形式语言与自动机　作业 3

1951112　林日中

习题 4.1.2　证明下列语言都不是正则的：

e)　由 0 和 1 构成的 ww 形式的串的集合，也就是某个串重复的串集合.

f)　由 0 和 1 构成的 ww^R 形式的串的集合，也就是由某个串后面跟着它的反转所构成的串的集合.（一个串的逆的形式化定义见 4.2.2 节.）

g)　由 0 和 1 构成的 $w\bar{w}$ 形式的串的集合，其中 $\bar{w}$ 是把 w 中所有的 0 都换成 1 同时把所有的 1 都换成 0 而得到的串，例如，$\overline{011} = 100$，因此 011100 是该语言中的一个串.

h)　所有由 0 和 1 构成的 $w1^n$ 形式的串的集合，其中 w 是由 0 和 1 构成的长度为 n 的串.

证明.

e)　假设这个语言 L 是正则的. 那么由泵引理，存在一常数 $n > 0$，对于语言 L 中每个满足 $|w| \geq n$ 的字符串 w，存在一组 x, y, z 使得 $w = xyz$ 且有 $|xy| \leq n, |z| \geq 1$，并且有 $\forall k \geq 0, xy^kz \in L$.

考察串 $w = 0^n1^n, w' = ww = 0^n1^n0^n1^n$，显然，有 $|w'| \geq n$. 按照泵引理对串 w' 生成一组分解 x, y, z. 因为 $|xy| \leq n$，那么泵 y 必仅由 0 组成；那么，由 L 是正则的，有 $xz \in L$，其中，$xz = 0^{n-|y|}1^n0^n1^n, |y| \geq 1$；显然，有 $xz \notin L$.

因此，存在由 0 和 1 构成的 ww 形式的串不满足泵引理，这与假设矛盾. 故这个语言不是正则的. □

f)　假设这个语言 L 是正则的. 那么由泵引理，存在一常数 $n > 0$，对于语言 L 中每个满足 $|w| \geq n$ 的字符串 w，存在一组 x, y, z 使得 $w = xyz$ 且有 $|xy| \leq n, |z| \geq 1$，并且有 $\forall k \geq 0, xy^kz \in L$.

考察串 $w = 0^n1, w' = ww^R = 0^n110^n$，显然有 $|w'| \geq n$. 按照泵引理对串 w' 生成一组分解 x, y, z. 因为 $|xy| \leq n$，那么泵 y 必仅由 0 组成；那么，由 L 是正则的，有 $xz \in L$，其中，$xz = 0^{n-|y|}110^n, |y| \geq 1$；显然，有 $xz \notin L$.

因此，存在由 0 和 1 构成的 ww^R 形式的串不满足泵引理，这与假设矛盾. 故这个语言不是正则的. □

g) 假设这个语言 L 是正则的. 那么由泵引理，存在一常数 $n>0$，对于语言 L 中每个满足 $|w|\ge n$ 的字符串 w，存在一组 x,y,z 使得 $w=xyz$ 且有 $|xy|\le n,|z|\ge 1$，并且有 $\forall k\ge 0,xy^kz\in L$.

考察串 $w=0^n,w'=w\bar{w}=0^n1^n$，显然有 $|w'|\ge n$. 按照泵引理对串 w' 生成一组分解 x,y,z. 因为 $|xy|\le n$，那么泵 y 必仅由 0 组成；那么，由 L 是正则的，有 $xz\in L$，其中，$xz=0^{n-|y|}1^n,|y|\ge 1$；显然，有 $xz\notin L$.

因此，存在由 0 和 1 构成的 $w\bar{w}$ 形式的串不满足泵引理，这与假设矛盾. 故这个语言不是正则的. □

h) 假设这个语言 L 是正则的. 那么由泵引理，存在一常数 $n>0$，对于语言 L 中每个满足 $|w|\ge n$ 的字符串 w，存在一组 x,y,z 使得 $w=xyz$ 且有 $|xy|\le n,|z|\ge 1$，并且有 $\forall k\ge 0,xy^kz\in L$.

考察串 $w=0^n,w'=w1^n=0^n1^n$，显然有 $|w'|\ge n$. 按照泵引理对串 w' 生成一组分解 x,y,z. 因为 $|xy|\le n$，那么泵 y 必仅由 0 组成；那么，由 L 是正则的，有 $xz\in L$，其中，$xz=0^{n-|y|}1^n,|y|\ge 1$；显然，有 $xz\notin L$.

