

FROM BEHAVIOR TO STRUCTURE

P. H. Roosen-Runge
© 2000

Suppose all we know about a system S is its behavior in response to stimulation or input. The only aspect of S which is specified is a function, call it B_S , which maps an input sequence i into an output sequence $B(i)$, with the interpretation that $B_S(i)$ represents the behavior or response of S for the given input. (This assumes that the system is *deterministic*: each input sequence determines a specific output sequence.)

The school of psychology known as Behaviorism believed that such behavioral functions were all a psychologist could or needed to know about animal behavior. This position rested on the belief that there was no sound way to reason from observable behavior to unobservable internal psychological structure, and hence psychology ought to give up theorizing about such structures.

In formal terms, the behaviorists were quite wrong. It is, in fact, possible to argue rigorously from external behavior to a unique minimal internal causal structure (in the form of a transition function)—a structure which is just sufficient to account completely for a particular behavioral function of the system. Furthermore, the more detailed the behavior, the more detailed the internal structure constructed from the behavior, and any two consistent behavioral descriptions can be combined to generate a single causal structure which “explains” them both. (In another sense, the behaviorists were right!—the construction is reversible, so minimal state descriptions of systems can be translated into completely equivalent totally behavioral descriptions, and we could, in principle, dispense with the state description.)

The *Myhill-Nerode Theorem* establishes the existence of a minimal state description for any finite-state (recognition) automaton and provides a construction of such a description directly from the input sequences recognized by the automaton. An elegant approach to its proof was developed originally by Rabin and Scott, in a famous paper, “Finite Automata and Their Decision Problems” [*IBM J. of Sys. Dev.*, 1960]. This is essentially the approach used here, but generalized to arbitrary transducers with no restriction on the number of states, and with an explicit algebraic formulation of what is required in order that the total system behavior (in the form of a mapping of input sequences to output sequences) can be realized by a causal state-transition function.

Our first task is to investigate ways of classifying distinct behaviors as *equivalent*:

To begin with, we stipulate that the inputs to the system come from a specific input set I , and the outputs from a specific behavioral repertoire R . We let I^* and R^* represent the sets of all possible sequences of elements of I and R , respectively; in algebraic language, I^* and R^* are the *free semi-groups generated by* I and R . I^* and R^* include the sequence of length 0 which we will denote by ϵ .

What assumptions can we make about the general structure of B_S ? We will assume that the system's behavior obeys the **causality principle**:

Future inputs cannot influence the response to current inputs; or to put it another way, observing more inputs to the system cannot alter the responses already observed.

In formal terms, this means that if $B_S(xw) = z$, then $z = B_S(x)y$ for some suffix y which we denote by $B_S(xw) / B_S(x)$. Then

$$B_S(xw) = B_S(x) (B_S(xw) / B_S(x))$$

To simplify the presentation that follows, we will also assume that

$$B_S(\epsilon) = \epsilon; \text{ that is, no output if no input.}$$

And since we will only consider a single fixed system S , the behavior function does not need to be indexed by the system, and we can just write $B(x)$.

Let E be any equivalence relation on I^* . If E satisfies the condition

$$x E y \text{ implies } xw E yw \text{ for all } w \text{ in } I^*,$$

then we will call E a *right-congruence*.

For each equivalence relation, there is a corresponding partition of I^* into equivalence classes, and vice-versa, for every partition, there is an equivalence relation whose equivalence classes are the elements of the partition. It will be convenient not to make too much of a distinction between a partition and its corresponding equivalence relation; in fact, we will use the same name for both. The definition of right-congruence can be extended directly to partitions as follows:

A partition P is a right-congruence if whenever x and $y \in X \in P$, $xw \in Y \in P$ implies $yw \in Y$, for all $w \in I^*$.

Now we classify input sequences in terms of the following relation:

$$x \sim y \text{ if for all } w \in I^*, B(xw) / B(x) = B(yw) / B(y).$$

This is obviously an equivalence relation; it is also an right-congruence. To show this, we need to show that

$$\text{if } x \sim y \text{ then for all } w \in I^*, xw \sim yw.$$

That is, for all $z \in I^*$,

$$B(xwz) / B(xw) = B(ywz) / B(yw).$$

$$\begin{aligned} \text{But } B(xwz) &= B(xw) (B(xwz) / B(xw)) \\ &= B(x) (B(xw) / B(x)) (B(xwz) / B(xw)), \end{aligned}$$

and $B(ywz) = B(y) (B(yw)/ B(y)) (B(ywz)/ B(yw))$.

