The Myhill-Nerode Theorem

Earlier in this course, we discussed indistinguishability between states of DFAs, leading up to a procedure to generate a minimal DFA. However, we left the details of the proof open. Here, we will cover them, along with a theorem that will allow us to reason more generally about languages.

Distinguishability

Recall that we defined two states of an automaton to be distinguishable as follows:

Definition 1. *In a DFA $D = (\Sigma, Q, \delta, q_0, F)$, states $p, q \in Q$ are* distinguishable *if there exists a string $w \in \Sigma^*$ such that $\delta^*(p, w) = p'$ and $\delta^*(q, w) = q'$ and either $p' \in F, q' \notin F$ or $p' \notin F, q' \in F$.*

Indistinguishability is defined similarly:

Definition 2. *In a DFA $D = (\Sigma, Q, \delta, q_0, F)$, states $p, q \in Q$ are* indistinguishable *if for all $w \in \Sigma^*$, $\delta^*(p, w) = p'$ and $\delta^*(q, w) = q'$ and either $p' \in F, q' \in F$ or $p' \notin F, q' \notin F$.*

We can adapt this notion of distinguishability to instead reason about languages directly.

Definition 3. *Given a language L, strings w and x are* distinguishable *(relative to L) if there exists a string $y \in \Sigma^*$ such that either $wy \in L$ and $xy \notin L$, or $wy \notin L$ and $xy \in L$.*

Likewise, we can also define indistinguishability for strings:

Definition 4. *Given a language L, strings w and x are* indistinguishable *(relative to L) if for all strings $y \in \Sigma^*$, either $wy \in L$ and $xy \in L$, or $wy \notin L$ and $xy \notin L$.*

In fact, this notion of indistinguishability is quite strong, in that it defines an equivalence relation:

Theorem 1. *Let L be a language and define the relation $\equiv_L$ on pairs of strings, such that $x \equiv_L y$ if and only if x and y are indistinguishable relative to L. For any language L, $\equiv_L$ is an equivalence relation.*

Proof. Let L be an arbitrary language. We need to show that $\equiv_L$ is reflexive, symmetric, and transitive.

- Let $x \in \Sigma^*$. Then for any $w \in \Sigma^*$, $xw \in L$ if and only if $xw \in L$, so x is indistinguishable from x.
- Let $x, y \in \Sigma^*$ such that $x \equiv_L y$ and let $w \in \Sigma^*$ be arbitrary. Then, because $x \equiv_L y$, either xw and yw are both in L, or they are both not in L. Either way, the condition for $y \equiv Lx$ is satisfied.
- Let $x, y, z \in \Sigma^*$ such that $x \equiv_L y$ and $y \equiv_L z$ – we seek to show that $x \equiv_L z$. Let $w \in \Sigma^*$ be arbitrary. There are two cases:
  - $xw \in L$: Then because $x \equiv_L y$, $yw \in L$. But because $y \equiv_L z$, we also have $zw \in L$.
  - $xw \notin L$: Then because $x \equiv_L y$, $yw \notin L$. But because $y \equiv_L z$, we also have $zw \notin L$.

  In both cases, $xw \in L$ if and only if $zw \in L$. □

Why does this relation matter? Well, it turns out that we can use it to produce a very robust method of determining whether L is regular.

Theorem 2 (Myhill-Nerode Theorem). *Let $L \subseteq \Sigma^*$ be a language. L is regular if and only if $\equiv_L$ has finitely many equivalence classes. Moreover, the number of classes is exactly the number of states in a minimal DFA recognizing L.*

Before we prove this theorem, we demonstrate its significance. We have tools like the pumping lemma, which show that certain languages *cannot* be regular. However, there are languages which satisfy the conditions of the pumping lemma, yet are still not regular. Consider the following language:

$$L = \{a^i b^j c^k \mid i, j, k \geq 0 \text{ and if } i = 1, \text{ then } j = k\}$$

L satisfies the condition of the pumping lemma with $p = 1$. However, it is not regular. Consider the language $L' = L(ab^*c^*)$. Note that if L were regular, then by closure under intersection, so too would be $L \cap L' = \{ab^i c^i \mid i \geq 0\}$. But we can show that the latter is not regular via the pumping lemma.