因此，存在由 0 和 1 构成的 $w1^n,|w|=n$ 形式的串不满足泵引理，这与假设矛盾. 故这个语言不是正则的. □

习题 4.1.3 证明下列语言都不是正则的：

a) 所有满足以下条件的串的集合：由 0 和 1 构成，开头的是 1，并且当我们把该串看作是一个整数时该整数是一个素数.

b) 所有满足以下条件的 0^i1^j 形式的串的集合：i 和 j 的最大公约数是 1.

证明.

a) 由初等数论知识，存在任意大的素数. 假设这个语言 L 是正则的. 那么由泵引理，存在一常数 $n>0$，对于语言 L 中每个满足 $|w|\ge n$ 的字符串 w，存在一组 x,y,z 使得 $w=xyz$ 且有 $|xy|\le n,|z|\ge 1$，并且有 $\forall k\ge 0,xy^kz\in L$.

考察素数 $q=x\cdot 2^{m+k}+y\cdot 2^k+z\in L$，$w_q,\left|w_q\right|\ge n$ 为 q 对应的二进制串，即 $w_q=xyz,|y|=m,|z|=k$. x,y,z 即为串 w_q 根据泵引理生成的一组分解. 那么，由 L 是正则的，有

$$w_p = xy^q z \in L, p = x \cdot 2^{qm+k} + y \cdot 2^k \cdot \sum_{j=0}^{q-1} 2^{jm} + z$$

故 p 也是素数.

由费马小定理 $2^{q-1} \equiv 1 (\bmod q)$，有

$$2^{qn} = 2^{(q-1)n} \cdot 2^n \equiv 2^n (\bmod q)$$

因此

$$2^{qn} - 1 \equiv 2^n - 1 (\bmod q)$$

因为 $2^n - 1 < q$，有

$$\frac{2^{qn} - 1}{2^n - 1} = 1 + 2^n + \cdots + 2^{(q-1)n} \equiv 1 (\bmod q)$$

因此

$$p \equiv x \cdot 2^{m+k} + y \cdot 2^k + z = q \equiv 0 (\bmod q)$$

即 $q|p$. 这与 p 是素数矛盾.

故 L 不是正则语言. □

b) 假设这个语言 L 是正则的. 那么由泵引理，存在一常数 $n > 0$，对于语言 L 中每个满足 $|w| \geq n$ 的字符串 w，存在一组 x, y, z 使得 $w = xyz$ 且有 $|xy| \leq n, |z| \geq 1$，并且有 $\forall k \geq 0, xy^k z \in L$.

考察串 $w = 0^q 1^{(q-1)!}$，其中，q 是不比 n 小的最小质数，那么显然有 $|w| \geq n$，且 $\gcd(q, (q-1)!) = 1$. 按照泵引理对串 w 生成一组分解 x, y, z. 因为 $|xy| \leq n$，那么泵 y 必仅由 0 组成；那么，由 L 是正则的，有 $xz \in L$，其中，$xz = 0^{q-|y|} 1^{(q-1)!}, |y| \geq 1$；而 $\gcd(q-k, (q-1)!) = q-k, k \geq 1$. 显然，有 $xz \notin L$.

因此，存在 $0^i 1^j, \gcd(i, j) = 1$ 形式的串不满足泵引理，这与假设矛盾. 故这个语言不是正则的. □

习题 4. 2. 2 如果 L 是一个语言，a 是一个符号，则 L/a（称作 L 和 a 的商）是所有满足如下条件的串 w 的集合：wa 属于 L. 例如，如果 $L = \{a, aab, baa\}$，则 $L/a = \{\varepsilon, ba\}$，证明：如果 L 是正则的，那么 L/a 也是. 提示：从 L 的 DFA 出发，考虑接受状态的集合.

证明. 因为 L 是正则的，那么存在 DFA $M = (Q, \Sigma, \delta, s, F)$ 使得 $L(M) = L$. 根据 M

构造一个新的 DFA $M' = (Q, \Sigma, \delta, s, F')$，它与 M 具有相同的状态集、转移函数和初始状态；不同之处是，接受状态 F' 被定义为

$$F' = \{q | \delta(q, a) \in F\}$$

即状态 q 是 M' 的接受状态，当且仅当 $\delta(q, a)$ 是 M 的接受状态.

那么，对于每一个到达接受状态 $q' \in F'$ 的输入 w，有 $wa \in L$，因为读取 w 后再读取一个 a，会跳转到 M 的接受状态 $q \in F$. 因此，自动机 M' 对应的语言 $L(M') = L/a$.

