On the State Minimization of Nondeterministic Finite Automata

TSUNEHIKO KAMEDA, MEMBER, IEEE, AND PETER WEINER, MEMBER, IEEE

Abstract—**The aim of this paper is to obtain a procedure for finding a minimum state nondeterministic finite automaton (NDA) equivalent to a given (in general, nondeterministic) finite automaton. Given a finite automaton $\mathscr{A}$, we derive from $\mathscr{A}$ a matrix of 1's and 0's, called a *reduced automaton matrix* RAM) of $\mathscr{A}$, in a certain way and show that each state of $\mathscr{A}$ corresponds to a *grid* over the RAM. A grid consists of a set of rows and a set of columns of an RAM such that only 1's appear at the intersections. It is also shown that the union of all the grids, each of which corresponds to a state of $\mathscr{A}$, covers all the 1 entries of an RAM.**

We then reverse this analysis in order to "synthesize" a minimum state NDA. First we construct an RAM of a given finite automaton; then we find a set of grids that covers all and only its 1 entries. With the help of the *intersection rule* described in the text, we construct an NDA whose states correspond to the grids in the set.

Unfortunately, this construction does not always yield an NDA which is equivalent to the given automaton. But it is shown that at least one of the NDA's thus constructed is equivalent to the given automaton. Therefore we present a minimization procedure which includes searching.

Index Terms—**Equivalence of automata, finite automata theory, nondeterministic automata, reduction of automata, regular events.**

INTRODUCTION

THE CONCEPT of a nondeterministic (finite) automaton (NDA) was formulated by Rabin and Scott in their classical paper published in 1959 [1]. Recently it has found an extensive use in connection with the study of formal languages. Also, some attempts have been made to investigate and understand properties of NDA's [2]–[6]. Because of their finiteness, most of the problems concerning finite automata are easy to solve, but there are a few problems which have resisted every attempt to solve—a notable example being the problem of minimum star height [11]. Another problem which has been open to date is to find a nonexhaustive procedure for finding a *minimum state NDA* equivalent to a given NDA. The first published attempt was made by Ott and Feinstein [2]. In this paper we attack the latter open problem.

Let N be the number of states of a given finite automaton. Any minimum state NDA equivalent to the given automaton has at most N states. It is clear that there are only finite number of NDA's which have at most N states. Therefore we can always resort to the exhaustive procedure. Namely, we can construct every NDA that has at most N states and pick one with the fewest states and equivalent to the given automaton. But this procedure is too lengthy to be practical. Thus our problem is to find a procedure which yields the result with a reasonable amount of work.

For an NDA there is no such thing as "canonical" NDA corresponding to the reduced DA which exists uniquely for any deterministic finite automaton (DA). However, we can look at our problem from the following viewpoint. Given an NDA we can always find a DA which is equivalent to it using the *subset construction*. Once an equivalent DA is constructed, we can reduce it to get a reduced DA which is equivalent to the given NDA. Therefore even with an NDA we can uniquely associate a reduced DA equivalent to it. Thus we have standardized NDA's in a way, and our problem now is how to find a minimum state NDA equivalent to a given reduced DA. This is the approach we are going to take in this paper.

The paper consists of two parts. In Part 1, we present some basic tools for analysis of finite automata, which are relevant to our later discussion, but quite general in nature. In Part 2, we attack the problem of state minimization. The reader is cautioned beforehand that the reduction of an NDA in the sense of [4] and [6] does not mean state minimization in our sense. According to *their* definition, an NDA is reduced if no two states are equivalent.

In this paper, by a *minimum NDA* we mean a minimum state NDA within a behavior-equivalence class.

PART 1: ANALYSIS

I. PRELIMINARIES

In the following we fix the input alphabet to be Σ. Thus Σ^* (the free monoid generated by Σ with unit e) denotes all finite input sequences, including the null sequence e of length 0. Any subset of Σ^* is called an *event*. We denote the *empty set* by $\varnothing$.

A (*finite*) *automaton* $\mathscr{A}$ over the input alphabet Σ is a quadruple

$$\mathscr{A} = (S, M, S_0, F)$$

where $S = \{s_1, s_2, \cdots, s_n\}$ $(\neq \varnothing)$ is the set of *states*, $M: S \times \Sigma \to 2^S$ (the set of all subsets of S) is the *transition function*, $S_0 (\neq \varnothing) \subseteq S$ is the set of *initial states*, and $F \subseteq S$ is the set of *final states*. If $S_0 = \{s_0\}$, and if $|M(s, \sigma)| = 1$ for $\forall s \in S$ and $\forall \sigma \in \Sigma$, then the automaton $\mathscr{A}$ is said to be a *deterministic automaton* (DA).[1] Otherwise $\mathscr{A}$ is said to be a *nondeterministic automaton* (NDA).

The domain of the transition function M may be extended

Manuscript received December 22, 1967; revised January 3, 1969 and October 23, 1969.

T. Kameda is with the Department of Electrical Engineering, University of Waterloo, Waterloo, Ontario, Canada.

P. Weiner is with the Department of Computer Science, Yale University, New Haven, Conn.

[1] Here | | denotes the size of the set inside. A DA is sometimes defined with the condition $|M(s, \sigma)| \leq 1$. But any such DA can be converted to a DA of our definition by introducing a "dead" state (nonfinal state which has no outgoing transition except to itself).

from $S \times \Sigma$ to $S \times \Sigma^*$ by a recursive definition. For $\forall s \in S$, $\forall \sigma \in \Sigma$, and $\forall x \in \Sigma^*$,

$$M(s, e) = s$$
$$M(s, x\sigma) = \bigcup_{s' \in M(s,x)} M(s', \sigma).$$

It is convenient to further extend the domain of M to $2^S \times \Sigma^*$ by

$$M(R, x) = \bigcup_{s \in R} M(s, x) \qquad \text{for } \forall R \in 2^S \text{ and } \forall x \in \Sigma^*.$$

In the following we shall assume, without mentioning it, that M is extended as above whenever necessary.

An input sequence $x \in \Sigma^*$ is *accepted* by $\mathscr{A}$, iff $M(S_0, x) \cap F \neq \varnothing$. The set

$$bh(\mathscr{A}) = \{x \in \Sigma^* | M(S_0, x) \cap F \neq \varnothing\}$$

is called the *behavior of* (or the *event recognized by*) $\mathscr{A}$. The *dual* of an automaton $\mathscr{A} = (S, M, S_0, F)$ is defined as

$$\overleftarrow{\mathscr{A}} = (S, \overleftarrow{M}, F, S_0),$$

where $\overleftarrow{M}$ is such that for $\forall \sigma \in \Sigma$, $\forall s_i$, and $\forall s_j \in S$,

$$s_i \in \overleftarrow{M}(s_j, \sigma) \Leftrightarrow s_j \in M(s_i, \sigma).$$

It is obvious that $\overleftarrow{\overleftarrow{\mathscr{A}}} = \mathscr{A}$.

The reverse of a sequence $x = \sigma_1 \sigma_2 \cdots \sigma_k$, $k \geq 1$ and $\sigma_i \in \Sigma$, is $\bar{x} = \sigma_k \cdots \sigma_2 \sigma_1$. The reverse of the null sequence e is defined to be itself, i.e., $\bar{e} = e$. The *reverse* of an event E is defined as

$$\overleftarrow{E} = \{\bar{x} | x \in E\}.$$

Note that for two events E and E', $E = E' \Leftrightarrow \overleftarrow{E} = \overleftarrow{E'}$.

In the sequel we assume that $(\forall s \in S)(\exists x \in \Sigma^*)[s \in M(S_0, x)]$. We also assume $F \neq \varnothing$, since otherwise $bh(\mathscr{A}) = \varnothing$ and all the problems concerning the behavior of such $\mathscr{A}$ are trivial.

II. Subset Construction

In this section we introduce an operation called the *subset construction*,[2] and show its rather interesting properties.

Definition 1: Given an automaton $\mathscr{A} = (S, M, S_0, F)$, the *succeeding event* and *preceding event* of a state s_i of $\mathscr{A}$, $\mathrm{sc}_{\mathscr{A}}(s_i)$ and $\mathrm{pr}_{\mathscr{A}}(s_i)$, respectively, are defined by[3]

$$\mathrm{sc}_{\mathscr{A}}(s_i) = bh((S, M, s_i, F))$$
$$\mathrm{pr}_{\mathscr{A}}(s_i) = bh((S, M, S_0, s_i)).$$

Therefore $\mathrm{sc}_{\mathscr{A}}(s_i) = \{x \in \Sigma^* | M(s_i, x) \cap F \neq \varnothing\}$ and $\mathrm{pr}_{\mathscr{A}}(s_i) = \{x \in \Sigma^* | s_i \in M(S_0, x)\}$. Two states of $\mathscr{A}$, s_i and s_j, are said to be *equivalent*, written $s_i \sim s_j$, iff $\mathrm{sc}_{\mathscr{A}}(s_i) = \mathrm{sc}_{\mathscr{A}}(s_j)$. A DA is *reduced* iff no two distinct states are equivalent.

Clearly, if $\mathscr{A}$ is a DA, then for all $i \neq j$, $\mathrm{pr}_{\mathscr{A}}(s_i) \cap \mathrm{pr}_{\mathscr{A}}(s_j) = \varnothing$.

