Handout #3
Finite-state automata

First some standard stage-setting definitions:

(1) For any set Σ, we define Σ^* as the smallest set such that:
- $\epsilon \in \Sigma^*$, and
- if $x \in \Sigma$ and $u \in \Sigma^*$ then $(x{:}u) \in \Sigma^*$.

We often call Σ an *alphabet*, call the members of Σ *symbols*, and call the members of Σ^* *strings*.

(2) For any two strings $u \in \Sigma^*$ and $v \in \Sigma^*$, we define $u +\!\!+\, v$ as follows:
- $\epsilon +\!\!+\, v = v$
- $(x{:}w) +\!\!+\, v = x{:}(w +\!\!+\, v)$

Although these definitions provide the "official" notation, I'll sometimes be slightly lazy and abbreviate '$x{:}\epsilon$' as 'x', and abbreviate both '$s{:}t$' and '$s +\!\!+\, t$' as just 'st' in cases where it should be clear what's intended.

I'll generally use x, y and z for individual symbols of an alphabet Σ, and use u, v and w for strings in Σ^*. This should help to clarify whether a ':' or a '$+\!\!+$' has been left out.

And a reminder: if we have some set S, then the powerset of S, written $\mathcal{P}(S)$, is the set of all subsets of S. This is in a sense similar to, but different from, the set S^*. For example, if $S = \{\text{a}, \text{b}\}$, then:

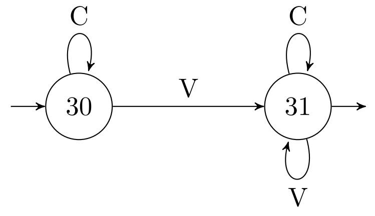
- $S^* = \{\epsilon, \text{a}, \text{b}, \text{aa}, \text{ab}, \text{ba}, \text{bb}, \text{aaa}, \text{aab}, \dots\}$
- $\mathcal{P}(S) = \{\{\}, \{\text{a}\}, \{\text{b}\}, \{\text{a}, \text{b}\}\}$

1 Finite-state automata, informally

Below, in (3) and (4), are graphical representations of two finite-state automata (FSAs). The circles represent *states*. The *initial* states are indicated by an "arrow from nowhere"; the *final* or *accepting* states are indicated by an "arrow to nowhere".

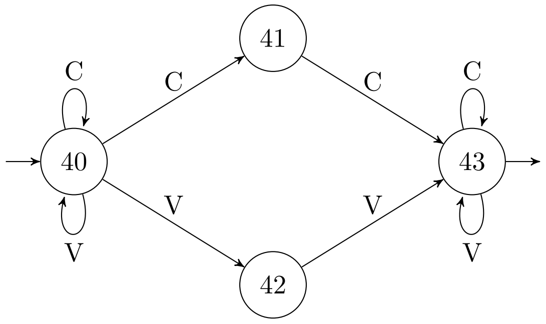
The FSA in (3) generates the subset of $\{\text{C}, \text{V}\}^*$ consisting of all and only strings that have at least one occurrence of 'V'.

(3)



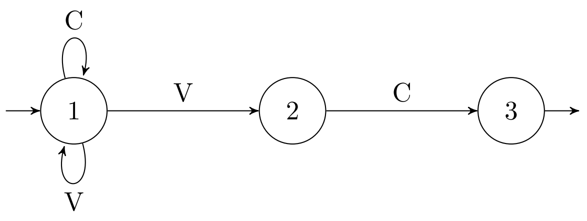
The FSA in (4) generates the subset of $\{\text{C}, \text{V}\}^*$ consisting of all and only strings that contain either two adjacent 'C's or two adjacent 'V's (or both).

(4)



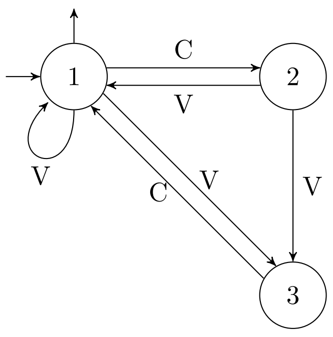
The FSA in (5) generates the subset of $\{\mathrm{C}, \mathrm{V}\}^*$ consisting of all and only strings which end in 'VC'.