So the pumping lemma gives what we might call a "one-sided" decision procedure: we can use it to categorically declare a language not regular, but can't use it to say that a language is regular. In contrast, the Myhill-Nerode theorem provides an exact characterization of whether or not a given language is regular.

Proof. Let L be an arbitrary language. We prove both directions of the implication.

Suppose L is regular. Then there is a DFA $(Q, \Sigma, \delta, q_0, F)$ which recognizes it. Recall that the extended transition function δ^* gives the final destination, when we give it a starting state and a string to process. We claim that for any $w, x \in \Sigma^*$, if $\delta^*(q_0, w) = \delta^*(q_0, x)$, then $w \equiv_L x$. This is because after processing these two strings, the DFA will end up at the same state, after which point their computations on any further extension will be identical. This means that the number of equivalence classes for $\equiv_L$ is at most the number of states in Q, which is finite.

In the reverse direction, suppose that $\equiv_L$ has only finitely many equivalence classes. We construct one state for each of these equivalence classes. For any string $x \in \Sigma^*$, let $[x]$ denote the equivalence class corresponding of x. Then the transition $\delta([x], a)$ is simply the equivalence class $[xa]$. By the definition of $\equiv_L$, this choice is well-defined, i.e., we get the same value regardless of which element of $[x]$ we chose as our representative (this makes $\equiv_L$ a *congruence* with respect to this operation). State $[x]$ will be final if and only if $x \in L$; again, this does not depend upon which member of the equivalence class we chose. Finally, we take the initial state to be $[\varepsilon]$. This defines a DFA whose language is exactly L, and which has exactly as many states as there are equivalence classes of $\equiv_L$.

To show that there can be no smaller DFA, we argue that any DFA accepting L must have at least one state per equivalence class of $\equiv_L$. Suppose towards contradiction that such a DFA D exists with strictly fewer states than the number of equivalence classes. Then by the pigeonhole principle, there would be two distinct equivalence classes $[x]$ and $[y]$ corresponding to the same state q in D. Since x and y are in different equivalence classes, they must be distinguishable. This means that there exists a string w such that exactly one of xw and yw is in L. But we have that:

$$\delta^*(q_0, xw) = \delta^*(\delta^*(q_0, x), w) = \delta^*(q, w) = \delta^*(\delta^*(q_0, y), w) = \delta^*(q_0, yw).$$

This is a state which must be both simultaneously accepting and rejecting, which is a contradiction. □

Observe that in the second half of the above proof, we somewhat explicitly constructed a *minimal automaton* recognizing L. We now have the tools to demonstrate its uniqueness (up to renaming of states), which we had previously postponed showing.

Theorem 3. *Let L be a regular language and let $D = (Q, \Sigma, \delta, q_0, F)$ be any minimal DFA recognizing L. Then for every state $q \in Q$, the set $S_q = \{w \mid \delta^*(q_0, w) = q\}$ is an equivalence class of $\equiv_L$.*

Proof. The proof of the Myhill-Nerode theorem demonstrates that it is impossible for any two strings in different equivalence classes to be sent to the same state. So all we need to show is that no two strings belonging to the same equivalence class can be sent to distinct states in D. Putting these two properties together shows that the sets S_q correspond exactly to equivalence classes of $\equiv_L$.

Suppose towards contradiction that D contains two states q, r, such that there are strings $x, y \in \Sigma^*$ satisfying:

- $\delta^*(q_0, x) = q$,
- $\delta^*(q_0, y) = r$, and
- $x \equiv_L y$.

Since x and y are indistinguishable relative to L, the states q and r must be indistinguishable (you should take a minute to convince yourself that this is true!). However, if D contains two indistinguishable states, it cannot be minimal – we can combine the two to produce a smaller DFA recognizing the same language. This gives the desired contradiction.

Putting this all together: If D is minimal, then the sets S_q which record "all strings which are valid transitions from q_0 to q" for all states $q \in Q$ are exactly the equivalence classes under $\equiv_L$. Since the transitions between states must also capture the structure $\delta([x], a) = [xa]$, there is only one possible DFA of this size. □

Acknowledgments

This reading was written by Dr. Shawn Ong.