Definition 2: Given an automaton $\mathscr{A} = (S, M, S_0, F)$, we define the *subset DA associated with* $\mathscr{A}$ to be $D(\mathscr{A}) = (P, M', p_0, F')$, where

$$p_0 = S_0$$
$$P = \{M(S_0, x) | x \in \Sigma^*\} = \{p_1, p_2, \cdots, p_m\}$$
$$F' = \{p \in P | p \cap F \neq \varnothing\}$$
$$M'(p_i, \sigma) = p_j \Leftrightarrow M(p_i, \sigma) = p_j \quad \text{for all } i, j, \text{ and } \forall \sigma \in \Sigma.$$

This construction is called the *subset construction.* We assume $p_i \neq p_j$ for $i \neq j$. If $p_i \in P$ does not contain any state of $\mathscr{A}$, then p_i is called an *empty state.*

Lemma 1: Let $\mathscr{A}$ be as defined in Definition 2 and let $\mathscr{B} = D(\mathscr{A})$. Given $\forall p \in P$, we have the following relations:

1) $\mathrm{sc}_{\mathscr{B}}(p) = \bigcup_{s \in p} \mathrm{sc}_{\mathscr{A}}(s)$
2) $\mathrm{pr}_{\mathscr{B}}(p) \subseteq \bigcap_{s \in p} \mathrm{pr}_{\mathscr{A}}(s)$
3) $\mathrm{pr}_{\mathscr{B}}(p) \cap \mathrm{pr}_{\mathscr{A}}(s) = \varnothing$ for $\forall s \in S - p$.

Proof:

1) $$\begin{aligned} \mathrm{sc}_{\mathscr{B}}(p) &= \{x | M'(p, x) \cap F \neq \varnothing\} \\ &= \{x | \bigcup_{s \in p} M(s, x) \cap F \neq \varnothing\} \\ &= \bigcup_{s \in p} \{x | M(s, x) \cap F \neq \varnothing\} \\ &= \bigcup_{s \in p} \mathrm{sc}_{\mathscr{A}}(s). \end{aligned}$$

2) $$\begin{aligned} \mathrm{pr}_{\mathscr{B}}(p) &= \{x | p = M'(p_0, x)\} \\ &= \{x | p = M(S_0, x)\} \\ &\subseteq \{x | p \subseteq M(S_0, x)\} \\ &= \{x | \bigwedge_{s \in p} [s \in M(S_0, x)]\} \\ &= \bigcap_{s \in p} \{x | s \in M(S_0, x)\} \\ &= \bigcap_{s \in p} \mathrm{pr}_{\mathscr{A}}(s). \end{aligned}$$

3) $(\forall s \notin p)[x \in \mathrm{pr}_{\mathscr{B}}(p) \Rightarrow s \notin M(S_0, x)]$.

But

$$s \notin M(S_0, x) \Leftrightarrow x \notin \mathrm{pr}_{\mathscr{A}}(s).$$

Therefore

$$(\forall s \notin p)[x \in \mathrm{pr}_{\mathscr{B}}(p) \Rightarrow x \notin \mathrm{pr}_{\mathscr{A}}(s)]. \qquad \text{Q.E.D.}$$

If we let $p = p_0 = S_0$ in part 1 of Lemma 1, we get the following.

Corollary (Rabin and Scott [1]): $bh(D(\mathscr{A})) = bh(\mathscr{A})$.

Lemma 2:

1) $\mathrm{sc}_{\mathscr{A}}(s) = \overleftarrow{\mathrm{pr}}_{\overleftarrow{\mathscr{A}}}(s)$
2) $\mathrm{pr}_{\mathscr{A}}(s) = \overleftarrow{\mathrm{sc}}_{\overleftarrow{\mathscr{A}}}(s)$.

Proof: Easy and left to the reader.

Corollary (Rabin and Scott [1]): $bh(\mathscr{A}) = \overleftarrow{bh}(\overleftarrow{\mathscr{A}})$.

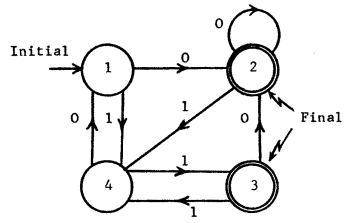
We now present an interesting but little known theorem, which was discovered by Brzozowski [8].

Theorem 1 (Brzozowski): Let $\mathscr{A} = (S, M, S_0, F)$ be a DA, not necessarily reduced. $D(\overleftarrow{\mathscr{A}})$ is a reduced DA equivalent $\mathscr{A}$.

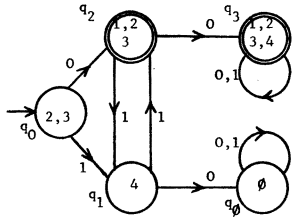
Proof: The fact that $\mathscr{C} = D(\overleftarrow{\mathscr{A}})$ is equivalent to $\mathscr{A}$ follows from the first corollary stated above. Now suppose that

[2] The operation was first used by Rabin and Scott [1] to construct a DA equivalent to a given NDA.

[3] The succeeding event and preceding event have been called by Brzozowski [8] the *derivative* and *predecessor*, respectively.



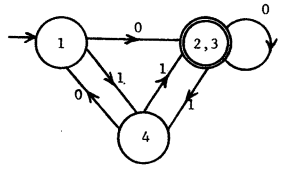
(a)



(b)

	q_0	q_1	q_2	q_3
1	0	0	1	1
2	1	0	1	1
3	1	0	1	1
4	0	1	0	1

(c)



(d)

Fig. 1. Reduction of a DA. (a) Given DA, $\mathscr{H}$. (b) $D(\overleftarrow{\mathscr{H}})$. (c) SCM of $\mathscr{H}$. (d) $\hat{\mathscr{H}}$.

two distinct states q_i and q_j of $\mathscr{C}$ were equivalent. By the hypothesis,

$$\mathrm{sc}_{\mathscr{C}}(q_i) = \mathrm{sc}_{\mathscr{C}}(q_j),$$

which implies (by Lemma 1) that

$$\bigcup_{s\in q_i} \mathrm{sc}^{\leftarrow}_{\mathscr{A}}(s) = \bigcup_{s\in q_j} \mathrm{sc}^{\leftarrow}_{\mathscr{A}}(s).$$

Using Lemma 2, we get

$$\bigcup_{s\in q_i} \overleftarrow{\mathrm{pr}}_{\mathscr{A}}(s) = \bigcup_{s\in q_j} \overleftarrow{\mathrm{pr}}_{\mathscr{A}}(s)$$

$$\overleftarrow{\bigcup_{s\in q_i} \mathrm{pr}_{\mathscr{A}}(s)} = \overleftarrow{\bigcup_{s\in q_j} \mathrm{pr}_{\mathscr{A}}(s)}$$

$$\bigcup_{s\in q_i} \mathrm{pr}_{\mathscr{A}}(s) = \bigcup_{s\in q_j} \mathrm{pr}_{\mathscr{A}}(s).$$

The last equality says that the union of the (nonempty) events in a subset of the set $\{\mathrm{pr}_{\mathscr{A}}(s)|s\in S\}$ is identical to the union of the events in another subset distinct from the first subset. This is a contradiction to the fact that the preceding events of a DA are mutually disjoint. Q.E.D.

Theorem 1 is significant in its own right and has some interesting applications, which are discussed in Kameda and Weiner [7].

III. State Composition Matrix and its Properties

In this section we shall investigate the relation among the succeeding events of the states of a finite automaton. For this purpose we introduce the concept of the *state composition matrix* and investigate its properties.

Definition 3: Given an automaton $\mathscr{A}=(S, M, S_0, F)$, let $\mathscr{C}=D(\overleftarrow{\mathscr{A}})=(Q, M'', q_0, F'')$. A *state composition matrix* (*SCM*) of $\mathscr{A}$ contains a row for each state $s_i\in S$ and a column for each nonempty state (refer to Definition 2) $q_i\in Q$, with the (i, j) element

$$c_{ij} = \begin{cases} 1 & \text{if } s_i\in q_j \\ 0 & \text{otherwise.} \end{cases}$$

A row of an SCM *covers* another row if the former has a 1 in every column in which the latter has a 1.

Note that an SCM of $\mathscr{A}$ is unique within permutation of rows and columns.

Example: If we apply the subset construction to the dual $\overleftarrow{\mathscr{H}}$ of the DA $\mathscr{H}$, shown in Fig. 1(a), we get the reduced DA shown in Fig. 1(b) (Theorem 1). An SCM of $\mathscr{H}$ is shown in Fig. 1(c). It is seen that rows 2 and 3 cover row 1.

Definition 4: Let $\mathscr{A}$ and $\mathscr{C}$ be as defined in Definition 3. A *characteristic event* of $\mathscr{A}$ is defined for each nonempty state $q_i\in Q$ as

$$\mathrm{ch}^{\mathscr{A}}_i = \overleftarrow{\mathrm{pr}}_{\mathscr{C}}(q_i).$$

Since $\mathscr{C}$ is a DA, we have (see the note after Definition 1)

$$\mathrm{ch}^{\mathscr{A}}_i \cap \mathrm{ch}^{\mathscr{A}}_j = \varnothing \qquad \text{for } i \neq j.$$

Thus $\mathrm{ch}^{\mathscr{A}}_i$, $i=1, 2, \cdots, m$ and $\Sigma^* - \mathrm{ch}(\mathscr{A})$ together form a partition of Σ^*, where $\mathrm{ch}(\mathscr{A}) = \bigcup_{i=1}^{m} \mathrm{ch}^{\mathscr{A}}_i$.

