Theory of Computation

Chapter 1

Regular Languages

1.1 Languages

Definition 1.1. An **alphabet** is a finite nonempty set of symbols.

Definition 1.2. Let Σ be an alphabet.

- A **string** over Σ is a finite sequence of symbols from Σ. The collection of all strings over Σ is denoted by Σ^*.
- The **length** of a string w, denoted by $|w|$, is the number of symbols it contains.
- The string containing no symbols is called the **empty string**, denoted by ϵ.

Definition 1.3. A subset of Σ^* is called a **language** over Σ.

1.2 Deterministic Finite State Automata

Definition 1.4. A **deterministic finite state automaton** (DFA) is a tuple

$$A = (Q, \Sigma, \delta, q_0, F),$$

where each component is as follows.

- Q is a finite set of **states**.
- Σ is a finite set of input symbols.
- $\delta : Q \times \Sigma \to Q$ is a function, called the **transition function**.
- $q_0 \in Q$ is called the **start state**.
- $F \subseteq Q$ is called the **accepting states**.

Definition 1.5. Let $A = (Q, \Sigma, \delta, q_0, F)$ be a DFA. For each string $w \in \Sigma^*$, we define $\delta_w : Q \to Q$ as follows, where $a \in \Sigma$ and $x \in \Sigma^*$.

- $\delta_\epsilon(p) = p$ for each $p \in Q$.
- $\delta_a(p) = \delta(p, a)$ for each $p \in Q$.
- $\delta_{xa}(p) = \delta(\delta_x(p), a)$ for each $p \in Q$.

Definition 1.6. Let $A = (Q, \Sigma, \delta, q_0, F)$ be a DFA.

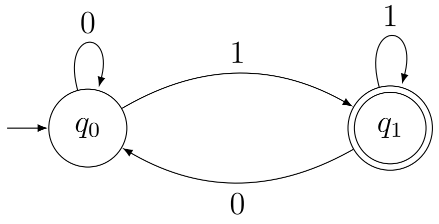
- We say that A accepts a string $w \in \Sigma^*$ if $\delta_w(q_0) \in F$.
- The **language** of A, denoted $L(A)$, is defined as the set of strings that are accepted by A.

Definition 1.7. A language L is **regular** if there exists a DFA A such that $L(A) = L$.

Example. Let $A = (\{q_0, q_1\}, \{0, 1\}, \delta, q_0, \{q_1\})$ be a DFA, where the transition function δ is as follows.

	0	1
q_0	q_0	q_1
q_1	q_0	q_1

Instead of using the formal definition, one can also draw a state diagram of A as follows.



It can be shown that a string $w \in \{0, 1\}^*$ is accepted by A if and only if w ends with 1. Thus, the language $L = \{w \in \{0, 1\}^* : w \text{ ends with } 1\}$ is regular.

Theorem 1.8. If L is a regular language over Σ, then $\Sigma^* \setminus L$ is also regular.

Proof. Let $A = (Q, \Sigma, \delta, q_0, F)$ be a DFA with $L = L(A)$. Let $A' = (Q, \Sigma, \delta, q_0, Q \setminus F)$. Then for each $w \in \Sigma^*$, we have

$$w \in L(A') \quad \Leftrightarrow \quad \delta_w(q_0) \in Q \setminus F \quad \Leftrightarrow \quad w \notin L(A).$$

Thus, $L(A') = \Sigma^* \setminus L(A)$, implying that $\Sigma^* \setminus L$ is regular. $\square$

Theorem 1.9. If L_1 and L_2 are regular languages over Σ, then $L_1 \cup L_2$ is also regular.

Proof. Let

$$A_1 = (Q_1, \Sigma, \delta^{(1)}, q_1, F_1) \quad \text{and} \quad A_2 = (Q_2, \Sigma, \delta^{(2)}, q_2, F_2)$$

be DFAs with $L_1 = L(A_1)$ and $L_2 = L(A_2)$. We construct the DFA

$$A = (Q_1 \times Q_2, \Sigma, \delta, (q_1, q_2), F)$$

as follows.

- $\delta((p, q), a) = (\delta^{(1)}(p, a), \delta^{(2)}(q, a))$ for each $p \in Q_1$, $q \in Q_2$ and $a \in \Sigma$.
- $F = \{(p, q) : p \in F_1 \text{ or } q \in F_2\}$.

