ it can be generalized to other classes of languages
- ↳ it is easy to apply to disprove regularity.

Problem: To show that

$$L_1 = \{a^n b^n \mid n \in \mathbb{N}\}$$

is not regular, we have to reason about all regular languages.

An argument like "L_1 can count" is not sufficient:

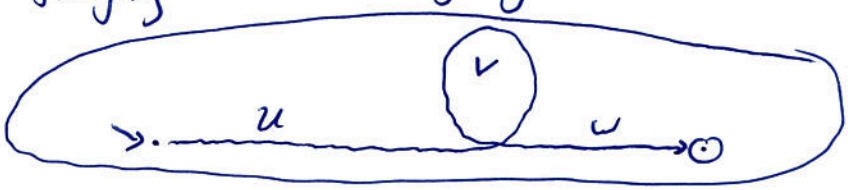
$L = \{w \in \{0,1\}^* \mid \#_0(w) = \#_1(w)\}$ is not regular but

$L' = \{w \in \{0,1\}^* \mid \#_{01}(w) = \#_{10}(w)\}$ is regular.

5.1 Pumping Lemma:

Idea:
- To show that L_1 is not regular, we prove that all regular languages have a structural property that is violated by L_1.
- The structural property states that words can be pumped if they are at least as long as a certain value. Pumping means we can indefinitely repeat an infix while staying in the language.

Intuition:



Theorem (Pumping Lemma):

For every $L \subseteq \Sigma^*$ there is a number $\rho_L \in \mathbb{N}$ so that for all $x \in L$ with $|x| \geq \rho_L$ there is a decomposition

$$x = uvw$$

satisfying the following:

(1) $|v| \geq 1$

(2) $|uv| \leq \rho_L$

(3) $uv^i w \in L$ for all $i \in \mathbb{N}$.

Note: · The number ρ_L depends on the choice of L.

· The representation of L does not go into the lemma.

Proof:

Let $L \subseteq \Sigma^*$ be regular with $L = L(A)$ and $A = (\Sigma, Q, q_0, \to, Q_F)$ an NFA.

We choose

$$\rho_L := |Q|.$$

Consider a word $x \in L$ with $|x| \geq \rho_L$.

Let $x = a_1 \dots a_r$.

Since A accepts x, we have a run

$$q_0 \xrightarrow{a_1} q_1 \xrightarrow{a_2} \dots \dots \xrightarrow{a_{r-1}} q_{r-1} \xrightarrow{a_r} q_r \in Q_F.$$

Since the run contains $r+1 > |x| \geq \rho_L = |Q|$ states, there is a state that repeats:

$$q_j = q_k \quad \text{with } j < k.$$

We consider the first such repetition,

which means

$$q_0, \dots, q_{k-1}$$

are all distinct.

Define $u := a_1 \dots a_j$

$v := a_{j+1} \dots a_k$

$w := a_{k+1} \dots a_r$.

With this choice, we have

(1) $v \neq \varepsilon$ since $j < k$.

(2) $|uv| \leq p_L$ since $q_0, \dots, q_{k-1}$ are all distinct.

(3) Automaton A can repeat the $q_j = q_k$-loop arbitrarily often when accepting. □

Example (Application of the pumping lemma):

Consider

$$L = \{ yy \mid y \in \{a,b\}^* \}.$$

To show that L is not regular, we reason towards a contradiction.

Assume L was regular.

Then there is a number $p_L \in \mathbb{N}$ with (1) to (3) as above.

Consider the word

$$x = a^{p_L} b\, a^{p_L} b \in L.$$

Since $|x| \geq p_L$,

we can decompose the word into

$x = uvw$ with (1) $v \neq \varepsilon$

(2) $|uv| \leq p_L$ and

(3) $\forall i \in \mathbb{N}: uv^i w \in L$.