Theorem 2:

$$\mathrm{sc}_{\mathscr{A}}(s_i) = \bigcup_{j|c_{ij}=1} \mathrm{ch}^{\mathscr{A}}_j.$$

Proof: Let $\mathscr{A}_i=(S, M, s_i, F)$. From the two corollaries in Section II, it follows that

$$bh(\mathscr{A}_i) = \overleftarrow{bh}(D(\overleftarrow{\mathscr{A}_i})).$$

Let $\mathscr{C}$, Q, M'', and q_0 be as defined above. Then

$$D(\overleftarrow{\mathscr{A}_i}) = (Q, M'', q_0, G),$$

where

$$G = \{q_j\in Q|s_i\in q_j\} = \{q_j\in Q|c_{ij}=1\}.$$

Therefore

$$bh(D(\overleftarrow{\mathscr{A}_i})) = \bigcup_{q_j\in G} \mathrm{pr}_{D(\overleftarrow{\mathscr{A}_i})}(q_j) = \bigcup_{j|c_{ij}=1} \mathrm{pr}_{\mathscr{C}}(q_j).$$

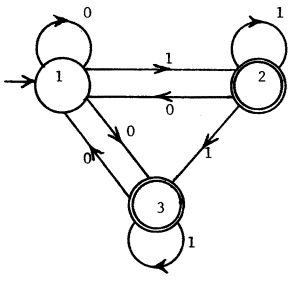
Thus it follows that

$$\mathrm{sc}_{\mathscr{A}}(s_i) = bh(\mathscr{A}_i) = \bigcup_{j|c_{ij}=1} \overleftarrow{\mathrm{pr}}_{\mathscr{C}}(q_j) = \bigcup_{j|c_{ij}=1} \mathrm{ch}^{\mathscr{A}}_j.$$ Q.E.D.

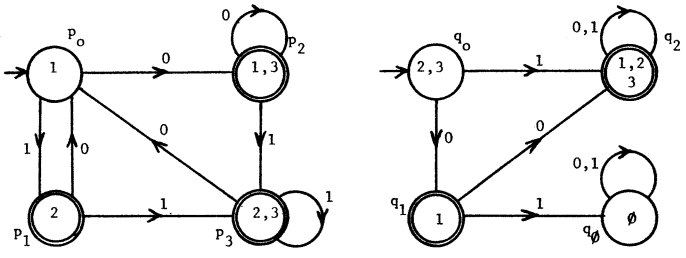
Corollary: Two states of an automaton are equivalent iff the corresponding rows in an SCM of the automaton have an identical pattern of 1's and 0's.

Example: The above corollary is useful for the reduction of a DA. It is seen that states 2 and 3 of the DA $\mathscr{H}$ of Fig. 1(a) are equivalent, since the corresponding rows in the SCM shown in Fig. 1(c) have an identical pattern of 1's and 0's. Thus the two states can be merged, yielding the reduced DA $\hat{\mathscr{H}}$ equivalent to $\mathscr{H}$ (Fig. 1(d)).

Definition 5: Let $\mathscr{A}$ and $\mathscr{A}'$ be two NDA's. A *configuration* of $\mathscr{A}$ is a subset of the states of $\mathscr{A}$. The configurations C and C' of $\mathscr{A}$ and $\mathscr{A}'$, respectively, are *equivalent*, written $C \sim C'$, iff $\bigcup_{s\in C} \mathrm{sc}_{\mathscr{A}}(s) = \bigcup_{s'\in C'} \mathrm{sc}_{\mathscr{A}'}(s')$.



(a)



(b) (c)

Fig. 2. Example for illustrating various concepts. (a) Given NDA, $\mathscr{A}$. (b) $\mathscr{B}=D(\mathscr{A})$. (c) $\mathscr{C}=D(\overleftarrow{\mathscr{A}})$.

	q_0	q_1	q_2
p_0		1	1
p_1	2		2
p_2	3	1	1,3
p_3	2,3		2,3

(a)

	q_0	q_1	q_2
p_0		1	1
p_1	1		1
p_2	1	1	1
p_3	1		1

(b)

Fig. 3. SM and EAM of $\mathscr{A}$. (a) SM of $\mathscr{A}$. (b) EAM of $\mathscr{A}$.

Theorem 2 says that a state s_i of $\mathscr{A}$ is equivalent to the configuration $\{q_j|c_{ij}=1\}$ of $\overleftarrow{\mathscr{C}}$.

Corollary to Theorem 2: $\mathrm{sc}_{\mathscr{A}}(s_i)\supseteq\mathrm{sc}_{\mathscr{A}}(s_j)$, iff the row (of an SCM of $\mathscr{A}$) corresponding to s_i covers the row corresponding to s_j.

Corollary to Theorem 2: Any state of an NDA, whose row in the SCM is the union[4] of some other rows, can be elimnated without changing the behavior of the NDA.

Proof: Let $\{s\}\sim C$. Modify the transition function M of $\mathscr{A}$ as follows: $M'(s', \sigma)=M(s', \sigma)\cup C-\{s\}$ for $\forall\sigma\in\Sigma$ and $\forall s'|s\in M(s', \sigma)$.

It is clear that for no $\{s'\}$ such that $s\in M(s', \sigma)$, $\mathrm{sc}_{\mathscr{A}}(s')$ is affected by this modification. Thus s can be eliminated without changing the behavior of the NDA. Q.E.D.

Definition 6: An NDA is called *irreducible* iff it has no state such that the state is equivalent to any other configuration.

It should be noted that there may be an NDA which is equivalent to an irreducible NDA but has fewer states. Therefore an irreducible NDA is, in general, different from a minimum state NDA. In fact, finding a minimum state NDA rather than an irreducible NDA is the objective of this paper.

We mention here that given any matrix of 1's and 0's, one can construct a DA which has an SCM identical to the given matrix [7].

IV. Reduced Automaton Matrix

The main objective of this section is to introduce a few matrices which are derived from an automaton. Throughout the section, we let $\mathscr{A}=(S, M, S_0, F)$ be a given automaton, $\mathscr{B}=D(\mathscr{A})=(P, M', p_0, F')$, and $\mathscr{C}=D(\overleftarrow{\mathscr{A}})=(Q, M'', q_0, F'')$. Also we let $s_k\in S$, $p_i\in P$, and $q_j\in Q$.

[4] Union is taken component-wise. The union operation is defined by $0\cup 0=0$ and $0\cup 1=1\cup 0=1\cup 1=1$.

Definition 7: A *states map* (*SM*) of $\mathscr{A}$ contains a row for each nonempty state of $\mathscr{B}$, and a column for each nonempty state of $\mathscr{C}$. The (i, j) entry contains $p_i\cap q_j$, or is blank if $p_i\cap q_j=\emptyset$. An *elementary automaton matrix* (*EAM*) of $\mathscr{A}$ is obtained from an SM of $\mathscr{A}$ by replacing each nonblank entry by a 1. Its (i, j) element is denoted by e_{ij}.

Example: We construct an SM and EAM of the NDA $\mathscr{A}$ of Fig. 2(a). Since we have $D(\mathscr{A})$ and $D(\overleftarrow{\mathscr{A}})$ available in Fig. 2(b) and (c), respectively, we simply follow Definition 7 to get an SM and EAM of $\mathscr{A}$, shown in Fig. 3(a) and (b), respectively.

Theorem 3:

$$\mathrm{sc}_{\mathscr{B}}(p_i)=\bigcup_{j|e_{ij}=1}\mathrm{ch}_j^{\mathscr{A}}.$$

Proof: By Theorem 2,

$$\mathrm{sc}_{\mathscr{A}}(s_k)=\bigcup_{j|c_{kj}=1}\mathrm{ch}_j^{\mathscr{A}}.$$

But by part 1 of Lemma 1,

$$\mathrm{sc}_{\mathscr{B}}(p_i)=\bigcup_{k|s_k\in p_i}\mathrm{sc}_{\mathscr{A}}(s_k),$$

and since it is clear that $\{j|c_{kj}=1$ for k such that $s_k\in p_i\}=\{j|e_{ij}=1\}$, we have

$$\mathrm{sc}_{\mathscr{B}}(p_i)=\bigcup_{k|s_k\in p_i}\left(\bigcup_{j|c_{kj}=1}\mathrm{ch}_j^{\mathscr{A}}\right)=\bigcup_{j|e_{ij}=1}\mathrm{ch}_j^{\mathscr{A}}.$$ Q.E.D.

Corollary: $\mathrm{sc}_{\mathscr{B}}(p_i)\supseteq\mathrm{sc}_{\mathscr{B}}(p_j)$, iff the row (of an EAM of $\mathscr{A}$) corresponding to p_i covers the row corresponding to p_j. By covering we mean the same as in an SCM. (See Definition 3.)