Since $x \sim y$, $B(xwz) / B(x) = B(ywz) / B(y)$,
and so

$$(B(xw)/ B(x)) (B(xwz)/ B(xw)) = (B(yw)/ B(y)) (B(ywz)/ B(yw)).$$

But $B(xw)/ B(x) = B(yw)/ B(y)$,

so $B(xwz)/ B(xw) = B(ywz)/ B(yw)$,

and $xw \sim yw$.

Let S be the partition of I^* induced by the relation \sim . This set of equivalence classes gives us the crucial ingredient for reconstructing S as a *transducer*, that is, as a state-transition system with an output function .

Define $\text{Trans} : S \times I \rightarrow S$ to be the function

$\text{Trans}(\text{state}, x) =$ the element of S which contains the sequence zx for some z state.

Why does it not matter which z we pick? Because the fact that \sim is a right-congruence implies that for all $z \sim \text{state}$, all sequences zx belong to the same equivalence class in S . So Trans is a well-defined function.

To bring out the connection between a state and the input sequences it contains, we represent a state as $[x]$ where x is an arbitrary sequence in that state.

We can then rewrite the definition for Trans as

$$\text{Trans} ([z], x) = [zx].$$

We still need to verify that Trans works properly with respect to input sequences, namely that

$$\text{Trans} ([z], xy) = \text{Trans} (\text{Trans} ([z], x), y)$$

for all input sequences x , and y . But this follows from the fact that concatenation of sequences is associative:

$$\begin{aligned} \text{Trans} ([z], xy) &= [z(xy)] = [(zx)y] = \text{Trans} ([zx], y) \\ &= \text{Trans} (\text{Trans} ([z], x), y). \end{aligned}$$

The argument so far works for **any** right congruence. Why then did we choose S as the set of states for S ? Because this guarantees that the transition system we have constructed satisfies two important properties:

- (1) it can be given an output function $\text{Output} : S \times I \rightarrow R^*$ which reproduces exactly the behavior of S ;

(2) it is minimal (in the cardinality of its state set) among transition systems satisfying (1).

To establish (1), we define $Output$ as follows:

$$\begin{aligned} Output([]D) &= \epsilon \\ Output([z], i) &= B(zi)/B(z), \text{ for } i \in I. \end{aligned}$$

For this to work, we have to show that the output function is well-defined; that is, the right-hand side is the same regardless of what sequence in $[z]$ is chosen as its representative. But that follows immediately from the definition of B .

Define $Output^*(x) = B(x)$,
 $Output^*(i) = Output([], i)$, for $i \in I$,
 $Output^*(xi) = Output^*(x) Output^*([x], i)$.

Then $Output^*(x) = B(x)$ for all $x \in I^*$. The proof rests on induction on the length of x . For if $x \in I$,

$$Output^*(x) = Output([], x) = B(x)/B() = B(x).$$

Otherwise,

$$\begin{aligned} Output^*(x) &= Output^*(yi) = Output^*(y) Output^*([y], i), \text{ for } x = yi, \\ &= B(y) (B(x)/B(y)) = B(x). \end{aligned}$$

What we have shown is that there is a way of defining $Trans$ and $Output$ functions such if we take $[]$ as the initial state of the transition system, then the concatenated outputs from the successive states reached by an input sequence x are exactly the behavioral responses $B(x)$ to that input by the system S . This establishes (1).

To show (2), let T be any transducer ($Trans_T, Output_T, initial$), and define

$$\begin{aligned} Output_T^*(x) &= Output_T(Trans_T(initial, x)), \\ Output_T^*(xi) &= Output_T^*(x) Output_T(Trans_T(initial, x), i). \end{aligned}$$

(The outputs of a transducer are only defined for state transitions which begin in its initial state, so we can assume, without loss of generality, that for all states s in $States_T$, $Trans_T(initial, x) = s$ for some x .)

