**CS351**

## Proving Languages not to be Regular

Before we show how languages can be proven not regular, first, how would we show a language is regular?

Although regular languages and automata are quite powerful in what they can represent, there are many problems that cannot be solved with a DFA. This may not seem to be the case from what we've done so far! But we will soon see many simple examples that are not regular languages.

## The Pumping Lemma

Our technique to prove nonregularity comes from a theorem called the Pumping Lemma. This theorem states that all regular languages have a special property. If we can show that a language does not have this property, then the language cannot be regular. The property states that all strings in the language can be "pumped" if they are at least as long as a special value, the **pumping length**. This means that each such string contains a section that can be repeated any number of times with the resulting string remaining in the language.

The pumping lemma states:

Let L be a regular language. Then there is a number p (the pumping length) where, if s is any string in L of length at least p, then s may be divided into three parts, s=xyz, satisfying the following conditions:

1. $y \neq \varepsilon$ (but x and y may be ε)
2. $|xy| \leq p$
3. for each $k \geq 0$, $xy^k z \in L$

This theorem says that when s is divided into xyz, either x or z may be empty, but y may not be empty. Further, the pieces x and y together must have length at most p. Finally the last condition says that we can always find a nonempty string y not too far from the beginning of s that can be "pumped", i.e. it can repeat any number of times.

Note that although $y \neq \varepsilon$, for the third condition, k may equal zero, giving us $y^0 = \varepsilon$, and then xz must be $\in$ L.

Proof: First, let's consider the simplest case of a regular language L. Let's say that there are no strings in L of length at least p. In this case, the theorem becomes **vacuously** true. The three conditions hold for all strings of length at least p if there aren't any such strings. For example, if L is composed of simply the finite set { a }, then we could pick p=2 and the theorem is vacuously true because there are no strings of length at least 2.

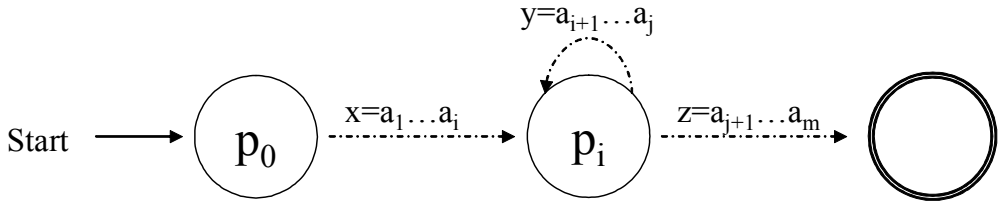
Now, suppose that L is regular. Then L=L(A) for some DFA A. Suppose that A has n states. Now, consider any string s of length n or more; say $s=a_1a_2\ldots a_m$ where $m \geq n$. For i=0,1,…n, define state $p_i$ to be $\delta\hat{}(q_0, a_1a_2\ldots a_i)$. That is, $p_i$ is the state that A is in after reading the first i symbols of s.

By the pigeonhole principle, since there are more input symbols then there are states, we must repeat some state more than once. Thus we can find two different integers i and j with $0 \leq i < j \leq n$, such that $p_i = p_j$. We can break s into s=xyz as follows:

1. $x = a_1a_2\ldots a_i$
2. $y = a_{i+1}a_{i+2}\ldots a_j$
3. $z = aj_{+1}a_{j+2}\ldots a_m$

In other words, string x takes us to state $p_i$. Then we somehow repeat back to that state; at this point we are at input symbol j, and the corresponding state is $p_j$ (where $p_i=p_j$). Finally we finish by moving to an accepting state. Since i is less than j (not less than or equal to j) this means that we must have at least one symbol for y, so y may not be empty.

We can visualize this behavior on the automaton:




If this behavior is possible, then it means that $xy^+z$ must be in the language. We've already shown that y may not be empty. Finally, if there are n states in the automaton, and $j \leq n+1$ (since y is not empty). In the extreme case, x is empty and y contains all the symbols, so $|xy| \leq n$ (in our Theorem, we used p instead of n).