Thus we can carry out the reduction of the DA $\mathscr{B}$, not only using an SCM but also using an EAM of $\mathscr{A}$. Namely two states of $\mathscr{B}$ which have an identical pattern of 1's and 0's in the corresponding rows of an EAM of $\mathscr{A}$ can be merged. We also note that the reduction of $\mathscr{C}$ can be carried out as well by merging any set of states whose corresponding

	$\{q_0\}$	$\{q_1\}$	$\{q_2\}$
$\{p_0\}$		1	1
$\{p_1, p_3\}$	2,3		2,3
$\{p_2\}$	3	1	1,3

(a)

	$\{q_0\}$	$\{q_1\}$	$\{q_2\}$
$\{p_0\}$		1	1
$\{p_1, p_3\}$	1		1
$\{p_2\}$	1	1	1

(b)

Fig. 4. RSM and RAM of $\mathscr{A}$. (a) RSM of $\mathscr{A}$. (b) RAM of $\mathscr{A}$.

columns have an identical pattern of 1's and 0's. This follows from the fact that the definitions of $\mathscr{B}$ and $\mathscr{C}$ are symmetric.

The above observation implies, in terms of an SM, the following. Two states of $\mathscr{B}(\mathscr{C})$ which have an identical pattern of blank entries in the corresponding rows (columns) of an SM of $\mathscr{A}$ can be merged. We shall call two rows (columns) *equivalent* if they have an identical pattern of blank entries.

Definition 8: A *reduced states map* (*RSM*) of $\mathscr{A}$ is obtained from an SM of $\mathscr{A}$ by merging all the equivalent rows and columns. The merging of two rows (columns) means the replacing of the two rows (columns) by a new row (column), each entry of which is the union of the entries in the corresponding column (row).

Example: We consider the NDA $\mathscr{A}$ of Fig. 2(a) again. An SM of $\mathscr{A}$ was obtained in Fig. 3(a). It is seen that the rows p_1 and p_3 are equivalent. Merging the two rows, we obtain an RSM shown in Fig. 4(a). No two columns are equivalent, since no two columns of Fig. 3(a) have an identical pattern of blank entries.

Definition 9: The reduced DA associated with $\mathscr{A}$, denoted $\hat{\mathscr{B}} = (\hat{P}, \hat{M}', \hat{p}_0, \hat{F}')$, is obtained from $\mathscr{B}$ as follows:

$$\hat{P} = \{[p] | p \in P\},$$

where $[p]$ is the equivalence class of the states of $\mathscr{B}$ containing p. Very often we shall consider $[p]$ as consisting of states of $\mathscr{A}$, namely $[p] = \{s | (\exists p' \in [p])[s \in p']\}$.

$$\hat{M}'([p_i], \sigma) = [p_j] \Leftrightarrow M'(p_i, \sigma) = p_j \quad \text{for} \quad \forall \sigma \in \Sigma.$$

$$\hat{p}_0 = [p_0]$$

$$\hat{F}' = \{[p] | p \in F\}.$$

Lemma 3: An RSM of $\mathscr{A}$ is obtained from $\hat{\mathscr{B}}$ and $\hat{\mathscr{C}}$ in the same manner as an SM is obtained from $\mathscr{B}$ and $\mathscr{C}$ (Definition 7).

Proof: Easy and left to the reader.

Definition 10: A *reduced automaton matrix* (*RAM*) of $\mathscr{A}$ is obtained from an RSM of $\mathscr{A}$ by replacing each nonblank entry by a 1.

Note that the reduced DA associated with $\overleftarrow{\mathscr{A}}$ and $\rightleftarrows$ are $\mathscr{C}$ and $\mathscr{B}$, respectively. Therefore the transpose of an RAM of $\mathscr{A}$ is an RAM of $\overleftarrow{\mathscr{A}}$.

Theorem 4: Let I be an equivalence class of automata. A unique (within permutation of the rows and columns) RAM exists for all automata in I.

Proof: Let $\mathscr{A} \in I$. By definition, an RAM of $\mathscr{A}$ is an SCM of $\hat{\mathscr{B}}$, which is unique for $\forall \mathscr{A} \in I$ except the names of the states. On the other hand, an SCM of a DA is unique within permutation of rows and columns. (See Theorems 1 and 2.) Hence follows the theorem. Q.E.D.

Note that the converse of Theorem 4 is not true. Namely, two automata having the same RAM may belong to different equivalence classes.

V. Grids and Covers

In this section, we introduce the concepts of a *grid* and a *cover*, and then relate them with states of an automaton.

Definition 11: Given an EAM (or RAM), if all the entries at the intersections of a set of rows $\{p_{i_1}, \cdots, p_{i_a}\}$ and a set of columns $\{q_{j_1}, \cdots, q_{j_b}\}$ are 1's, then this set of 1's is said to form a *grid*. We represent the grid by $g = (p_{i_1}, \cdots, p_{i_a}; q_{j_1}, \cdots, q_{j_b})$. The grid g is said to *contain* the pair (p_i, q_j), if $i \in \{i_1, \cdots, i_a\}$ and $j \in \{j_1, \cdots, j_b\}$. A grid g_1 *contains*, or is *an extension of* another grid g_2, written $g_1 \supseteq g_2$, if all the pairs contained in g_2 are also contained in g_1.

Definition 12: A set of grids forms a *cover* if every 1 in the EAM (or RAM) belongs to at least one grid in the set. A *minimum cover* is a cover which consists of the minimum number of grids.

Definition 13: Given a cover over an EAM (or RAM), a *cover map* (*CM*) is obtained by replacing each 1 in the EAM (or RAM) by the names of all the grids (in the given cover) it belongs to.

If we connect the same name with horizontal and vertical lines, we will see a "grid" in the conventional sense of the word

Theorem 5: Given an automaton $\mathscr{A}$,

1) an SM (RSM) of $\mathscr{A}$ is a CM, namely, the states of $\mathscr{A}$ appear as a cover over an EAM (RAM) of $\mathscr{A}$, and
2) in the CM above, all the initial states of $\mathscr{A}$ appear in the row corresponding to $p_0(\hat{p}_0)$ and all the final states of $\mathscr{A}$ appear in the column corresponding to q_0 ($\hat{q}_0$). (For the definitions of p_0, $\hat{p}_0$, q_0, and $\hat{q}_0$, see Section IV.)

Proof: 1) It is clear that any 1 in an EAM (RAM) belongs to at least one grid which corresponds to a state of $\mathscr{A}$. Suppose a state s of $\mathscr{A}$ appears at two points in a CM, which are neither in the same row nor the same column. That is to say, $s \in p_{i_1} \cap q_{j_1}$ and $s \in p_{i_2} \cap q_{j_2}$, where $i_1 \neq i_2$ and $j_1 \neq j_2$. These relations imply that $s \in p_{i_1} \cap q_{j_2}$ and $s \in p_{i_2} \cap q_{j_1}$. Therefore s must appear also at (i_1, j_2) and (i_2, j_1). The case where the two points (i_1, j_1) and (i_2, j_2) are in the same row or in the same column is trivial. 2) follows from the construction of $D(\mathscr{A})$ and $D(\overleftarrow{\mathscr{A}})$. Q.E.D.

Example: It is seen that in Figs. 3(a) and 4(a), the states of $\mathscr{A}$ appear as a cover, and that the initial state 1 appears

in the row corresponding to p_0 ($\hat{p}_0$) and the final states 2 and 3 appear in the column corresponding to q_0 ($\hat{q}_0$).

PART 2: SYNTHESIS

In Part 2 we attack the problem of finding a minimum NDA equivalent to a given automaton, which is the final objective of the whole paper. One time-honored method of synthesis is to turn analysis "inside out." We shall resort to this method here. One vital tool of our analysis in Part 1 was the subset construction. In the following section we introduce an inverse operation of the subset construction and investigate its properties.

VI. Intersection Rule

To begin with, we recall the following relationship of a given NDA $\mathscr{A}=(S, M', S_0, F')$ to its subset DA $D(\mathscr{A})=(P, M, p_0, F)$:[5]

1) $p_0=S_0$
2) $(\forall s\in S)(\forall p\in P)[s\in F'\Rightarrow[s\in p\Rightarrow p\in F]]$, i.e., all final states of $\mathscr{A}$ are only contained in final states of $D(\mathscr{A})$
3) $(\forall\sigma\in\Sigma)(\forall s,\ \forall s'\in S)(\forall p\in P)[s'\in M'(s,\ \sigma)\Rightarrow[s\in p\Rightarrow s'\in M(p,\sigma)]]$, i.e., if there is a transition from s to s' under σ in $\mathscr{A}$, then there also is a transition under σ in $D(\mathscr{A})$ from any subset containing s to a subset containing s'.

Now suppose we are only shown $D(\mathscr{A})$. Is it possible to reconstruct $\mathscr{A}$? This question motivates the following two definitions.

Definition 14: Let $\mathscr{B}=(P, M, p_0, F)$ be a DA. The pair $\langle Z, f\rangle$ is called a *subset assignment* to $\mathscr{B}$ if Z is a finite set and $f: P\to 2^Z-\{\varnothing\}$ is a function. Such an f is called a *subset assignment function.* The *natural subset assignment* to the subset DA $D(\mathscr{A})$ is $\langle S, f\rangle$, where S is the set of states of $\mathscr{A}$ and $f(p)=\{s|s\in p\}$ for $\forall p\in P$. The *subset assignment associated with a cover* over an EAM (or RAM) of an NDA $\mathscr{A}$ assigns to each state of $\mathscr{B}=D(\mathscr{A})$ (or $\hat{\mathscr{B}}$) the set of grids appearing in the corresponding row of the EAM (or RAM).

According to this definition, the natural subset assignment to $D(\mathscr{A})$ is the subset assignment associated with the cover over an SM (considered as a CM over an EAM) of $\mathscr{A}$.