(5)



If we think of state 1 as indicating syllable boundaries, then FSA in (6) generates sequences of syllables of the form '(C)V(C)'. The string 'VCV', for example, can be generated via two different paths, 1-1-2-1 and 1-3-1-1, corresponding to different syllabifications.

(6)



2 Formal definition of an FSA

(7) A finite-state automaton (FSA) is a five-tuple $(Q, \Sigma, I, F, \Delta)$ where:

- Q is a finite set of states;
- Σ, the alphabet, is a finite set of symbols;
- $I \subseteq Q$ is the set of initial states;
- $F \subseteq Q$ is the set of ending states; and
- $\Delta \subseteq Q \times \Sigma \times Q$ is the set of transitions.

So strictly speaking, (4) is a picture of the following mathematical object:

(8) $$\big(\ \{40,41,42,43\},\quad \{\mathrm{C},\mathrm{V}\},\quad \{40\},\quad \{43\},$$
$$\{(40,\mathrm{C},40),(40,\mathrm{C},41),(40,\mathrm{V},40),(40,\mathrm{V},42),(41,\mathrm{C},43),(42,\mathrm{V},43),(43,\mathrm{C},43),(43,\mathrm{V},43)\}\ \big)$$

You should convince yourself that (4) and (8) really do contain the same information.

Now let's try to say more precisely what it means for an automaton $M = (Q, \Sigma, I, F, \Delta)$ to generate/accept a string.

(9) For M to generate a string of three symbols, say $x_1x_2x_3$, there must be four states q_0, q_1, q_2, and q_3 such that
- $q_0 \in I$, and
- $(q_0, x_1, q_1) \in \Delta$, and
- $(q_1, x_2, q_2) \in \Delta$, and
- $(q_2, x_3, q_3) \in \Delta$, and
- $q_3 \in F$.

.

(10) More generally, M generates a string of n symbols, say $x_1x_2 \dots x_n$, iff: there are $n+1$ states q_0, q_1, q_2, $\dots q_n$ such that
- $q_0 \in I$, and
- for every $i \in \{1, 2, \dots, n\}$, $(q_{i-1}, x_i, q_i) \in \Delta$, and
- $q_n \in F$.

To take a concrete example:

(11) The automaton in (4)/(8) generates the string 'VCCVC' because we can choose q_0, q_1, q_2, q_3, q_4 and q_5 to be the states 40, 40, 41, 43, 43 and 43 (respectively), and then it's true that:
- $40 \in I$, and
- $(40, \text{V}, 40) \in \Delta$, and
- $(40, \text{C}, 41) \in \Delta$, and
- $(41, \text{C}, 43) \in \Delta$, and
- $(43, \text{V}, 43) \in \Delta$, and
- $(43, \text{C}, 43) \in \Delta$, and
- $43 \in F$.

Side remark: Note that abstractly, (10) is not all that different from:

(12) A tree-based grammar will generate a string $x_1x_2 \dots x_n$ iff: there is some collection of nonterminal symbols that we can choose such that
- those nonterminal symbols and the symbols x_1, x_2, etc. can all be clicked together into a tree structure in ways that the grammar allows, and
- the nonterminal "at the top" is the start symbol.

(Much more on this in a few weeks!)

We'll write $\mathcal{L}(M)$ for the set of strings generated by an FSA M. So stated roughly, the important idea is:

(13) $w \in \mathcal{L}(M)$

$$\iff \bigvee_{\text{all possible paths } p} \Big[\text{string } w \text{ can be generated by path } p\Big]$$

$$\iff \bigvee_{\text{all possible paths } p} \Big[\bigwedge_{\text{all steps } s \text{ in } p} \big[\text{step } s \text{ is allowed and generates the appropriate part of } w\big]\Big]$$

It's handy to write $I(q_0)$ in place of $q_0 \in I$, and likewise for F and Δ. Then one way to make (13) precise is:

(14) $x_1x_2 \dots x_n \in \mathcal{L}(M)$

$$\iff \bigvee_{q_0 \in Q} \bigvee_{q_1 \in Q} \cdots \bigvee_{q_{n-1} \in Q} \bigvee_{q_n \in Q} \Big[I(q_0) \wedge \Delta(q_0, x_1, q_1) \wedge \cdots \wedge \Delta(q_{n-1}, x_n, q_n) \wedge F(q_n)\Big]$$

But it's practical and enlightening to break this down in a couple of different ways.