**Applications of the Pumping Lemma**

To use the pumping lemma to prove that a language L is not regular:

- First assume that L is regular in order to obtain a contradiction.
- Then use the pumping lemma to guarantee the existence of a pumping length p such that all strings of length at least p can be pumped
- Find a string s in L that has length p or greater but cannot be pumped
    - This is demonstrated by considering all ways of dividing s into x,y, and z and showing that one or more of the conditions of the pumping lemma are violated
    - Typically we will show that condition 3 is violated
- Since the existence of s contradicts the pumping lemma if L was regular means that L is not regular.

The tricky part is coming up with the string s. This may take some creative thinking and trial and error, because for a non-regular language, some strings may fit the pumping lemma conditions, while others won't. If a string works, you may need to pick another one until you find one that leads to the contradiction.

You can also think of this as a "game" where one side is trying to prove the language non-regular and the other trying to thwart the first player. In the game Player 1 is trying to show the language is not regular by finding a contradiction with condition 3, while Player 2 is trying to thwart Player 1:

- Player 1 picks the language L to be proved non-regular
- Player 2 picks p, the pumping length. Player 1 doesn't know what p is, so player 1 must devise a game plan that works for all possible p's (i.e. we must leave p as a variable)
- Player 1 picks string s, where s is in L
- Player 2 divides s into x,y, and z, obeying the constraints that are stipulated in the pumping lemma ($y \neq \varepsilon$, $|xy| \leq p$, $xy^*z \in L$).
- Player 1 "wins" by picking k, which may be a function of p, x, y, or z, such that $xy^kz$ is not in L.

Example 1:

Let B be the language $\{0^n1^n \mid n \geq 0\}$. Let's use the pumping lemma to show that this is not regular.

Assume that B is regular. Let p be the pumping length selected by the adversary (Player 2 above). Choose s to be the string $0^p1^p$. This string is obviously a member of B. It also has length at least p, so it is a valid choice.

The pumping lemma guarantees for a regular language that s can be split into three pieces, s=xyz, according to the constraints of the pumping lemma. Our adversary has the following choices:

1. The string y consists only of 0's. Then we pick k=2. By condition 3, $xy^2z$ should also be in B. But this results in the string xyyz. Since we added more 0's with the addition of another y, this is not in L.
2. The string y consists only of 1's. Then we pick k=2 and by the same logic as above, the string xyyz would have more 1's than 0's and this is also not in B.
3. The string y consists of 1's and 0's. Then we pick k=2 and xyyz may have the same number of 1's and 0's but now the 0's and 1's are not in the desired order (we needed to have all the 0's come before the 1's). Therefore, this string is not in B either.

Thus a contradiction is unavoidable despite how the string s is split up by the adversary. This means that language B is not regular.

In this example, finding the string s was easy because any string in B of length p or more would work. In general though, we might have to be more careful about how we select s.

Example 2:

Let C = { w | w has an equal number of 0's and 1's }. Use the pumping lemma to show that this language is not regular.

Assume that C is regular and let p be the pumping length selected by the adversary. Choose s to be the string $(01)^p$. This is clearly of length at least p and is also in the language.

Our adversary splits the string into an x, y, and z. Let's say the adversary splits it into x=ε, y=01, and z=$(01)^{p-1}$. Can we find a value k such that $xy^kz$ is not in C? If k=0 then we just get the string xz, which is the string z. z has an equal number of 0's and 1's so it is in C. If k=1 then we get the string xyz, which is also in C. If we pick k=2 then we get the string xyyz, which is also in C. No matter what value of k we pick, each resulting string is still in the language. This means that we didn't pick the right string (or that the language actually is regular).