Definition 15: Let $\mathscr{B}=(P, M, p_0, F)$ be a DA, and let $\langle Z, f\rangle$ be a subset assignment to $\mathscr{B}$. Then $I(Z, f, \mathscr{B})$ is the NDA (Z, N, Z_0, G), where for $\forall z\in Z$, $\forall p\in P$, and $\forall\sigma\in\Sigma$,

1) $Z_0=f(p_0)$
2) $z\in G\Leftrightarrow[z\in f(p)\Rightarrow p\in F]$
3) $z'\in N(z,\sigma)\Leftrightarrow[z\in f(p)\Rightarrow z'\in f(M(p,\sigma))]$.

$I(Z, f, \mathscr{B})$ is called the NDA obtained by the *intersection rule* from $\mathscr{B}$.

Definition 15, among other things, specifies the transition function N of $I(Z, f, \mathscr{B})$. It is easy to see that the definition amounts to

$$N(z,\sigma)=\bigcap_{p|z\in f(p)} f(M(p,\sigma)) \quad \text{for } \forall z\in Z \text{ and } \forall\sigma\in\Sigma.$$

[5] Note that we have switched the previous notations from M to M' and F to F', and vice versa, since in this section we are mainly concerned with $D(\mathscr{A})$ and would like to avoid primed symbols.

This explains why the name *intersection rule* has been chosen, and also makes the application of the rule very easy.

Example: We apply the intersection rule to the DA $\mathscr{B}$ of Fig. 2(b). The natural subset assignment $\langle S, f\rangle$ is shown in the figure, where S is the set of states of $\mathscr{A}$. Let us determine $N(3, 0)$ as an example:

$$\begin{aligned} N(3,0) &= \bigcap_{p|3\in f(p)} f(M(p,0)) \\ &= f(M(p_2,0))\cap f(M(p_3,0)) \\ &= f(p_2)\cap f(p_0) \\ &= \{1,3\}\cap\{1\}=\{1\}. \end{aligned}$$

Now a question presents itself: Is the intersection rule really an inverse of the subset construction? Or more precisely, is $bh(I(S, f, \mathscr{B}))=bh(\mathscr{B})$, where $\langle S, f\rangle$ is the natural subset assignment? Later we shall answer this question affirmatively. (See Lemma 5.)

Then the next natural question is: How can one find a subset assignment $\langle Z, f\rangle$ for which $bh(I(Z, f, \mathscr{B}))=bh(\mathscr{B})$ and such that the size of Z is minimum? As a matter of fact, answering this question is precisely the ultimate objective of this paper. For the time being we shall drop the minimality requirement on Z and investigate properties of various subset assignments.

Definition 16: A subset assignment $\langle Z, f\rangle$ to a DA $\mathscr{B}$ is *legitimate* iff $bh(I(Z, f, \mathscr{B}))=bh(\mathscr{B})$. A cover is *legitimate* iff its associated subset assignment is legitimate.

Thus if $\mathscr{B}=D(\mathscr{A})$ for some NDA $\mathscr{A}$, the legitimacy of $\langle Z, f\rangle$ implies $bh(I(Z, f, \mathscr{B}))=bh(\mathscr{A})$. In the rest of this section we let $\mathscr{B}$ and $\langle Z, f\rangle$ be as defined in Definition 15.

Lemma 4: For $\forall\langle Z, f\rangle$, $bh(I(Z, f, \mathscr{B}))\subseteq bh(\mathscr{B})$.

Proof: See Appendix I.

Lemma 5: The natural subset assignment is legitimate.

Proof: Let $\mathscr{A}=(S, M', S_0, F')$ and $\mathscr{B}=D(\mathscr{A})$. We want to show $bh(I(S, f, \mathscr{B}))=bh(\mathscr{B})$ for the natural subset assignment $\langle S, f\rangle$. Let $I(S, f, \mathscr{B})=(S, N, Z_0, G)$. Then for $\forall\sigma\in\Sigma$, $\forall s\in S$ and $p\in P$,

1) $Z_0 = p_0 = S_0$
2) $s\in F'\Rightarrow[s\in p\Rightarrow p\in F]\Rightarrow s\in G$,

so that

$$F'\subseteq G.$$

3) $s'\in M'(s,\sigma)\Rightarrow[s\in p\Rightarrow s'\in M(p,\sigma)]$
$\Rightarrow s'\in N(s,\sigma)$,

so that

$$M'(s,\sigma)\subseteq N(s,\sigma).$$

Thus it is clear that $M'(S_0, x)\cap F'\neq\varnothing\Rightarrow N(Z_0, x)\cap G\neq\varnothing$. This implies that $bh(\mathscr{A})\subseteq bh(I(S, f, \mathscr{B}))$. But $bh(\mathscr{A})=bh(\mathscr{B})$ by the corollary to Lemma 1 and $bh(I(S, f, \mathscr{B}))\subseteq bh(\mathscr{B})$ by Lemma 4. Hence $bh(I(S, f, \mathscr{B}))=bh(\mathscr{B})$. Q.E.D.

Note that $I(S, f, \mathscr{B})$ is not necessarily identical to $\mathscr{A}$, as is exemplified by the following example.

Example: Fig. 2(b) shows $\mathscr{B}=D(\mathscr{A})$ with the natural subset assignment $\langle S, f\rangle$, where S is the set of states of $\mathscr{A}$.

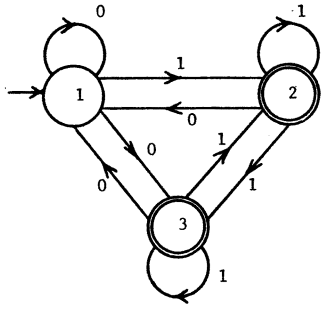


Fig. 5. NDA obtained from $\mathscr{B}$ by the intersection rule.

$I(S, f, \mathscr{B})$ is shown in Fig. 5. Note that $I(S, f, \mathscr{B}) \neq \mathscr{A}$ although the two NDA's are equivalent, as may be verified by constructing $D(I(S, f, \mathscr{B}))$.

The implication of the following theorem is that the search for a legitimate subset assignment to any DA $\mathscr{B}$ may be restricted to the subset assignments to $\hat{\mathscr{B}}$.

Theorem 6: Let $\hat{\mathscr{B}} = (\hat{P}, \hat{M}, \hat{p}_0, \hat{F})$ be the reduced DA equivalent to $\mathscr{B}$. If $\langle Z, f \rangle$ is a legitimate subset assignment to $\mathscr{B}$, then $\langle Z, \hat{f} \rangle$ is a legitimate subset assignment of $\hat{\mathscr{B}}$, where

$$f([p]) = \bigcup_{p' \in [p]} f(p').$$

Proof: Let $I(Z, f, \mathscr{B}) = (Z, N, Z_0, G)$ and $I(Z, \hat{f}, \hat{\mathscr{B}}) = (Z, N', Z_0', G')$. For $\forall \sigma \in \Sigma$, $\forall z \in Z$, and $\forall p \in P$, we have the following relations.

1) $$\begin{aligned} z \in Z_0 &\Leftrightarrow z \in f(p_0) \\ &\Rightarrow z \in \hat{f}([p_0]) \\ &\Rightarrow z \in Z_0', \end{aligned}$$

so that

$$Z_0 \subseteq Z_0'.$$

2) $$\begin{aligned} z \in G &\Leftrightarrow [z \in f(p) \Rightarrow p \in F] \\ &\Leftrightarrow [z \in \hat{f}([p]) \Rightarrow [p] \in \hat{F}] \\ &\Leftrightarrow z \in G', \end{aligned}$$

so that

$$G = G'.$$

3) $$\begin{aligned} z' \in N(z, \sigma) &\Leftrightarrow [z \in f(p) \Rightarrow z' \in f(M(p, \sigma))] \\ &\Rightarrow [z \in f([p]) \Rightarrow z' \in f(M([p], \sigma))] \\ &\Rightarrow z' \in N'(z, \sigma) \end{aligned}$$

so that

$$N(z, \sigma) \subseteq N'(z, \sigma).$$

We may deduce, as in the proof of Lemma 5, that $bh(I(Z, f, \mathscr{B})) \subseteq bh(I(Z, \hat{f}, \hat{\mathscr{B}}))$. But $bh(I(Z, \hat{f}, \hat{\mathscr{B}})) \subseteq bh(\hat{\mathscr{B}})$ by Lemma 4 and $bh(\mathscr{B}) = bh(I(Z, f, \mathscr{B}))$ by our assumption, which together yield $bh(I(Z, \hat{f}, \hat{\mathscr{B}})) \subseteq bh(I(Z, f, \mathscr{B}))$. Hence $bh(I(Z, \hat{f}, \hat{\mathscr{B}})) = bh(I(Z, f, \mathscr{B})) = bh(\mathscr{B})$. Q.E.D.

Example: The $\hat{\mathscr{B}}$ for our running example $\mathscr{A}$ is shown in Fig. 6(a). Note that $I(S, \hat{f}, \hat{\mathscr{B}})$ shown in Fig. 6(b) is equivalent but not identical to $I(S, f, \mathscr{B})$.