2.1 Forward values

For any FSA M there's a two-place predicate fwd_M, relating states to strings in an important way:

(15) $\text{fwd}_M(w)(q)$ is true iff there's a path through M from some initial state to the state q, emitting the string w

Given a way to work out $\text{fwd}_M(w)(q)$ for any string and any state, we can easily use this to check for membership in $\mathcal{L}(M)$:

(16)
$$w \in \mathcal{L}(M) \iff \bigvee_{q_n \in Q} \Big[\text{fwd}_M(w)(q_n) \wedge F(q_n) \Big]$$

We can represent the predicate fwd_M in a table. Each column shows fwd_M values for the *entire prefix* consisting of the header symbols to its *left*. The first column shows values for the empty string.

(17) Here's the table of fwd_M values for prefixes of the string 'CVCCVVC' for the FSA in (5).

State		C	V	C	C	V	V	C
1	1	1	1	1	1	1	1	1
2	0	0	1	0	0	1	1	0
3	0	0	0	1	0	0	0	1

Notice that filling in the values in the leftmost column is easy: this column just says which states are initial states. And with a little bit of thought you should be able to convince yourself that, in order to fill in a column of this table, you only need to know:

- the values in the column immediately to its left, and
- the symbol immediately to its left.

More generally, this means that:

(18) The fwd_M values for a non-empty string $x_1 \dots x_n$ depend only on
- the fwd_M values for the string $x_1 \dots x_{n-1}$, and
- the symbol x_n.

This means that we can give a recursive definition of fwd_M:

(19)
$$\begin{aligned} \text{fwd}_M(\epsilon)(q) &= I(q) \\ \text{fwd}_M(x_1 \dots x_n)(q) &= \bigvee_{q_{n-1} \in Q} \Big[\text{fwd}_M(x_1 \dots x_{n-1})(q_{n-1}) \wedge \Delta(q_{n-1}, x_n, q) \Big] \end{aligned}$$

This suggests a natural and efficient algorithm for calculating these values: write out the table, start by filling in the leftmost column, and then fill in other columns from left to right. This is where the name "forward" comes from.

2.2 Backward values

We can do all the same things, flipped around in the other direction.

For any FSA M there's a two-place predicate bwd_M, relating states to strings in an important way:

(20) $\text{bwd}_M(w)(q)$ is true iff there's a path through M from the state q to some ending state, emitting the string w

Given a way to work out $\text{bwd}_M(w)(q)$ for any string and any state, we can easily use this to check for membership in $\mathcal{L}(M)$:

(21)
$$w \in \mathcal{L}(M) \iff \bigvee_{q_0 \in Q} \Big[I(q_0) \wedge \text{bwd}_M(w)(q_0) \Big]$$

We can represent the predicate bwd_M in a table. Each column shows bwd_M values for the *entire suffix* consisting of the header symbols to its *right*. The last column shows values for the empty string.

(22) Here's the table of bwd_M values for suffixes of the string 'CVCCVVC' for the FSA in (5).

State	C	V	C	C	V	V	C	
1	1	1	1	1	1	1	0	0
2	0	0	0	0	0	0	1	0
3	0	0	0	0	0	0	0	1

In this case, filling in the last column is easy, and each other column can be filled in simply by looking at the values immediately to its right.

(23) The bwd_M values for a non-empty string $x_1 \dots x_n$ depend only on
- the bwd_M values for the string $x_2 \dots x_n$, and
- the symbol x_1.

So bwd_M can also be defined recursively.

(24)
$$\text{bwd}_M(\epsilon)(q) = F(q)$$
$$\text{bwd}_M(x_1 \dots x_n)(q) = \bigvee_{q_1 \in Q} \Big[\Delta(q, x_1, q_1) \wedge \text{bwd}_M(x_2 \dots x_n)(q_1) \Big]$$

2.3 Forward values and backward values together

Now we can say something beautiful:

(25)
$$uv \in \mathcal{L}(M) \iff \bigvee_{q \in Q} \Big[\text{fwd}_M(u)(q) \wedge \text{bwd}_M(v)(q) \Big]$$

And in fact (16) and (21) are just special cases of (25), with u or v chosen to be the empty string:

(26)
$$w \in \mathcal{L}(M) \iff \bigvee_{q \in Q} \Big[\text{fwd}_M(w)(q) \wedge \text{bwd}_M(\epsilon)(q) \Big] \iff \bigvee_{q \in Q} \Big[\text{fwd}_M(w)(q) \wedge F(q) \Big]$$

(27)
$$w \in \mathcal{L}(M) \iff \bigvee_{q \in Q} \Big[\text{fwd}_M(\epsilon)(q) \wedge \text{bwd}_M(w)(q) \Big] \iff \bigvee_{q \in Q} \Big[I(q) \wedge \text{bwd}_M(w)(q) \Big]$$

3 Interchangeable subexpressions

Now forget about FSAs for a moment, and just consider sets of strings "out of the blue". We'll connect things back to FSAs shortly, in section 4.

(28) Given some stringset $L \subseteq \Sigma^*$, the L-remainders of a string u are all the strings v such that $uv \in L$. I'll write $\text{rem}_L(u)$ for the set of L-remainders of u, so we can write this definition in symbols as: $\text{rem}_L(u) = \{v \mid v \in \Sigma^*, uv \in L\}$.

This definition may seem unfamiliar and a bit awkward, but the underlying idea is quite simple and *very powerful* — you will probably recognize that it's been hiding somewhere inside your existing understanding of how grammars (of any sort!) work. Roughly, $\text{rem}_L(u)$ gives us a handle on "all the things we're still allowed to do, if we've done u so far". Some examples will clarify.

(29) If $L_1 = \{\text{cat}, \text{cap}, \text{cape}, \text{cut}, \text{cup}, \text{dog}\}$, then:
 a. $\text{rem}_{L_1}(\text{ca}) = \{\text{t}, \text{p}, \text{pe}\}$
 b. $\text{rem}_{L_1}(\text{c}) = \{\text{at}, \text{ap}, \text{ape}, \text{ut}, \text{up}\}$
 c. $\text{rem}_{L_1}(\text{cap}) = \{\epsilon, \text{e}\}$
 d. $\text{rem}_{L_1}(\text{d}) = \{\text{og}\}$

(30) If $L_2 = \{\text{ad}, \text{add}, \text{baa}, \text{bad}, \text{cab}, \text{cad}, \text{dab}, \text{dad}\}$, then:
 a. $\text{rem}_{L_2}(\text{c}) = \text{rem}_{L_2}(\text{d}) = \{\text{ab}, \text{ad}\}$
 b. $\text{rem}_{L_2}(\text{a}) = \{\text{d}, \text{dd}\}$
 c. $\text{rem}_{L_2}(\text{da}) = \{\text{b}, \text{d}\}$
 d. $\text{rem}_{L_2}(\text{ad}) = \{\epsilon, \text{d}\}$

When we notice that $\text{rem}_{L_2}(\text{c}) = \text{rem}_{L_2}(\text{d})$, this tells us something useful about how we can go about designing a grammar to generate the stringset L_2: such a grammar *doesn't need to care about* the distinction between starting with 'c' and starting with 'd', because for any string v that you choose, cv and dv will either both be in L_2 or both not be in L_2. An initial 'c' and an initial 'd' are *interchangeable subexpressions.*

(31) Given a stringset $L \subseteq \Sigma^*$ and two strings $u \in \Sigma^*$ and $v \in \Sigma^*$, we define a relation $\equiv_L$ such that: $u \equiv_L v$ iff $\text{rem}_L(u) = \text{rem}_L(v)$.

Some slightly more linguistics-ish examples:

(32) Suppose that $\Sigma = \{\text{C}, \text{V}\}$, and L is the subset of Σ^* containing all strings that contain at least one 'V'. Then:
 a. C $\equiv_L$ CC, because both can only be followed by strings that fulfill the requirement for a 'V'.
 b. VC $\equiv_L$ CV, because both can be followed by anything at all.
 c. So two strings are L-equivalent iff they either both do or both don't contain a 'V'.