故如果 L 是正则的，L/a 也是正则的. □

习题 4.2.7 如果 $w = a_1 a_2 \cdots a_n$ 和 $x = b_1 b_2 \cdots b_n$ 是同样长度的串，定义 $alt(w, x)$ 是把 w 和 x 交叉起来且以 w 开头所得到的串，即 $a_1 b_1 a_2 b_2 \cdots a_n b_n$. 如果 L 和 M 是语言，定义 $alt(L, M)$ 是所有形式为 $alt(w, x)$ 的串的集合，其中 w 是 L 中的任意串，而 x 是 M 中与 w 等长的任意串. 证明：如果 L 和 M 都是正则的，那么 $alt(L, M)$ 也是.

证明. 要证 $alt(L, M)$ 是正则的，只需证可以通过保留正则性的操作，将 L 和 M 转换为 $alt(L, M)$.

我们假设 L 和 M 的字母表 Σ_L 和 Σ_M 是不相交的. 如果它们不是不相交的，我们可以简单地对其中一种语言的字母表进行同态；然后，我们只需在最后做相应的逆同态.

首先，我们定义同态 h_1 如下：

$$\forall a \in \Sigma_L, h_1(a) = a$$

$$\forall b \in \Sigma_M, h_2(b) = \varepsilon$$

那么，${h_1}^{-1}(L)$ 是包含 L 中的字符串的语言，其中任意插入了 Σ_M中的符号.

类似地，我们定义同态 h_2 如下：

$$\forall b \in \Sigma_M, h_1(b) = b$$

$$\forall a \in \Sigma_L, h_2(a) = \varepsilon$$

因此，$alt(L, M)$ 可以表示为

$$alt(L, M) = {h_1}^{-1}(L) \cap {h_2}^{-1}(M) \cap (\Sigma_L \Sigma_M)^*$$

因此，$alt(L, M)$ 可以通过保留正则性的操作从 L 和 M 得到.

故如果 L 和 M 都是正则的，那么 $alt(L, M)$ 也是正则的. □

习题 4.2.8 设 L 是一个语言，定义 $half(L)$ 是所有 L 中串的前一半构成的集合，

即 $\left\{w \middle| \text{对于某个满足 } |x| = |w| \text{ 的 } x\text{，} wx \text{ 属于 } L\right\}$. 例如，如果 $L = \{\varepsilon, 0010, 011, 010110\}$，则 $half(L) = \{\varepsilon, 00, 010\}$. 注意，长度为奇数的串对于 $half(L)$ 没有贡献. 证明：如果 L 是正则的，那么 $half(L)$ 也是.

证明. 因为 L 是正则的，那么存在 DFA $A = (Q_A, \Sigma, \delta_A, q_0, F_A)$ 使得 $L(A) = L$. 对于 $half(L)$ 构造 DFA $B = (Q_B, \Sigma, \delta_B, [q_0, F_A], F_B)$ 如下：

- $Q_B =$

$$\left\{[q,S] \in Q_A \times P(Q_A) \middle| \forall w \in \Sigma^*, q = \hat{\delta}_A(q_0, w), S = \left\{p \in Q_A \middle| \hat{\delta}_A(p, w') \in F_A, w' \in \Sigma^{|w|}\right\}\right\}$$

- $\delta_B([q,S], a) = [\delta_A(q,a), T]$，其中 $a \in \Sigma$，$T = \{p \in Q_A | \forall b \in \Sigma, \delta_A(p,b) \in S\}$
- $F_B = \{[q,S] \in Q_B | q \in S\}$

以这种构造方式，我们得到：状态 $[q,S]$ 是某个串 w 可以让 A 从初始状态 q_0 转移到状态 q，并且 S 中的状态均可以在接收某个与 w 等长的串后转移至 A 的接受状态。

那么，F_B 中的所有元素都具有“DFA A 接收某个长度的串到达状态 q 后，可再次接收另一个等长的串到达接受状态”的性质。因此，$L(B) = half(L)$。

故如果 L 是正则的，那么 $half(L)$ 也是正则的。□

习题 4.2.9 我们把习题 4.2.8 推广到能够决定取走串中多大部分的一系列函数. 如果 f 是一个整数函数，定义 $f(L)$ 为 $\left\{w \middle| \text{对某个满足 } |x| = f(|w|) \text{ 的 } x\text{，} wx \text{ 属于 } L\right\}$. 例如，和运算 $half$ 对应的 f 是恒等函数 $f(n) = n$，因为 $half(L)$ 的定义中有 $|x| = |w|$. 证明：如果 L 是正则的，那么对于以下的 f，$f(L)$ 也是正则的：

a) $f(n) = 2n$（也就是取走串的前三分之一）.

b) $f(n) = n^2$（也就是取走的长度是没取走部分长度的平方根）.

c) $f(n) = 2^n$（也就是取走的长度是剩下长度的对数）.

