2.7 Regular Expressions

Let Σ be an alphabet, and let Γ be the following alphabet:

$$\Gamma := \Sigma \,\dot{\cup}\, \{\emptyset, \varepsilon, +, *, (,)\}.$$

The regular expressions $\mathrm{RA}(\Sigma)$ on Σ are defined as follows, where a language $L(r) \subseteq \Sigma^*$ is associated to each expression r:

(1)	$\emptyset \in \mathrm{RA}(\Sigma)$	$: L(\emptyset) := \emptyset,$
(2)	$\varepsilon \in \mathrm{RA}(\Sigma)$	$: L(\varepsilon) := \{\varepsilon\},$
(3)	$\forall a \in \Sigma : a \in \mathrm{RA}(\Sigma)$	$: L(a) := \{a\}.$
(4)	For all $r, s \in \mathrm{RA}(\Sigma), (r + s) \in \mathrm{RA}(\Sigma)$	$: L(r + s) := L(r) \cup L(s).$
(5)	For all $r, s \in \mathrm{RA}(\Sigma), (rs) \in \mathrm{RA}(\Sigma)$	$: L(rs) := L(r) \cdot L(s).$
(6)	For all $r \in \mathrm{RA}(\Sigma), (r^*) \in \mathrm{RA}(\Sigma)$	$: L(r^*) := (L(r))^*.$

Example:

$r = (0+1)^*00(0+1)^*$: $L(r) = \{\, u \in \{0,1\}^* \mid u \text{ contains the factor } 00 \,\}$.

$s = (0+\varepsilon)(1+10)^*$: $L(s) = \{0,1\}^* \smallsetminus L(r)$.

$t = (a+b)((a+b)(a+b))^*$: $L(t) = \{\, u \in \{a,b\}^* \mid |u| \equiv 1 \mod 2 \,\}$.

If $L = \{x_1, x_2, \ldots, x_m\}$, then $L = L((x_1 + (x_2 + (x_3 + \cdots + x_n) \cdots)))$.

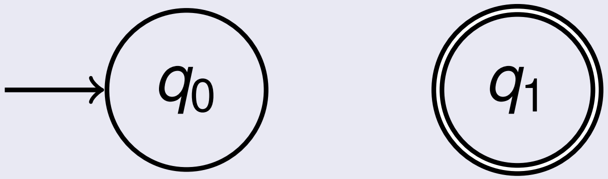
Theorem 2.30

From a regular expression r, an ε-NFA A_r can be constructed such that $L(A_r) = L(r)$.

Proof.

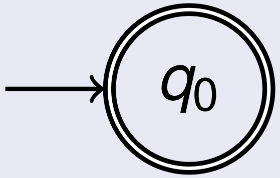
By induction on the number of operations used to build r:

(1) For $r = \emptyset$, we take the DFA A_r:

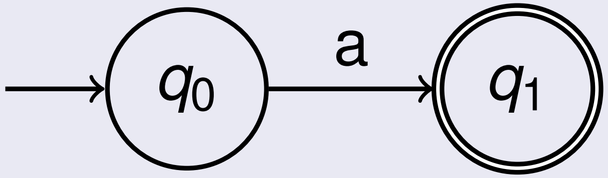


Proof of Theorem 2.30 (cont.)

(2) For $r = \varepsilon$, we take the DFA A_r:



(3) For $r = a$ $(\in \Sigma)$, we take the DFA A_r:



(4-6) By the induction hypothesis, there are ε-NFAs A_1 and A_2 such that $L(r_1) = L(A_1)$ and $L(r_2) = L(A_2)$. By Theorem 2.21, $\mathcal{L}(\varepsilon\text{-NFA})$ is closed under union, product, and Kleene star. Hence, from A_1 and A_2, we can construct ε-NFAs for $L(r_1 + r_2)$, $L(r_1 r_2)$, and $L(r_1^*)$.

□

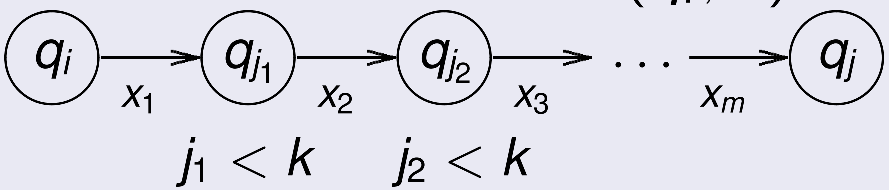
Theorem 2.31

From a DFA A, one can construct a regular expression r such that $L(A) = L(r)$.

