Nonregular languages

Consider the following language:

\[ B = \{0^n1^n \mid n \geq 0\}. \]

Is there a DFA that recognises \( B \)? No!

Intuition: an automaton should “remember” how many zeros it has seen.
It needs an infinite number of states for doing so.

Another example

\[ C = \{w \mid w \text{ has equal number of 0’s and 1’s}\}. \]

Pumping lemma

All regular languages have a certain property: each string of sufficiently large length contains a substring, which can be repeated any number of times, with the resulting strings remaining in the language.

**Pumping lemma.** If \( L \) is a regular language then there exists a number \( p \) such that if \( w \in L, |w| \geq p, \) then \( w \) could be represented as \( w = xyz \) and the following three conditions are satisfied:

1. for all \( i \geq 0 \) it holds that \( xy^i z = x y y \cdots y z \in L; \)

\[ i \text{ times} \]
2. $|y| > 0$;
3. $|xy| \leq p$.

Here $|w|$ denotes the length of the string $w$. Either $x$ or $z$ (or both) can be $\varepsilon$, but not $y$.

Idea of the proof. Let $p$ be the number of states of the automaton recognising $L$. Consider the run of the automaton on the input $w$ of length $|w| = n$:

If $n \geq p$, there should be a state $q_j$ which is repeated twice. We can repeat the substring between the appearances of $q_j$ arbitrary number of times.

Proof. Let $M = (Q, \Sigma, \delta, q_0, F)$ be a DFA recognising $L$, and $p$ be a number of states of $M$. Let $w = w_1w_2\ldots w_n$ be a string in the language $L$ with $n \geq p$. Let $q_0 = q_{i_0} \to q_{i_1} \to q_{i_2} \to \ldots \to q_{i_{n-1}} \to q_i = q_0$ be a sequence of states that $M$ enters when processing $w$: for each $j = 0, 1, \ldots, n - 1$ we have $\delta(q_{i_j}, w_j) = q_{i_{j+1}}$.

There are $n+1 \geq p+1$ states in this sequence. Since automaton has only $p$ states, amongst the first $p+1$ states $q_0, q_{i_1}, \ldots, q_{i_p}$ there exists a state $q_j$ that appears at least twice:

$q_0 = q_{i_0} \xrightarrow{w_1} \ldots \xrightarrow{w_l} q_{i_l} = q_j \xrightarrow{w_{l+1}} \ldots \xrightarrow{w_r} q_{i_r} = q_j \xrightarrow{w_{r+1}} \ldots \xrightarrow{w_n} q_i = q_0$.

Here $q_{i_l}$ is the first appearance of $q_j$ and $q_{i_r}$ is the second appearance of it ($l < r \leq p$). Denote:

$x = w_1w_2\ldots w_l$,
$y = w_{l+1}w_{l+2}\ldots w_r$,
$z = w_{r+1}w_{r+2}\ldots w_n$. 

2
Substring $x$ takes $M$ from $q_0$ to $q_j$, $y$ takes $M$ from $q_j$ to $q_j$ (forms a loop) and $z$ takes $M$ from $q_j$ to $q_a$. Therefore, $M$ should accept $xy^iz$ for any $i \geq 0$. We proved the first condition of the lemma.

Since $l < r$ then $|y| = r - l > 0$. This proves the second condition of the lemma.

And finally to prove the third condition of the lemma, we note that $|xy| = r \leq p$.

**Example 1.** Show that the language

$$B = \{0^n1^n \mid n \geq 0\}$$

is not regular.

Assume, to the contrary, that $B$ is regular. Let $p$ be the length given by the pumping lemma.

Take $w = 0^p1^p \in B$. Since $|w| \geq p$, we can write $w = xyz$, such that for all $i \geq 0$ it holds $xy^iz \in B$. From the condition 3 of the pumping lemma $|xy| \leq p$ and thus $|y| \leq |xy| \leq p$. So $y$ is situated entirely in the first part of $w$ and, hence, it contains only zeros. I.e. $xyyz$ has more zeros and therefore it does not belong to the language $B$. This contradicts the pumping lemma. Thus, the assumption that $B$ is regular was wrong and $B$ is not a regular language.