证明. 因为 L 是正则的，那么存在 DFA $A = (Q_A, \Sigma, \delta_A, q_0, F_A)$ 使得 $L(A) = L$. 对于 $half(L)$ 构造 DFA $B = (Q_B, \Sigma, \delta_B, [q_0, F_A], F_B)$ 如下：

- $Q_B =$

$$\left\{[q,S] \in Q_A \times P(Q_A) \middle| \forall w \in \Sigma^*, q = \hat{\delta}_A(q_0, w), S = \left\{p \in Q_A \middle| \hat{\delta}_A(p, w') \in F_A, w' \in \Sigma^{f(|w|)}\right\}\right\}$$

- $\delta_B([q,S], a) = [\delta_A(q,a), T]$，其中 $a \in \Sigma$，$T = \{p \in Q_A | \forall b \in \Sigma, \delta_A(p,b) \in S\}$
- $F_B = \{[q,S] \in Q_B | q \in S\}$

以这种构造方式，我们得到：状态 $[q,S]$ 是某个串 w 可以让 A 从初始状态 q_0 转移到状态 q，并且 S 中的状态均可以在接收某个长度为 $f(|w|)$ 的串后转移至 A 的接受状态。

那么，F_B 中的所有元素都具有“DFA A 接收某个长度的串到达状态 q 后，可再次接收另一个长度为 $f(n)$ 的串到达接受状态”的性质。因此，$L(B)=f(L)$。

故如果 L 是正则的，那么 $f(L)$ 也是正则的。□

习题补充 1 给出如下的正则文法 G，求出对应的 DFA M，使得 $L(M)=L(G)$.

（1） $G_1=(V,T,P_1,S)$

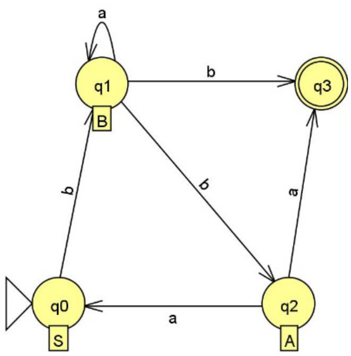
$P_1: S\to bB, B\to aB \mid bA \mid b, A\to a \mid aS$

（2） $G_2=(V,T,P_2,S)$

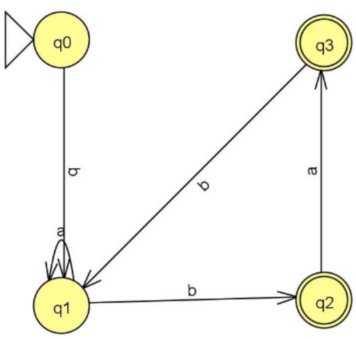
$P_2: S\to aS \mid bB \mid a, B\to bA \mid aB \mid aS$

解答.

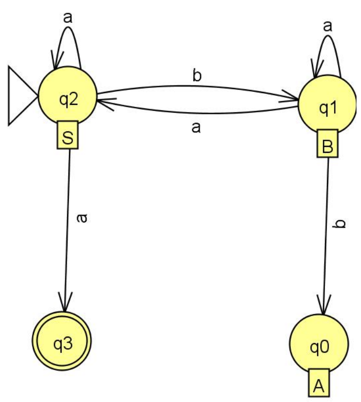
（1）文法 G_1 对应的 NFA M_1 如下图所示.



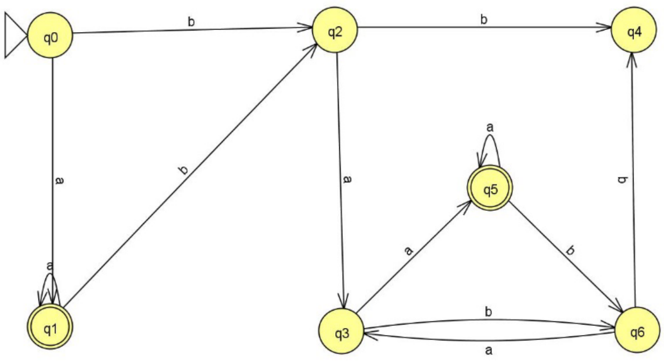
经确定化后的 DFA M_1' 如下图所示.