It can be shown that for each string $w \in \Sigma^*$, we have

$$\begin{aligned} w \in L(A) \quad &\Leftrightarrow \quad \delta_w((q_1, q_2)) \in F \\ &\Leftrightarrow \quad \delta_w^{(1)}(q_1) \in F_1 \text{ or } \delta_w^{(2)}(q_2) \in F_2 \\ &\Leftrightarrow \quad w \in L(A_1) \text{ or } w \in L(A_2). \end{aligned}$$

Thus, $L(A) = L(A_1) \cup L(A_2)$, implying that $L_1 \cup L_2$ is regular. $\square$

Corollary 1.10. If L_1 and L_2 are regular languages over Σ, then $L_1 \cap L_2$ is also regular.

Proof. Straightforward since by De Morgan's law we have

$$L_1 \cap L_2 = \Sigma^* \setminus ((\Sigma^* \setminus L_1) \cup (\Sigma^* \setminus L_2)).$$

$\square$

1.3 Nondeterministic Finite State Automata

Definition 1.11. A **nondeterministic finite state automaton** (NFA) is a tuple

$$A = (Q, \Sigma, \delta, q_0, F),$$

where each component is as follows.

- Q is a finite set of **states**.
- Σ is a finite set of input symbols.
- $\delta : Q \times \Sigma \times Q$ is a relation, called the **transition relation**.
- $q_0 \in Q$ is called the **start state**.
- $F \subseteq Q$ is called the **accepting states**.

Definition 1.12. Let $A = (Q, \Sigma, \delta, q_0, F)$ be an NFA. For each string $w \in \Sigma^*$, we define $\delta_w \subseteq Q \times Q$ as follows, where $a \in \Sigma$ and $x \in \Sigma^*$.

- $\delta_\epsilon = \{(p, q) : p = q\}$.
- $\delta_a = \{(p, q) : (p, a, q) \in \delta\}$.
- $\delta_{xa} = \{(p, q) : (p, r) \in \delta_x$ and $(r, q) \in \delta_a$ for some $r \in Q\}$.

Definition 1.13. Let $A = (Q, \Sigma, \delta, q_0, F)$ be an NFA.

- We say that A accepts a string $w \in \Sigma^*$ if there exists $q \in F$ such that $(q_0, q) \in \delta_w$.
- The **language** of A, denoted $L(A)$, is defined as the set of strings that are accepted by A.

Theorem 1.14. For every NFA A, there is a DFA A' with $L(A') = L(A)$.

Proof. Let $A = (Q, \Sigma, \delta, q_0, F)$. We construct $A' = (\mathcal{P}(Q), \Sigma, \Delta, \{q_0\}, \Phi)$ as follows.

- $\Delta : \mathcal{P}(Q) \times \Sigma \to \mathcal{P}(Q)$ is the function with

  $$\Delta_a(P) = \bigcup_{p \in P} \{q \in Q : (p, q) \in \delta_a\}$$

  for any $P \subseteq Q$ and $a \in \Sigma$.
- $\Phi = \{P \subseteq Q : P \cap F \neq \varnothing\}$.

Now we prove that for any $w \in \Sigma^*$, for any $q \in Q$ and for any $P \subseteq Q$, we have $q \in \Delta_w(P)$ if and only if $(p, q) \in \delta_w$ for some $p \in P$. For the induction basis, let $w = \epsilon$, and we have

$$q \in \Delta_\epsilon(P) \quad \Leftrightarrow \quad q \in P \quad \Leftrightarrow \quad (p, q) \in \delta_\epsilon \text{ for some } p \in P.$$

For the induction step, let $w = xa$, where x is any string and a is any symbol. Note that by the construction of Δ, we have $q \in \Delta_a(P)$ if and only if $(p, q) \in \delta_a$ for some $p \in P$. Thus, we can conclude that

$$\begin{aligned} q \in \Delta_{xa}(P) \quad &\Leftrightarrow \quad q \in \Delta_a(\Delta_x(P)) \\ &\Leftrightarrow \quad (r, q) \in \delta_a \text{ for some } r \in \Delta_x(P) \\ & \quad \text{and } (p, r) \in \delta_x \text{ for some } p \in P \\ &\Leftrightarrow \quad (p, q) \in \delta_{xa} \text{ for some } p \in P. \end{aligned}$$