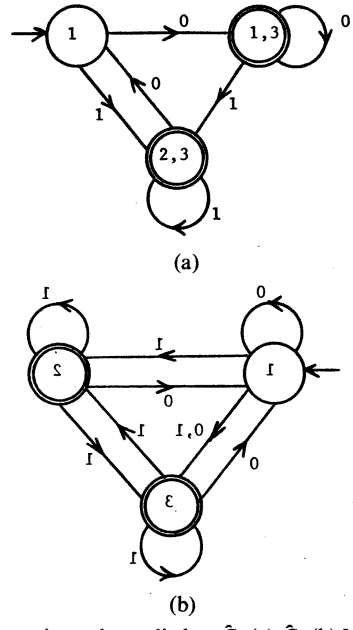


Fig. 6. Intersection rule applied to $\hat{\mathscr{B}}$. (a) $\hat{\mathscr{B}}$. (b) NDA obtained from $\hat{\mathscr{B}}$ by the intersection rule.

VII. Subset Assignments Associated with Covers

First we state a basic theorem which follows as a corollary to our previous theorems.

Theorem 7: Given any automaton $\mathscr{A}$, there exists a legitimate cover over an RAM of $\mathscr{A}$ such that the NDA obtained by the intersection rule using the subset assignment associated with the cover is a minimum NDA.

Proof: Follows from Theorems 4, 6, and Lemma 5. Choose a minimum NDA in the equivalence class I of Theorem 4. Its RSM shows a cover with the desired property. Q.E.D.

Definition 17: A legitimate subset assignment $\langle Z, f \rangle$ to a DA such that the size of Z is minimum possible is called a *minimum legitimate subset assignment.* A *minimum legitimate cover* is a legitimate cover consisting of the minimum number of grids.

Theorem 7 implies that we have only to consider covers over an RAM of a given NDA, in order to obtain a minimum NDA. Next we shall show that we can further restrict the domain of search for a minimum legitimate cover.

Definition 18: A grid is called a *prime grid* iff it cannot be properly contained in any other grid.

Lemma 6: For any grid there exists at least one prime grid that contains it.

Lemma 7: Let $\mathscr{A} = (S, M', S_0, F')$ be an NDA, $\mathscr{B} = (P, M, p_0, F)$ be $D(\mathscr{A})$, and let $s_i \in S$ and $p_k \in P$ such that $s_i \notin p_k$. The grid over an EAM (or RAM) of $\mathscr{A}$ corresponding to s_i (see Theorem 5) can be extended to the row of the EAM (or RAM) corresponding to p_k iff

$$sc_{\mathscr{A}}(s_i) \subseteq sc_{\mathscr{B}}(p_k).$$

Proof: Note

$$sc_{\mathscr{A}}(s_i) = \bigcup_{j | c_{ij} = 1} ch_j^{\mathscr{A}} \quad \text{(Theorem 2)}$$

and

$$\mathrm{sc}_{\mathscr{B}}(p_k) = \bigcup_{l|e_{kl}=1} \mathrm{ch}_l^{\mathscr{A}} \quad \text{(Theorem 3).}$$

From the constructions of an SCM of $\mathscr{A}$ and EAM of $\mathscr{A}$, it should be clear that

$$s_i \text{ is extendable to } p_k \Leftrightarrow \{j|c_{ij}=1\} \subseteq \{l|e_{kl}=1\}.$$

Thus we have

$$s_i \text{ is extendable to } p_k \Leftrightarrow \mathrm{sc}_{\mathscr{A}}(s_i) \subseteq \mathrm{sc}_{\mathscr{B}}(p_k). \quad \text{Q.E.D.}$$

Lemma 8: Let $\mathscr{A}$ and $\mathscr{B}$ be as defined in Lemma 7. If $\mathrm{sc}_{\mathscr{A}}(s_i) \subseteq \mathrm{sc}_{\mathscr{B}}(p_k)$, then $(\forall\sigma\in\Sigma)(\forall s\in M'(s_i,\sigma))[\mathrm{sc}_{\mathscr{A}}(s) \subseteq \mathrm{sc}_{\mathscr{B}}(M(p_k,\sigma))]$.

Proof: It is clear that

$$\mathrm{sc}_{\mathscr{A}}(s_i) = \Big(\bigcup_{\sigma\in\Sigma} \sigma\Big(\bigcup_{s\in M'(s_i,\sigma)} \mathrm{sc}_{\mathscr{A}}(s)\Big)\Big) \cup \delta(s_i)$$

where

$$\delta(s_i) = e \text{ if } s_i \in F' \text{ and } \varnothing \text{ otherwise.}$$

Similarly we have

$$\mathrm{sc}_{\mathscr{B}}(p_k) = \Big(\bigcup_{\sigma\in\Sigma} \sigma\, \mathrm{sc}_{\mathscr{B}}(M(p_k,\sigma))\Big) \cup \delta(p_k)$$

where

$$\delta(p_k) = e \text{ if } p_k \in F \text{ and } \varnothing \text{ otherwise.}$$

By taking the derivative (Brzozowski [9]) of both sides of

$$\Big(\bigcup_{\sigma\in\Sigma} \sigma\Big(\bigcup_{s\in M'(s_i,\sigma)} \mathrm{sc}_{\mathscr{A}}(s)\Big)\Big) \cup \delta(s_i) \subseteq \Big(\bigcup_{\sigma\in\Sigma} \sigma\, \mathrm{sc}_{\mathscr{B}}(M(p_k,\sigma))\Big) \cup \delta(p_k)$$

with respect to σ, we get

$$\bigcup_{s\in M'(s_i,\sigma)} \mathrm{sc}_{\mathscr{A}}(s) \subseteq \mathrm{sc}_{\mathscr{B}}(M(p_k,\sigma)).$$

Hence follows the lemma. Q.E.D.

Theorem 8: If a legitimate cover is given, we can derive a legitimate cover consisting only of prime grids which are extensions of the grids in the given cover. Moreover, the number of the prime grids in the new cover is no more than the number of grids in the given cover.

Proof: See Appendix II.

Corollary: In finding a minimum legitimate cover, we have only to consider prime grids over an RAM.

VIII. Minimization Procedure

With the preparations in the previous sections, we now are ready to describe our minimization procedure of NDA's. We present two lemmas before describing our procedure. The first lemma, which is presented without proof since it is simple, usually simplifies the finding of prime grids a great deal.

Lemma 9: If a row (column) of an RAM can be expressed as a union[4] of some other rows (columns), the elimination of the row (column) does not affect the finding of prime grids, provided that the eliminated row (column) is included in the first (second) coordinate of every prime grid that can be extended to that row (column).

When we eliminate either rows or columns only, as described in Lemma 9, we can eliminate them one by one without going back to the original RAM. Note, however, that when both rows and columns are eliminated, we have to make sure that the eliminated row (or column) is a union of some other rows (or columns) in the original RAM.

	q_0	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8	q_9
p_0		α	α							
p_1			β	β						
p_2				γ	γ					
p_3						δ	δ			
p_4	ϵ					δ	δ,ϵ			
p_5	ϵ						ϵ	η	η	
p_6								η	η,ζ	ζ
p_7									ζ	ζ

(a)

State \ Input		0	1
Initial	p_0	p_4	$p_\emptyset$
	p_1	p_5	$p_\emptyset$
	p_2	p_6	p_1
	p_3	p_2	p_7
Final	p_4	p_2	p_7
Final	p_5	p_1	p_3
	p_6	p_0	p_3
	p_7	p_0	$p_\emptyset$
	$p_\emptyset$	$p_\emptyset$	$p_\emptyset$

(b)

Fig. 7. Example of illegitimate minimum cover. (a) CM. (b) DA associated with the CM.

Lemma 10: There exists an RAM such that no minimum cover over it is legitimate.[6]

Proof: We shall give an example of such an RAM. Fig. 7 shows the reduced DA associated with an NDA (not shown) and a minimum cover over the RAM of the NDA. It is clear that the cover shown is the only minimum cover. That is is not legitimate can be shown by actually constructing an NDA using the intersection rule. Q.E.D.

Minimization Procedure:

1) Given an NDA, construct its RAM and find a minimum cover.
2) Let i_0 be the number of grids in the minimum cover found above and set $i=i_0$
   a) For each cover containing i prime grids, test whether it is legitimate.
   b) If no legitimate cover is found, set $i=i+1$, and go to step a.

[6] This lemma was proved by Professor S. Even of Harvard University, who also provided us with an example in the proof.

It is obvious that the process should terminate after a finite number of cycles. In fact, an upper bound on the number of cycles is given by MIN(l, n, m), where l, n, m are, respectively, the number of states in the given NDA, that in the reduced DA associated with the given NDA, and that in the reduced DA associated with the dual of the given NDA.

Example: We apply our procedure to the NDA $\mathscr{A}$ of Fig. 2(a). The reduced DA $\hat{\mathscr{B}}$ associated with $\mathscr{A}$, shown in Fig. 8(a), was obtained from $\mathscr{B}$ (Fig. 2(b)). An RAM of $\mathscr{A}$ was obtained before (Fig. 4(b)). In this case we have a unique minimum cover shown in Fig. 8(b). Fig. 8(c) shows the subset assignment associated with this cover. Using the intersection rule, we obtain the NDA shown in Fig. 8(d). It can be easily verified that the cover we used is legitimate, by applying the subset construction of the NDA of Fig. 8(d) and showing that it is equivalent to $\mathscr{B}$.

In the above example, the only minimum cover turned out to be legitimate. At present we do not know how typical this is, although we do know that this is not always the case (see Lemma 10).