Proof.

Let $A = (Q, \Sigma, \delta, q_0, F)$ be a DFA for $L = L(A) \subseteq \Sigma^*$.
Assume that $Q = \{q_0, q_1, q_2, \ldots, q_n\}$.
For all $i, j \in \{0, 1, \ldots, n\}$ and all $k \in \{0, 1, \ldots, n+1\}$,

$$R^k_{i,j} := \{\, x \in \Sigma^* \mid \delta(q_i, x) = q_j, \text{ and for all } v, w \in \Sigma^+, \text{ if } x = vw \text{ and } \delta(q_i, v) = q_\ell, \text{ then } \ell < k \,\}.$$



Thus, $x \in R^k_{i,j}$ iff the DFA A, starting in state q_i and reading input x, reaches state q_j, and all states q_ℓ encountered during this computation satisfy the condition that $\ell < k$.

Then $L = \bigcup_{q_j \in F} R^{n+1}_{0,j}$.

Proof of Theorem 2.31 (cont.)

$R^0_{i,j} = \{ a \in \Sigma \mid \delta(q_i, a) = q_j \} \qquad (i \neq j)$

$R^0_{i,i} = \{ a \in \Sigma \mid \delta(q_i, a) = q_i \} \cup \{\varepsilon\}$

Thus: $R^0_{i,j}$ are finite languages, that is,
there exist regular expressions $\alpha^0_{i,j}$ for them.

$R^{k+1}_{i,j} = R^k_{i,j} \cup R^k_{i,k}(R^k_{k,k})^* R^k_{k,j}$

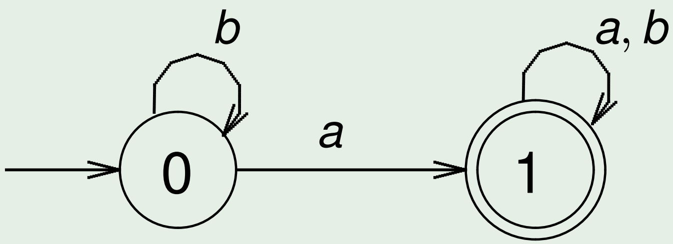
$\alpha^{k+1}_{i,j} := (\alpha^k_{i,j} + \alpha^k_{i,k}(\alpha^k_{k,k})^* \alpha^k_{k,j})$

Now $L = \bigcup_{q_j \in F} R^{n+1}_{0,j}$,

that is, if $F = \{q_{i_1}, q_{i_2}, \ldots, q_{i_m}\}$, then
$\gamma = (\alpha^{n+1}_{0,i_1} + (\alpha^{n+1}_{0,i_2} + \cdots + \alpha^{n+1}_{0,i_m})).$ □

Example:

A:



$R^k_{i,j} : i, j \in \{0, 1\}, k \in \{0, 1, 2\}.$

Example (cont.):

$$R^0_{0,0} = \{\varepsilon, b\}, \quad R^0_{0,1} = \{a\}, \quad R^0_{1,0} = \emptyset, \quad R^0_{1,1} = \{\varepsilon, a, b\}$$

$$R^1_{0,0} = R^0_{0,0} \cup R^0_{0,0} \cdot (R^0_{0,0})^* \cdot R^0_{0,0} = \{\varepsilon, b\} \cup \{\varepsilon, b\} \cdot \{\varepsilon, b\}^* \cdot \{\varepsilon, b\} = b^*$$

$$R^1_{0,1} = R^0_{0,1} \cup R^0_{0,0} \cdot (R^0_{0,0})^* \cdot R^0_{0,1} = \{a\} \cup \{\varepsilon, b\} \cdot \{\varepsilon, b\}^* \cdot \{a\} = b^* \cdot a$$

$$R^1_{1,0} = R^0_{1,0} \cup R^0_{1,0} \cdot (R^0_{0,0})^* \cdot R^0_{0,0} = \emptyset \cup \emptyset \cdot \{\varepsilon, b\}^* \cdot \{\varepsilon, b\} = \emptyset$$

$$R^1_{1,1} = R^0_{1,1} \cup R^0_{1,0} \cdot (R^0_{0,0})^* \cdot R^0_{0,1} = \{\varepsilon, a, b\} \cup \emptyset \cdot \{\varepsilon, b\}^* \cdot \{a\} = \{\varepsilon, a, b\}$$