(33) Suppose that $\Sigma = \{\text{C}, \text{V}\}$, and L is the subset of Σ^* containing all strings that have two adjacent 'C's or two adjacent 'V's (or both). Then
 a. C $\equiv_L$ CVC $\equiv_L$ CVCVC, because these all require remainders that have two adjacent 'C's *or* two adjacent 'V's *or* an initial 'C'.
 b. V $\equiv_L$ VCV $\equiv_L$ VCVCV, because these all require remainders that have two adjacent 'C's *or* two adjacent 'V's *or* an initial 'V'.
 c. CC $\equiv_L$ VCVCVVCVC

(34) Suppose that $\Sigma = \{\text{C}, \text{V}\}$, and L is the subset of Σ^* containing all strings that *do not* have two adjacent occurrences of 'V'. Then:

a. CCCC $\equiv_L$ VC, because both can be followed by anything without adjacent 'V's.
b. CCV $\equiv_L$ V, because both can be followed by anything without adjacent 'V's that does not begin with 'V'.
c. CCV $\not\equiv_L$ CCC, because only the latter can be followed by 'VC'.
d. In fact: two strings are L-equivalent iff they end with the same symbol!

(35) Suppose that Σ is the set of English words, and L is the set of all grammatical English word-sequences. Then (probably?):
a. John $\equiv_L$ the brown furry rabbit
b. John $\equiv_L$ Mary thinks that John
c. John $\not\equiv_L$ the fact that John

4 The Myhill-Nerode Theorem

We can connect this idea of equivalent subexpressions back to forward values in an FSA. First one quick helper definition:

(36) For any FSA $M = (Q, \Sigma, I, F, \Delta)$ and any string $u \in \Sigma^*$, we define the *forward set* of u in M to be the set of states $\text{fwdSet}_M(u) = \{q \mid q \in Q, \text{fwd}_M(u)(q)\}$.

Now here's the important connection:

(37) For any FSA $M = (Q, \Sigma, I, F, \Delta)$ and for any two strings $u \in \Sigma^*$ and $v \in \Sigma^*$, if $\text{fwdSet}_M(u) = \text{fwdSet}_M(v)$ then $u \equiv_{\mathcal{L}(M)} v$.

Given any particular stringset L, we can think of the relation $\equiv_L$ as sorting out all possible strings into buckets (or "equivalence classes"): two strings belong in the same bucket iff they are equivalent prefixes. So what (37) says is that for an FSA to generate L it must be arranged so that fwdSet only maps two strings to the same state-sets if those two strings are equivalent prefixes; the machine can ignore distinctions between bucket-mates, but only between bucket-mates.

This idea of ignoring at least some distinctions is exactly what makes a grammar different from a list of strings. *The challenge in writing grammars is always about ignoring the "irrelevant" distinctions in order to allow creativity, while tracking the "relevant" distinctions.*[1]

And now we can put our finger on the capacities/limitations of finite-state automata.

(38) **The Myhill-Nerode Theorem:** Given a particular stringset L, there is an FSA that generates L iff the relation $\equiv_L$ sorts strings into only finitely-many buckets.

Why is this, exactly?

4.1 Why do FSAs make only finitely-many distinctions?

Well, if we have a particular FSA whose set of states is Q, then there are only finitely many distinct subsets of Q that fwdSet_M can map strings to; specifically, there are $2^{|Q|}$ of them. So there are only finitely-many "candidate forward sets", meaning that the FSA is necessarily making only those finitely-many distinctions.

[1] McCulloch & Pitts (1943, pp.130–131) put this nicely in their classic paper: "our knowledge of the world including ourselves, is incomplete . . . This ignorance, implicit in all our brains, is the counterpart of abstraction which renders our knowledge useful." See also chapters 2–4 of Minsky's book *Computation: Finite and Infinite Machines* (1967) for good discussion of this very general point.

Having noticed this, it's very easy to convince ourselves that no FSA can generate the stringset $L = \{a^n b^n \mid n > 0\}$. Notice that a $\not\equiv_L$ aa, and aa $\not\equiv_L$ aaa, and so on. In fact any string of 'a's is non-equivalent to each of the different-length strings of 'a's, so this stringset sorts strings into infinitely-many buckets, one bucket for each length. There is no way for an FSA M to be set up such that $\text{fwdSet}_M(a^j) \neq \text{fwdSet}_M(a^k)$ whenever $j \neq k$; any FSA will incorrectly collapse the distinction between two such strings of 'a's.

4.2 Why can any finitely-many distinctions be captured with an FSA?

On the other hand, if we have a particular stringset L whose equivalence relation $\equiv_L$ makes only finitely-many distinctions, then there is a straightforward way to construct a minimal FSA whose states track exactly those distinctions.