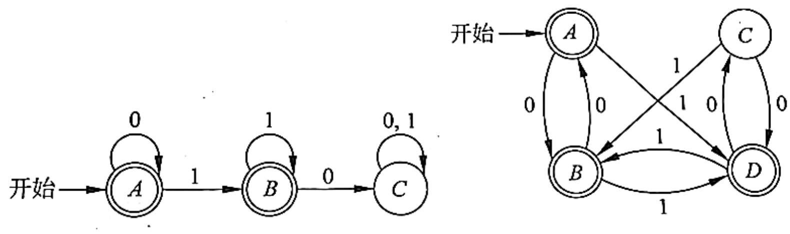
（2）文法 G_2 对应的 NFA M_2 如下图所示.



经确定化后的 DFA M_2' 如下图所示.



习题补充 2 给出下图描述的两个 DFA M, 分别求出对应的正则文法 G, 使得 $L(G) = L(M)$.



a) DFA M_1

b) DFA M_2

解答.

a) DFA M_1 对应的文法 $G_1 = (V, T, P_1, S)$

$$P_1: S \to 0S \mid 1A \mid \varepsilon$$

$$B \to 0B \mid 1B$$

$$A \to 1A \mid 0B \mid \varepsilon$$

b) DFA M_2 对应的文法 $G_2 = (V, T, P_2, S)$

$$P_2: S \to 1C \mid 0A \mid \varepsilon$$

$$A \to 0S \mid 1C \mid \varepsilon$$

$$B \to 0C \mid 0A$$

$$C \to 1A \mid 0B \mid \varepsilon$$

习题补充 3 Let $L_1 \subseteq \{0,1,2\}^*$ be a regular language, we can consider L_1 as a subset of integers under base 3, let L_2 be the corresponding set of L_1 over $\{0,1\}^*$ (i.e. under base 2), for example if $L_1 = \{11,12,121\}$, then $L_2 = \{100,101,10000\}$. Question: is L_2 a regular language?

Solution. L_2 **may or may not** be a regular language, depending on L_1. We illustrate it with cases a) and b).

a) Taking $L_1 = \{11\}, L_2 = \{100\}$ as an example, both L_1 and L_2 are obviously regular.

b) We use $[m]_2$ and $[m]_3$ to denote the decimal values corresponding to a string m consisting of symbols in $\{0,1\}$ or $\{0,1,2\}$ (i.e., an integer string under base 2 or 3), respectively.

Apparently, 2^* is a regular language, and $2^* \in L_1$. Thus

$$\forall k \in \mathbb{N}, 2^k \in L_1, [2^k] = 3^k - 1$$

Let $L = \{w \big| [w]_2 = 3^k - 1, k \in \mathbb{N}\}$. Now we are to prove that L is not a regular language.

Suppose language L is regular. Then the pumping lemma says that there is an n so that for all strings w with length $|w| \geq n$, there is a decomposition $w = xyz$ so that $|xy| \leq n, |y| \geq 0$, and xy^kz is also in this language for all $k \geq 0$.

Now, consider string w where $[w]_2 = 3^n - 1$. Then apparently $|w| \geq n$, so there is a decomposition $w = xyz$ that satisfies the criteria in the pumping lemma. Thus

$$[w]_2 = [x]_2 \cdot 2^{|y|+|z|} + [y]_2 \cdot 2^{|z|} + [z]_2 \tag{1}$$

Let $w' = xy^kz \in L, k \geq 0$, we have

$$[w']_2 = [x]_2 \cdot 2^{k|y|+|z|} + [y]_2 \cdot 2^{|z|} \cdot \sum_{i=0}^{k-1} 2^{i|y|} + [z]_2 \tag{2}$$

We conduct $(1) - (2)$, where

$$[w']_2 - [w]_2 = 2^{|y|+|z|} \cdot \frac{2^{(k-1)|y|} - 1}{2^{|y|} - 1} \cdot \left([x]_2 \cdot \left(2^{|y|} - 1\right) + [y]_2\right) \tag{3}$$

It is clear that $[w]_2 < [w']_2$ and $w, w' \in L$, so $[w']_2 - [w]_2$ must be a multiple of 3^n, i.e., $[w']_2 - [w]_2$ contains a factor 3^n. Based on the fact that $2^{|y|+|z|}$ does not contain factor 3, and that $[x]_2 \cdot \left(2^{|y|} - 1\right) + [y]_2 < [w]_2 < 3^n$, $[w']_2 - [w]_2$ does **not** contain factor 3^n, which produces a contradiction.

Therefore, L is not regular, then neither is L_2. □