In the above form our minimization procedure requires that every cover be tested in some systematic way until a legitimate cover is found. This may take the following steps. We let $\hat{\mathscr{B}}=(P, M, p_0, F)$ be the reduced DA associated with the given NDA.

1) Apply the intersection rule to $\hat{\mathscr{B}}$ to obtain $\mathscr{M}=I(Z, f, \hat{\mathscr{B}})=(Z, N, Z_0, G)$, where $\langle Z, f\rangle$ is the subset assignment associated with the chosen cover.
2) Construct $D(\mathscr{M})$.
3) Reduce $D(\mathscr{M})$ and test if it is equivalent to $\hat{\mathscr{B}}$.

We shall show that the steps 2 and 3 can be replaced by a simpler test.

Lemma 11: Given the CM of a legitimate cover over an RAM, let $\hat{\mathscr{B}}$ and $\mathscr{M}$ be as defined above. Suppose the cover contains a grid $z=(p_1, \cdots, p_a; q_1, \cdots, q_b)$.[7] Then

$$\mathrm{sc}_{\mathscr{M}}(z) \subseteq \bigcup_{k=1}^{b} \mathrm{ch}_k^{\hat{\mathscr{B}}}.$$

Proof: The RAM mentioned in the theorem can be considered as an SCM of $\hat{\mathscr{B}}$. (See the proof of Theorem 4.) Since z is a grid, the SCM of $\hat{\mathscr{B}}$ has 1's at the intersections of rows, $p_1, \cdots, p_a$ and the columns $q_1, \cdots, q_b$. Therefore it follows from Theorem 2 that

$$\mathrm{sc}_{\hat{\mathscr{B}}}(p_i) \subseteq \bigcup_{k=1}^{b} \mathrm{ch}_k^{\hat{\mathscr{B}}}, \qquad i = 1, 2, \cdots, a. \tag{1}$$

Consider now the automata, $\mathscr{M}_z=(Z, N, z, G)$ and $\hat{\mathscr{B}}_i=(P, M, p_i, F)$ for $i=1, 2, \cdots, a$. It follows from Lemma 4 that

$$bh(\mathscr{M}_z) \subseteq bh(\hat{\mathscr{B}}_i), \qquad i = 1, 2, \cdots, a,$$

which means

$$\mathrm{sc}_{\mathscr{M}}(z) \subseteq \mathrm{sc}_{\hat{\mathscr{B}}}(p_i), \qquad i = 1, 2, \cdots, a.$$

[7] Strictly speaking, we should use double indexing like $p_{i_1}, \cdots, p_{i_a}$, but no confusion should arise here.

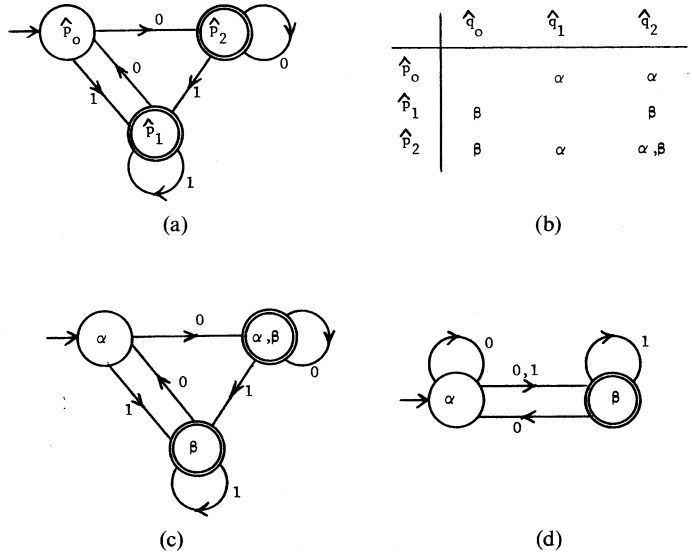


Fig. 8. Minimization of $\mathscr{A}$ of Fig. 2(a). (a) $\hat{\mathscr{B}}$. (b) CM. (c) $\hat{\mathscr{B}}$ with $\langle Z, f\rangle$. (d) $I(Z, f, \hat{\mathscr{B}})$.

This, together with (1), completes the proof. Q.E.D.

Now let $D(\mathscr{M})=(U, L, u_0, H)$.

Lemma 12: For $\forall u \in U$ and $\forall p \in P$,

$$u \sim p \Leftrightarrow (\forall \sigma \in \Sigma)[L(u, \sigma) \sim M(p, \sigma)] \wedge [u \in H \Leftrightarrow p \in F].$$

Proof: Similar to the proof of Lemma 8. Q.E.D.

Definition 19: A set of grids in a row of a CM *spans* the row iff at least one grid in the set appears in each nonblank entry in the row.

Theorem 9: Using the current notations, we have the following.

1) $(\exists \sigma \in \Sigma)[L(u, \sigma)$ does not span the row $M(p, \sigma)$ of the RAM $\vee [M(p, \sigma) \in G \wedge L(u, \sigma) \notin H]] \Rightarrow \sim p$.
2) $(\forall x \in \Sigma^*)[L(u, x)$ spans the row $M(p_0, x) \wedge [M(p_0, x) \in G \Leftrightarrow L(u_0, x) \in H]] \Rightarrow bh(\mathscr{M}) = bh(\hat{\mathscr{B}})$.

Proof: 1) Suppose $L(u, \sigma)$ does not span the row $M(p, \sigma)$. Then

$$\begin{aligned} \mathrm{sc}_{D(\mathscr{M})}(L(u, \sigma)) &= \bigcup_{z \in L(u, \sigma)} \mathrm{sc}_{\mathscr{M}}(z) && \text{(Lemma 1)} \\ &\subset \bigcup_{z \in f(L(u, \sigma))} \mathrm{sc}_{\mathscr{M}}(z) && \text{(Lemma 11)} \\ &\subseteq \mathrm{sc}_{\hat{\mathscr{B}}}(M(p, \sigma)) && \text{(Lemma 11).} \end{aligned}$$

Note that the second line above is a proper cotainment. Thus $u \sim p$ by Lemma 12. The rest follows from the fact that

$$M(p, \sigma) \in G \wedge L(u, \sigma) \notin H \Rightarrow L(u, \sigma) \sim M(p, \sigma).$$

2) The condition implies that there is a homomorphism $\phi: U \to P$ such that

$$(\forall u \in U)[\phi(L(u, \sigma)) = M(\phi(u), \sigma)]$$

and

$$u \in H \Leftrightarrow \phi(u) \in F.$$

Therefore it follows that $bh(\mathscr{M}) = bh(D(\mathscr{M})) = bh(\hat{\mathscr{B}})$. Q.E.D.

Now we can replace the steps 2 and 3 above by the following.

2') Consider the list of pairs (u_i, p_i), where $u_i \in U$, $p_i \in P$, starting with (u_0, p_0).
3') Given a pair (u, p), test $L(u, \sigma)$ and $M(p, \sigma)$ against the condition of part 1 of Theorem 9 for each $\sigma \in \Sigma$. If it is satisfied, then $u \sim p$ and the cover under test is illegitimate. Otherwise add the pair $(L(u, \sigma), M(p, \sigma))$ to the list, if it is not already there.
4') Repeat 3') for each new pair. If the process terminates without detecting illegitimacy, then the cover under test is legitimate.

Conclusion

After obtaining some basic results in Part 1, in Part 2 we presented a minimization procedure for NDA's, which is more efficient than the only existing procedure, namely the exhaustive search. Except for few examples, most of the samples we have tried have the property that there exists a minimum cover that is legitimate. But at the time of writing, we do not know in quantitative terms how often it is the case. At present we are looking for a more efficient method of finding legitimate covers.

Appendix I

Proof of Lemma 4

Let $lg(x)$ be the *length* of the sequence x with the provision that $lg(e)=0$. We first prove

$$N(Z_0, x) \subseteq f(M(p_0, x)) \tag{2}$$

by the induction on $lg(x)$.

1) $lg(x)=0$: This implies $x=e$. Thus we have

$$N(Z_0, e) = Z_0 = f(M(p_0, e)).$$

2) Assume true for all x such that $lg(x)<l$. Let $x=\sigma_1\sigma_2\cdots\sigma_{l-1}$ and $x'=x\sigma_l$, where $\sigma_i \in \Sigma$ for $i=1, 2, \cdots, l$.

$$\begin{aligned} N(Z_0, x') &= N(N(Z_0, x), \sigma_l) \\ &= \bigcup_{z \in N(Z_0, x)} N(z, \sigma_l) \\ &\subseteq \bigcup_{z \in f(M(p_0, x))} N(z, \sigma_l) \\ &= \bigcup_{z \in f(M(p_0, x))} \Big(\bigcap_{p \mid z \in f(p)} f(M(p, \sigma_l)) \\ &\subseteq f(M(M(p_0, x), \sigma_l)) \\ &= f(M(p_0, x')). \end{aligned}$$

We now can prove the lemma as follows.
For $\forall x \in \Sigma^*$,

$$\begin{aligned} x \in bh(I(Z, f, \mathscr{B})) &\Leftrightarrow N(Z_0, x) \cap G \neq \varnothing \\ &\Rightarrow f(M(p_0, x)) \cap G \neq \varnothing \quad \text{by (2)} \\ &\Rightarrow M(p_0, x) \in F \\ &\Rightarrow x \in bh(\mathscr{B}). \end{aligned}$$

Hence

$$bh(I(Z, f, \mathscr{B})) \subseteq bh(\mathscr{B}). \qquad \text{Q.E.D.}$$

Appendix II

Proof of Theorem 8

Let $\langle Z, f \rangle$ be the subset assignment associated with the given cover and let $\mathscr{M} = I(Z, f, \mathscr{B})$ and $\mathscr{N} = D(\mathscr{M}) = (U, L, u_0 -)$. (The entries which are not important in our present discussion are indicated by $-$.) Let $\hat{\mathscr{N}} = (\hat{U}, \hat{L}, \hat{u}_0, -)$ be the reduced *DA* associated with $\mathscr{N}$. We now consider an RSM of $\mathscr{M}$ and the subset assignment $\langle Z, h \rangle$ associated with it. We know by Theorem 6 that $I(Z, h, \hat{\mathscr{N}})$ is equivalent to $\mathscr{N}$, and hence to $\mathscr{M}$. The fact that the grid (in an RSM of $\mathscr{M}$) corresponding to $z \in Z$ can be extended to the row corresponding to $u \in \hat{U}$ means that $sc_{\mathscr{M}}(z) \subseteq sc_{\hat{\mathscr{N}}}(u)$ (Lemma 7).