Since $|uv| \leq p_L$ and $x = a^{p_L} b\, a^{p_L} b$,
u and v only consist of a's.
By Property (3), word

$$uvvw = a^{p_L + |v|} b\, a^{p_L} b \quad \text{with } |v| \geq 1$$

has to be in L. ↯ There is no way to split this word into yy. □

Remark:

- The pumping lemma (can also be) applied in contraposition:

$$\forall p_L \in \mathbb{N}\ \exists x \in L \text{ with } |x| \geq p_L$$
$$\forall \text{ decompositions } x = uvw \text{ with } |v| \geq 1 \text{ and } |uv| \leq p_L$$
$$\exists i \in \mathbb{N}: uv^i w \notin L$$
$$\Rightarrow L \text{ is not regular.}$$

- This is a game between defender ($\forall$) and spoiler ($\exists$):

Spoiler: I believe L is not regular.
Defender: Why not? Here is my number p_L.
Spoiler: I don't think so, here is my word x with $|x| \geq p_L$.
Does your p_L work with this?
Defender: For sure, take my decomposition
$x = uvw$
with $|v| \geq 1$ and $|uv| \leq p_L$.
Spoiler: Hh, but now you lost,
here is $i \in \mathbb{N}$ with $uv^i w \notin L$.

The pumping lemma states that

- if L is regular, defender has a winning strategy.

When applied in contraposition, it states that

- if spoiler has a winning strategy, then L is not regular.

A winning strategy is a strategy
to win no matter how well the opponent plays.

Example:

- Application of the pumping lemma in contraposition:

$$L_1 = \{a^n b^n \mid n \in \mathbb{N}\}.$$

Consider a number $p_{L_1} \in \mathbb{N}$.
We pick the word $x = a^{p_{L_1}} b^{p_{L_1}}$ with $|x| \geq p_{L_1}$.
Consider a decomposition

$$x = uvw \quad \text{with} \quad |v| \geq 1 \text{ and } |uv| \leq p_{L_1}.$$

Since $|uv| \leq p_{L_1}$, we know that u and v only consist of a's.

Take $i = 0$:

$$uv^0w = uw = a^{p_{L_1} - |v|} b^{p_{L_1}} \quad \text{with} \quad |v| \geq 1.$$

Since this word is not in L_1,
we can conclude that L_1 is not regular. □

- The following example uses the closure properties of regular languages:

$$L = \{w \in \{a, b\}^* \mid \#_a(w) = \#_b(w)\}.$$

If L was regular, then

$$L \cap a^*b^* = \{a^n b^n \mid n \in \mathbb{N}\} = L_1$$

was regular. We just showed that L_1 is not regular,
hence L cannot be regular. □

5.2 Ultimate Periodicity

Goal: Formalize the idea that regular languages cannot count (only count modulo).

Definition:
A set $U \subseteq \mathbb{N}$ is called *ultimately-periodic*,
if $\exists n \geq 0 \; \exists p > 0$:

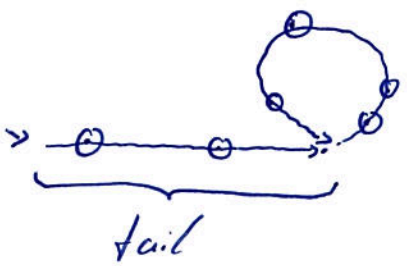
$$\forall m \geq n: \quad m \in U \text{ iff } m+p \in U.$$

Number p is called the *period* of U.

Idea: Except for an initial part,
numbers are in and out of U according to a repeating pattern.

Theorem:
Let $L \subseteq a^*$.
Then L is regular iff $\text{length}(L) := \{|w| \mid w \in L\}$
is ultimately periodic.

Proof:
"$\Rightarrow$" If L is regular, $L = L(A)$ for some DFA A.
Since there is only one letter, the DFA has the shape



Choose $n := |\text{tail}|$ = number of transitions in tail
$p := |\text{loop}|$.

"$\Leftarrow$" Build a DFA of tail length n
and loop of length p. □

Corollary:

Let $L \subseteq \Sigma^*$ be regular.

Then length(L) is an ultimately periodic set.

Proof:

Consider the homomorphism induced by

$$h : \Sigma \to \{a\}$$

(with $h(b) := a$ for all $b \in \Sigma$).

Then $h(w) = a^{|w|}$.

Since h preserves the length, we have

$$\text{length}(L) = \text{length}(h(L)).$$

But $h(L) \subseteq a^*$ is a regular language

(since the regular languages are closed under homomorphisms).

Hence, by the above theorem

length($h(L)$) is ultimately periodic

and so is length(L). □

Note:

Ultimately periodic sets are a first step towards the theory of Parikh images, semi-linear sets, and Presburger arithmetic.

Main results:

↳ Parikh(CFL) = Parikh(REG) = semi-linear sets
= Presburger-definable sets.

↳ Closure under $\cup, \cap, \bar{}, {}^*, h, h^{-1}$.