Then if $Output_T^* = B$, we can construct a total surjective (onto) function

$$g: States_T \rightarrow S.$$

Therefore the cardinality of $States_T$ is at least as large as the cardinality of S . Since this holds for any T , S is a **minimal** state set for any transducer which reproduces the behavior of S .

To construct the function g , we use the states of T to classify inputs:

define $x \sim_T y$ if $Trans_T(\text{initial}, x) = Trans_T(\text{initial}, y)$.

Then \sim_T is obviously a right-congruence, since $x \sim_T y$ implies

$Trans_T(\text{initial}, xw) = Trans_T(\text{initial}, yw)$, for all w .

Furthermore, there is an obvious correspondence between the equivalence classes of $Trans_T$ and the states of $States_T$; namely, if a state

$Trans_T(\text{initial}, x) = s$ for some input sequence x ,

let $f(s) = [x] \sim_T$. Then f is surjective. (Why?)

Now the crucial point: if $x \sim_T y$ then $x \sim_T y$. To show this, we show that

$Output_{T^*}(xw)/Output_{T^*}(x) = Output_{T^*}(yw)/Output_{T^*}(y)$.

The proof is by induction on the length of w .

For if $Trans_T(\text{initial}, x) = Trans_T(\text{initial}, y)$, then

$Output_T(Trans_T(\text{initial}, x), i) = Output_T(Trans_T(\text{initial}, y), i)$

so

$Output_{T^*}(xi)/Output_{T^*}(x) = Output_{T^*}(yi)/Output_{T^*}(y)$,

for $i \leq l$.

Suppose that by the induction hypothesis,

$Output_{T^*}(xw)/Output_{T^*}(x) = Output_{T^*}(yw)/Output_{T^*}(y)$,

for w of length n .

Then since

$Trans_T(\text{initial}, xw) = Trans_T(\text{initial}, yw)$,

$Output_T(Trans_T(\text{initial}, xw), i) = Output_T(Trans_T(\text{initial}, yw), i)$

and

$Output_{T^*}(xwi)/Output_{T^*}(xw) = Output_{T^*}(ywi)/Output_{T^*}(yw)$,

for all $i \leq l$, we can conclude that

$(Output_{T^*}(xw)/Output_{T^*}(x))(Output_{T^*}(xwi)/Output_{T^*}(xw))$
 $= (Output_{T^*}(xwi)/Output_{T^*}(x))$

$$\begin{aligned}
 &= (\text{Output}_{\mathcal{T}^*}(yw)/\text{Output}_{\mathcal{T}^*}(y))(\text{Output}_{\mathcal{T}^*}(ywi)/\text{Output}_{\mathcal{T}^*}(yw)) \\
 &= (\text{Output}_{\mathcal{T}^*}(ywi)/\text{Output}_{\mathcal{T}^*}(y)),
 \end{aligned}$$

which proves that if the equality holds for all w of length n , it holds for all w_i of length $n+1$; so it holds for all w .

But since $x \sim_{\mathcal{T}} y$ implies $x \sim y$, the partition associated with $\sim_{\mathcal{T}}$ refines the partition \mathbf{S} associated with \sim ; every element $[x] \sim_{\mathcal{T}}$ is contained in exactly one element $X \in \mathbf{S}$. Call this element $h([x])$. Note that h is total and surjective (why?). Now define $g = h \circ f$ on $\text{States}_{\mathcal{T}}$. Since f is surjective and h is total on the equivalence classes of $\sim_{\mathcal{T}}$, g is defined for every such class, and since h is surjective, so is g .

We have now established that there exists a surjective function g from $\text{States}_{\mathcal{T}}$ to \mathbf{S} and thus the cardinality of \mathbf{S} must be less than or equal to the cardinality of $\text{States}_{\mathcal{T}}$. But \mathcal{T} was an arbitrary transducer realizing the behavior function B . So $(\mathbf{S}, \text{Trans}, \text{Output})$ is a minimal transducer realizing B .

Thus, starting only with a behavioral description of \mathbf{S} as a system which maps inputs to outputs, we have shown that it has an equivalent structural description as a transducer with a minimal set of states.