We modify the legitimate subset assignment function h to get another subset assignment function h', which will be shown to be legitimate, and to be the subset assignment function associated with a cover containing only prime grids.

Case 1: In an RSM of $\mathscr{M}$ there is no grid z which can be extended to a row u such that $z \notin u$. In this case no vertical extension of any grid is possible. (See the comment after Definition 13.) We now extend all the grids horizontally as far as possible, so that all the grids are converted to prime grids. In this process the function is not changed. Therefore $h'=h$ and now we have a legitimate cover containing only prime grids.

Case 2: There exists a grid z extendable to a row u such that $z \notin u$. In this case we define h' by the following recursive definition.

a) $h'(u) = h(u) \cup \{z\}$.

This corresponds to the extension of z to u.

b) $h'(\hat{L}(u, \sigma)) = h(\hat{L}(u, \sigma)) \cup K(z, \sigma)$ for $\forall \sigma \in \Sigma$

where K is the transition function of $I(Z, h, \mathscr{N}')$. This makes sure that

$$(\forall \sigma \in \Sigma)(\forall z \in Z)\big[K(z, \sigma) \subseteq K'(z, \sigma)\big] \tag{3}$$

where K' is the transition function of $I(Z, h', \hat{\mathscr{N}})$. Note that the extensions implied by 2) are always possible because of Lemmas 7 and 8.

c) For each $z' \in K(z, \sigma) - h(\hat{L}(u, \sigma))$, replace z and u in a) by z' and $\hat{L}(u, \sigma)$. This makes sure that (3) holds true for each extension implied by b).

After this process terminates, we find a new pair z and u which satisfies the condition of Case 2, and apply the process. Finally we will arrive at Case 1 and all the grids will be extended to prime grids. By (3) and Lemma 4, we conclude that $\langle Z, h' \rangle$ is legitimate. Q.E.D.

Acknowledgment

The conditions in Theorem 6.2 of [7] are necessary but not sufficient. The authors would like to thank Dr. C. Carrez [10] for pointing out this fact, and Prof. S. Even of Harvard University for a number of useful discussions and for his contribution of Lemma 10. Thanks are also due to the referees who read the manuscript very carefully and offered numerous helpful suggestions.

References

[1] M. O. Rabin and D. Scott, "Finite automata and their decision problems," *IBM J. Res. and Develop.*, vol. 3, pp. 114–125, April 1959.
[2] G. H. Ott and N. H. Feinstein, "Design of sequential machines from their regular expressions," *J.ACM*, vol. 8, pp. 585–600, October 1961.
[3] P. H. Starke, "Einige Bemerkungen über nicht-deterministische Automaten," *EIK*, vol. 2, pp. 61–82, 1966.
[4] S. Neuber and P. H. Starke, "Über die Reduktion nicht-deterministischer Automaten," *EIK*, vol. 3, pp. 351–362, 1967.
[5] P. H. Starke, "Hüllenoperationen für nicht-deterministische Automaten," *EIK*, vol. 3, pp. 283–294, 1967.
[6] A. Schmitt, "Theorie der nicht-deterministischen und unvollständigen Mealey-Automaten," *Computing*, vol. 4, pp. 56–74, 1969.
[7] T. Kameda, "On the reduction of non-deterministic automata," Ph. D. dissertation, Department of Electrical Engineering, Princeton University, Princeton, N. J. Also, T. Kameda and P. Weiner, Computer Science Lab., Department of Electrical Engineering, Princeton University, Tech. Rept. 57, February 1968.
[8] J. A. Brzozowski, "Canonical regular expressions and minimal state graphs for definite events," *Proc. Symp. on Math. Theory of Automata*, vol. 12, Brooklyn, N. Y.: Brooklyn Polytechnic Institute, 1963, pp. 529–561.
[9] J. A. Brzozowski, "Derivatives of regular expressions," *J. ACM*, vol. 5, pp. 481–494, October 1964.
[10] C. Carrez, Faculte des Sciences, Lille, France, private communication.
[11] L. C. Eggan, "Transition graphs and the star height of regular events," *Mich. Math. J.*, vol. 10, pp. 385–397, 1963.

A Transform Approach to Logic Design

ROBERT J. LECHNER

Abstract —**This paper describes a new approach to the design of combinational logic using large-scale integrated (LSI) circuit technology. A simple "prototype" logic function of n binary variables is imbedded within an array of at most $(n+1)$ rows and columns. The cells of this array contain two-input EXCLUSIVE-OR gates, and its rows are fed by the input variables and logical "1." Its column outputs are first-degree polynomial functions of the input variables. These functions supply inputs to, and modify the output of, the prototype in order to realize the desired function. These transformations form a group; specifically, the largest subgroup of the $(n+1)$-dimensional affine group such that input variable encodings are not affected by feedback from the function's output.**

This approach to logic design complements rather than replaces conventional multilevel logic design. Its relative complexity is strongly dependent on the specific function (or set of functions) to be realized. In some cases, complexity is reduced; in others it is increased. Basically, EXCLUSIVE-OR gates have been introduced into the logic designer's "bag of tricks" in a particularly effective way: as an array rather than as separate components. This provides practical advantages, such as economical LSI array structures and effective new computational tools for the logic designer.

The number of prototypes required to generate all functions of n arguments is equal to the number of "prototype" equivalence classes (or P-classes) into which the group of feedback-free affine transformations partitions the space of all n-input, single-output switching functions. For $n=3$, 4, or 5, the required number of prototypes is 3, 8, or 48, respectively. As n becomes larger, the required number of prototypes increases as $2^{2^n-(n+1)^2}$ and the number of different functions that can be generated from a single randomly selected prototype approaches $2^{(n+1)^2}$. All transformation groups which have previously appeared in the literature on combinational switching theory are subgroups of the group considered herein, and further subdivide the prototype classes.

Computer programs have been written which identify the prototype associated with any given four-input function and almost all five-input functions. These programs have been used to find explicit prototype functions or canonical representatives for 46 out of the 48 equivalence classes for $n=5$. The relative size of each prototype class has also been estimated. The programs provide the logic designer with a constructive procedure for selecting an encoding transformation which can realize the desired function from its prototype. These design algorithms are based on iterated computation of the discrete Fourier transform and represent the first constructive application of abstract harmonic analysis to the synthesis of combinational logic.

Index Terms —**Affine group, combinational logic, equivalence classes, Fourier transform, large-scale integration, switching theory.**

Manuscript received October 28, 1968. The work described in this paper was supported by the 1967 Independent Research Program of Sylvania Electronic Systems. This paper was presented at the 9th Annual Symposium on Switching and Automata Theory, Schenectady, N. Y., October 1968.

The author was with Sylvania Electronic Systems, Needham Heights, Mass. 02194. He is now with Honeywell EDP, Waltham, Mass. 02154.

Functional Equivalence Under Affine Groups

Algebraic and Physical Representations

LET X represent the n-dimensional vector space (modulo two) over the two-element field $GF(2)=\{0, 1\}$. Elements of X will be represented as binary row vectors $x=(x_1, x_2, \cdots, x_n)$. Let $F: X \rightarrow Z$ (henceforth denoted by F) represent the space of all two-valued functions with domain X and range $Z=\{0, 1\}$. Let $x(i)$ represent the element of X which is also the binary representation of i, $0 \leq i < 2^n$, denote $f(x(i))$ by f_i, and let x' denote the column vector which is the transpose of the (row) vector x.

This section will discuss the action of affine transformations on the direct sum $X+Z$. This direct sum is a vector space of dimension $n+1$ over the field $GF(2)$. An affine transformation on X can be represented as $T(x)=xA+b$ where A is an $n \times n$ matrix over $GF(2)$, b is an arbitrary vector in X, and addition is modulo two. Such a transformation is in the affine group if and only if it is 1 to 1, which means that A must be nonsingular (mod 2).

An affine transformation on Z can be represented in the same way. However, since Z is a one-dimensional space over $GF(2)$, the only nonsingular linear transformation on Z is the identity. Therefore, the affine transformation group consists of only two elements—the identity and addition of