Consider again the stringset from (33), consisting of all strings with either two adjacent 'C's or two adjacent 'V's (or both). This stringset's equivalence relation sorts strings into four buckets:

(39)
a. a bucket containing strings that have either two adjacent 'C's or two adjacent 'V's;
b. a bucket containing strings that don't have two adjacent 'C's or 'V's, but end in 'C';
c. a bucket containing strings that don't have two adjacent 'C's or 'V's, but end in 'V';
d. a bucket containing only the empty string.

Having noticed this we can mechanically construct an appropriate FSA — known as the *minimal FSA* for this stringset — which has one state corresponding to each bucket. The crucial idea here is that if $u \equiv_L v$, then $ux \equiv_L vx$ for any $x \in \Sigma$, i.e. adding a symbol at the end can't "break" an equivalence.

(40)
$$\begin{tikzcd}
 & & \text{(39b)} \arrow[dr, "C"] \arrow[dd, bend left, "V"] & & \\
{} \arrow[r] & \text{(39d)} \arrow[ur, "C"] \arrow[dr, "V"'] & & \text{(39a)} \arrow[loop above, "C"] \arrow[loop below, "V"] \arrow[r] & {} \\
 & & \text{(39c)} \arrow[ur, "V"'] \arrow[uu, bend left, "C"] & &
\end{tikzcd}$$

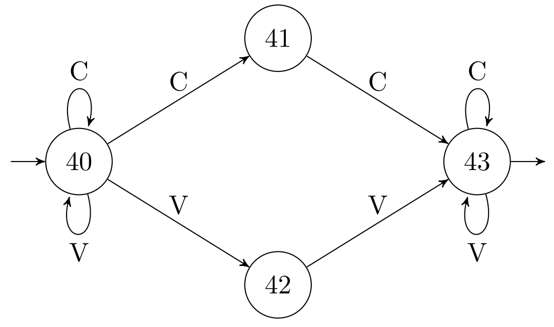
This strategy produces a specific kind of automaton, a *deterministic automaton.* In a deterministic automaton, there are never two arcs leading out of the same state that are labeled with the same symbol; each string corresponds to at most one path through the states, and so fwdSet only ever produces singleton sets or the empty set.

5 One last loose end: Determinization and minimization

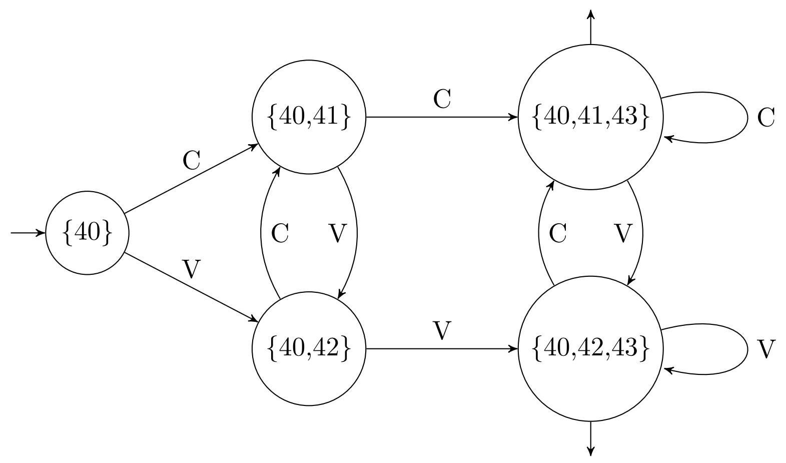
Any FSA M can be converted into an equivalent deterministic one M', by setting up the states of M' to correspond to sets of the states of M. Then, for any string u, the one state in $\text{fwdSet}_{M'}(u)$ will be the one corresponding to the set $\text{fwdSet}_M(u)$.

Applying this procedure to the FSA in (4) (repeated below), which generates the stringset in (33), produces the new FSA in (41).

(4)



(41)



It turns out that the two states $\{40, 41, 43\}$ and $\{40, 42, 43\}$ here “do the same thing”: in either case, it's possible to end, it's possible to transition to state $\{40, 41, 43\}$ on a ‘C’, and it's possible to transition to state $\{40, 42, 43\}$ on a ‘V’ (and that's all). Collapsing these two states will produce the minimal FSA in (40).