$$R^2_{0,0} = R^1_{0,0} \cup R^1_{0,1} \cdot (R^1_{1,1})^* \cdot R^1_{1,0} = b^* \cup b^* a \cdot \{\varepsilon, a, b\}^* \cdot \emptyset = b^*$$

$$R^2_{0,1} = R^1_{0,1} \cup R^1_{0,1} \cdot (R^1_{1,1})^* \cdot R^1_{1,1} = b^* a \cup b^* a \cdot \{\varepsilon, a, b\}^* \cdot \{\varepsilon, a, b\} = b^* a \cdot \{a, b\}^*$$

$$R^2_{1,0} = R^1_{1,0} \cup R^1_{1,1} \cdot (R^1_{1,1})^* \cdot R^1_{1,0} = \emptyset \cup \ldots \cdot \emptyset = \emptyset$$

$$R^2_{1,1} = R^1_{1,1} \cup R^1_{1,1} \cdot (R^1_{1,1})^* \cdot R^1_{1,1} = \{\varepsilon, a, b\} \cup \{\varepsilon, a, b\} \cdot \{\varepsilon, a, b\}^* \cdot \{\varepsilon, a, b\} = \{a, b\}^*.$$

$$L(A) = R^2_{0,1} = b^* \cdot a \cdot \{a, b\}^* = \{ w \in \{a, b\}^* \mid |w|_a \geq 1 \} \qquad \square$$

Corollary 2.32 (Kleene's Theorem)

A language L is regular iff there exists a regular expression r such that $L(r) = L$.

A substitution $\varphi : \Sigma^* \to 2^{\Delta^*}$ is called regular, if $\varphi(a) \in \mathrm{REG}(\Delta)$ for all $a \in \Sigma$.

Corollary 2.33

The language class REG is closed under regular substitutions, that is, if $L \in \mathrm{REG}(\Sigma)$ and if $\varphi : \Sigma^* \to 2^{\Delta^*}$ is a regular substitution, then $\varphi(L) \in \mathrm{REG}(\Delta)$.

Proof.

For $L \in \mathrm{REG}(\Sigma)$, there exists a regular expression $r \in \mathrm{RA}(\Sigma)$ such that $L = L(r)$.

For each letter $a \in \Sigma$, there exists a regular expression $r_a \in \mathrm{RA}(\Delta)$ such that $\varphi(a) = L(r_a)$.

Proof of Corollary 2.33 (cont.)

Let $s \in \text{RA}(\Delta)$ be the regular expression that we obtain from r by replacing each occurrence of each letter $a \in \Sigma$ by the expression r_a.

Claim:

$L(s) = \varphi(L)$.

Proof by induction on the structure of r:

1. If $r = \emptyset$, then $L = \emptyset = \varphi(L)$ and $s = \emptyset$, that is, $\varphi(L) = L(s)$.
2. If $r = \varepsilon$, then $L = \{\varepsilon\} = \varphi(L)$ and $s = \varepsilon$.
3. If $r = a \in \Sigma$, then $L = \{a\}$ and $s = r_a$. Hence,

$$\varphi(L) = \varphi(a) = L(r_a) = L(s).$$

Proof (cont.)

4. If $r = (r_1 + r_2)$, then $L = L(r_1) \cup L(r_2)$ and $s = (s_1 + s_2)$. By the ind. hyp., $L(s_i) = \varphi(L(r_i))$, $i = 1, 2$. Hence,
$$\varphi(L) = \varphi(L(r_1)) \cup \varphi(L(r_2)) = L(s_1) \cup L(s_2) = L(s).$$

5. If $r = (r_1 r_2)$, then $L = L(r_1) \cdot L(r_2)$ and $s = (s_1 s_2)$. By the ind. hyp., $L(s_i) = \varphi(L(r_i))$, $i = 1, 2$. Hence,
$$\varphi(L) = \varphi(L(r_1)) \cdot \varphi(L(r_2)) = L(s_1) \cdot L(s_2) = L(s).$$

6. If $r = (r_1^*)$, then $L = (L(r_1))^*$ and $s = (s_1^*)$. By the ind. hyp., $L(s_1) = \varphi(L(r_1))$. Hence,
$$\varphi(L) = (\varphi(L(r_1)))^* = (L(s_1))^* = L(s).$$

It follows that $\varphi(L) \in \mathrm{REG}(\Delta)$. □