Finally we prove that $L(A') = L(A)$, which is given by

$$\begin{aligned} w \in L(A') \quad &\Leftrightarrow \quad \Delta_w(\{q_0\}) \in \Phi \\ &\Leftrightarrow \quad \Delta_w(\{q_0\}) \cap F \neq \varnothing \\ &\Leftrightarrow \quad q \in \Delta_w(\{q_0\}) \text{ for some } q \in F \\ &\Leftrightarrow \quad (p, q) \in \delta_w \text{ for some } q \in F \text{ and } p \in \{q_0\} \\ &\Leftrightarrow \quad (q_0, q) \in \delta_w \text{ for some } q \in F \\ &\Leftrightarrow \quad w \in L(A). \end{aligned}$$

□

Theorem 1.15. If L_1 and L_2 are regular languages over Σ, then $L_1 L_2$ is also regular.

Proof. Let $A_1 = (Q_1, \Sigma, \delta^{(1)}, q_1, F_1)$ and $A_2 = (Q_2, \Sigma, \delta^{(2)}, q_2, F_2)$ be NFAs such that $L_1 = L(A_1)$ and $L_2 = L(A_2)$. We construct an NFA

$$A = (Q_1 \cup Q_2, \Sigma, \delta, q_1, F)$$

as follows.

- $\delta = \delta^{(1)} \cup \delta^{(2)} \cup \{(p, a, q) \in Q \times \Sigma \times Q : p \in F_1 \text{ and } (q_2, a, q) \in \delta^{(2)}\}$.
- If $q_2 \in F_2$, let $F = F_1 \cup F_2$. Otherwise, let $F = F_2$.

It can be shown that $L(A) = L(A_1)L(A_2)$, and thus $L_1 L_2$ is regular. □

Theorem 1.16. If L is a regular language over Σ, then L^* is also regular.

Proof. Let $A = (Q, \Sigma, \delta, q_0, F)$ be an NFA with $L = L(A)$. We construct an NFA

$$A' = (Q \cup \{q_0'\}, \Sigma, \delta', q_0', F \cup \{q_0'\})$$

with

$$\delta' = \delta \cup \{(p, a, q) \in Q \times \Sigma \times Q : p \in F \cup \{q_0'\} \text{ and } (q_0, a, q) \in \delta\}.$$

It can be shown that $L(A') = (L(A))^*$, and thus L^* is regular. □

1.4 Regular Expressions

Definition 1.17. Let Σ be an alphabet. A **regular expression** over Σ is a string in the minimal language over $\Sigma \cup \{\varnothing, \epsilon, {}^*, +, (,)\}$ that satisfies the following conditions.

1. $\varnothing$ is a regular expression.
2. ϵ is a regular expression.
3. If $a \in \Sigma$, then a is a regular expression.
4. If e_1 and e_2 are regular expressions, then so is $(e_1 e_2)$.
5. If e_1 and e_2 are regular expressions, then so is $(e_1 + e_2)$.
6. If e is a regular expression, then so is $(e)^*$.

Definition 1.18. A regular expression e over an alphabet Σ defines a language $L(e)$ as follows.

1. $L(\varnothing) = \varnothing$.
2. $L(\epsilon) = \{\epsilon\}$.
3. $L(a) = \{a\}$ for each $a \in \Sigma$.
4. $L((e_1 e_2)) = L(e_1)L(e_2)$ for each regular expressions e_1 and e_2.
5. $L((e_1 + e_2)) = L(e_1) \cup L(e_2)$ for each regular expressions e_1 and e_2.
6. $L((e)^*) = L(e^*)$ for each regular expression e.

Remark. From now on, we may omit parentheses if there is no ambiguity.

Lemma 1.19. If e is a regular expression over an alphabet Σ, then $L(e)$ is regular.

Proof. It can be easily shown that $\varnothing$ and $\{\epsilon\}$ are regular. Moreover, $\{a\}$ is regular for each $a \in \Sigma$. Thus, by Theorem 1.9, Theorem 1.15 and Theorem 1.16, we can conclude that for all regular expressions e, $L(e)$ is regular. $\square$

Lemma 1.20. If L is a regular language over an alphabet Σ, then there is a regular expression e over Σ such that $L(e) = L$.