**Practise session**

1. Show that the language

$$L = \{w \mid w \text{ has equal number of 0’s and 1’s}\}$$

is not regular.

2. Show that the language $L = \{ss \mid s \in \{0,1\}^*\}$ is not regular.

3. Show that the language $L = \{1^n^2 \mid n \geq 0\}$ is not regular.

4. Take $L = \{0^n1^m \mid n > m\}$. Prove that $L$ is not regular.
Solutions to the practise session questions

1. Show that the language

\[ L = \{ w \mid w \text{ has equal number of 0's and 1's} \} \]

is not regular.

Solution. Assume, to the contrary, that \( L \) is regular and let \( p \) be the “pumping length”, given by the pumping lemma. Take \( w = 0^p1^p \in L \). Clearly, \( |w| \geq p \).

Then we can apply pumping lemma and represent \( w \) as \( w = xyz \) (with the three conditions holding).

From the condition 3, \( |xy| \leq p \), therefore \( y \) should contain only zeros (because first \( p \) characters of \( w \) are zeros). Hence, \( xyyz \) contains more zeros than \( xyz \). And, since in \( xyz \) number of zeros and ones was the same, \( xyyz \) contains more zeros than ones. But from the condition 1 of the lemma, \( xyyz \) should also belong to the language \( L \), i.e. should have equal number of zeros and ones. Contradiction!? Therefore the assumption that \( L \) is regular was wrong and \( L \) is not regular.

2. Show that the language \( L = \{ ss \mid s \in \{0, 1\}^* \} \) is not regular.

Solution. Assume, to the contrary, that \( L \) is regular and \( p \) is its “pumping length”. Take \( w = 0^p1^p \in L \). Obviously, \( |w| \geq p \).

Then there should exist \( x, y, z \), such that \( w = xyz \) and conditions of the pumping lemma hold. Since \( |xy| \leq p \), \( y \) consists of zeros only. Therefore \( xyyz \) does not have the form \( ss \) as its first block of zeros is longer than the second block of zeros. So \( xyyz \notin L \). Contradiction to the lemma’s condition 1! That’s why the assumption that \( L \) is regular was wrong.

3. Show that the language \( L = \{ 1^{n^2} \mid n \geq 0 \} \) is not regular.

Solution. Assume, to the contrary, that \( L \) is regular and \( p \) is its “pumping length”. Take \( w = 1^{p^2} \in L; \) \( |w| = p^2 \geq p \). We could represent \( w \) as \( w = xyz \), where for all \( i \geq 0 \) it holds that \( x y^i z \in L \).

Consider the string \( x y^2 z \). As \( |y| \leq |xy| \leq p \), then \( |xy^2 z| = |xyz| + |y| \leq p^2 + p \).

On the other hand, since \( |y| > 0 \), we have \( |xy^2 z| = |xyz| + |y| > |xyz| = p^2 \). We have:

\[ p^2 < |xy^2 z| \leq p^2 + p = p(p + 1) < (p + 1)^2, \]

which means that the length of \( xy^2 z \) is strictly between two squares of consecutive integers and therefore it cannot be a square of any integer. Thus, \( xy^2 z \notin L \). Contradiction! Therefore \( L \) is not regular.
4. Take $L = \{0^n1^m \mid n > m\}$. Prove that $L$ is not regular.

*Solution.* We repeat the same proof scheme:

- Assume that $L$ is regular and $p$ is its “pumping length”.
- Take $w = 0^{p+1}1^p \in L$. Clearly, $|w| \geq p$.
- We can apply the pumping lemma. I.e. there exist $x, y, z$, such that $w = xyz$ and the conditions of the pumping lemma are satisfied.
- From the pumping lemma’s condition 1, for all $i \geq 0$ we have that $xy^iz \in L$. In particular, take $i = 0$. Therefore $xz$ should be in $L$ too.\(^1\)
- Since $|xy| \leq p$, $y$ consists of zeros only. Therefore removing $y$ from $xyz$ decreases the number of zeros by at least one and $xz$ will have not more than $p$ zeros. This contradicts the definition of $L$ and, hence, $xz \notin L$.
- We end up with contradiction. This means that original assumption was wrong and $L$ is not a regular language.

\(^1\)But in the next steps we will show the opposite.