Proof. Since L is regular, there exists a DFA $A = (Q, \Sigma, \delta, q_0, F)$ with $L(A) = L$. Suppose that $Q = \{p_1, p_2, \ldots, p_n\}$ with $p_1 = q_0$. For any $i, j \in \{1, \ldots, n\}$ and for any $k \in \{0, \ldots, n\}$, let $L_{ij}^{(k)}$ denote the language of strings w such that

- $\delta_w(p_i) = p_j$, and
- for each string x with $\epsilon \sqsubset x \sqsubset w$, we have $\delta_x(p_i) = p_\ell$ for some $\ell \in \{1, \ldots, k\}$.

We are going to prove that for all $i, j \in \{1, \ldots, n\}$ and $k \in \{0, \ldots, n\}$, there exists a regular expression $e_{ij}^{(k)}$ such that

$$L\big(e_{ij}^{(k)}\big) = L_{ij}^{(k)}.$$

The proof is by induction on k. For the induction basis, let $k = 0$. Let $\Pi_{ij} \subseteq \Sigma$ denote the set of symbols a with $\delta_a(p_i) = p_j$. If $i \neq j$, we have

$$L_{ij}^{(0)} = \bigcup_{a \in \Pi_{ij}} \{a\},$$

and thus we can construct $e_{ij}^{(0)}$ by

$$e_{ij}^{(0)} = \sum_{a \in \Pi_{ij}} a.$$

(If $\Pi_{ij} = \varnothing$, then the summation is defined as $\varnothing$.) If $i = j$, we have

$$L_{ii}^{(0)} = \{\epsilon\} \cup \bigcup_{a \in \Pi_{ii}} \{a\},$$

and thus we can construct $e_{ii}^{(0)}$ by

$$e_{ii}^{(0)} = \epsilon + \sum_{a \in \Pi_{ii}} a.$$

Now for the induction step, let $k \geq 1$. Suppose that $w \in L_{ij}^{(k)}$. If there is no string x with $\epsilon \sqsubset x \sqsubset w$ such that $\delta_x(p_i) = p_k$, then we have

$$w \in L_{ij}^{(k-1)}.$$

Otherwise, let $x_0, x_1, \ldots, x_\ell$ be all strings with $\epsilon \sqsubset x_0 \sqsubset x_1 \sqsubset \cdots \sqsubset x_\ell \sqsubset w$ such that

$$\delta_{x_0}(p_i) = \delta_{x_1}(p_i) = \cdots = \delta_{x_\ell}(p_i) = p_k.$$

Let $u_0, u_1, \ldots, u_{\ell+1}$ be strings such that

$$w = u_0 u_1 \cdots u_{\ell+1},$$

and $x_h = u_0 u_1 \cdots u_h$ for each $h \in \{0, \ldots, \ell\}$. Since $u_0 \in L_{ik}^{(k-1)}$, $u_{\ell+1} \in L_{kj}^{(k-1)}$, and $u_h \in L_{kk}^{(k-1)}$ for each $h \in \{1, \ldots, \ell\}$, it follows that

$$w \in L_{ik}^{(k-1)} \left(L_{kk}^{(k-1)}\right)^* L_{kj}^{(k-1)}.$$

As a result, we have

$$L_{ij}^{(k)} \subseteq L_{ij}^{(k-1)} \cup L_{ik}^{(k-1)} \left(L_{kk}^{(k-1)}\right)^* L_{kj}^{(k-1)},$$

implying

$$L_{ij}^{(k)} = L_{ij}^{(k-1)} \cup L_{ik}^{(k-1)} \left(L_{kk}^{(k-1)}\right)^* L_{kj}^{(k-1)}.$$

Therefore, we can construct $e_{ij}^{(k)}$ by

$$e_{ij}^{(k)} = e_{ij}^{(k-1)} + e_{ik}^{(k-1)} \left(e_{kk}^{(k-1)} \right)^* e_{kj}^{(k-1)}.$$

Now we can construct the regular expression e with $L(e) = L$ as follows. Let Φ be the set of integers $j \in \{1, \dots, n\}$ such that $p_j \in F$. Since

$$L = \bigcup_{j \in \Phi} L_{1j}^{(n)},$$

we can construct e by

$$e = \sum_{j \in \Phi} e_{1j}^{(n)},$$

which completes the proof. □

Theorem 1.21. Let Σ be an alphabet. A language L over Σ is regular if and only if there is a regular expression e over Σ such that $L(e) = L$.

Proof. Straightforward by Lemma 1.19 and Lemma 1.20. □