So let's try again by picking s=$0^p1^p$. This string is also clearly of length at least p and is also in the language. The adversary breaks up the string s into xyz. But we know that since $|xy| \leq p$, then x and y must consist entirely of 0's. Based on condition 3, if we let k=2, then xyyz will have more 0's then there are 1's so this is not in C. Similarly, if we let k=0, then xz will have fewer 0's then 1's so this is also not in C (we can pick k equal to any value except 1).

But what if the adversary picks x=ε and z= ε? That is, the string y contains the entire string. Then it would seem that $xy^kz$ will still be in C, since y will contain an equal number of 0's and 1's. But since |xy| must be ≤ p this selection is not possible for y to equal $0^p1^p$, given s=$0^p1^p$.

Example 3:

Let D = {ww | w ∈ $\{0,1\}^*$}. In other words, pick w equal to any finite sequence of 0's and 1's. Then allow only those strings that have this word in it back to back. Show that D is non-regular using the pumping lemma.

Assume that D is regular and let p be the pumping length selected by the adversary. Choose s to be the string $0^p0^p$. This is clearly of length at least p and is also in the language.

Our adversary splits the string into an x, y, and z. Let's say the adversary splits it into x=ε, y=00, and z=$0^{2p-2}$. Can we find a value k such that $xy^kz$ is not in D? The answer is

no, for any value of k the resulting string is still in D. Obviously this was not a good choice of a string s. Let's pick another one.

Choose s to be the string $0^p10^p1$. This is clearly a member of D and has length of at least p. Our adversary splits the string into an x,y, and z. Once again, since $|xy| \leq p$, we must have the case that x and y consist entirely of zeros. If we pump y by letting k=2, then we now have more zeros in the first half of the string then in the second half, so the resulting string is no longer in D. Therefore, the language is not regular.

Example 4:

Let $E = \{0^i1^j \mid i > j\}$. Show this is non-regular using the pumping lemma.

Assume that E is regular and let p be the pumping length selected by the adversary. Choose s to be the string $0^{p+1}1^p$. This is clearly of length at least p and is also in the language.

Our adversary splits the string into an x, y, and z. Once again, since $|xy| \leq p$, both x and y must consist entirely of zeros. If we pump y up, then we end up with strings that are still in L. But if we pump y down by allowing k=0, then we get the string xz. Removing the y decreases the number of 0's in s. But we constructed s so that there is exactly one more zero than one. So by removing a zero, we now have the same or fewer zero's than ones, so the resulting string is no longer in E. Therefore, the language is not regular.

Example 5:

Let $F = \{ 1^{n^2} \mid n >= 0 \}$. That is, each string consists of 1's and is of length that is a perfect square:

ε, 1, 1111, 111111111 , 1111111111111111, etc. (0, 1, 4, 9, 16, 25, 36, …)

Notice that the gap between the length of the string grows in the sequence. Large members of this sequence cannot be near each other. If we subtract off the difference in length between successive elements, we get 1, 3, 5, 7, 9, 11, 13, etc. For position i where i >0, we get the difference from position i and i-1 as 2*i -1.

Assume that F is regular and let p be the pumping length selected by the adversary. Let $m = p^2$. Choose s to be the string $1^m$. This string is clearly in F and is at least of length p.

The adversary picks some strings x,y, and z. Now consider the two strings $xy^kz$ and $xy^{k+1}z$. Both of these strings must be in F. These two strings differ only by a single repetition of y, or |y| which we know must be ≤ p. However, note that the length of the strings accepted by the language grows in length by 2i-1, not by some fixed amount |y|.

Each time we pump the string we increase the value of i, so we can always find a value of i to make 2i-1 larger than |y|. Consequently, we can always pick a large enough k such that assuming $xy^kz$ is in the language, $xy^{k+1}z$ cannot be in the language because the length differential will be too small to equal to 2i-1.

**Closure Properties and DFA Minimization**

While we have seen that some languages are not regular, certain operations on regular languages are guaranteed to produce regular languages. The following theorems are referred to as the **closure properties** of regular languages. Closure refers to some operation on a language, resulting in a new language that is of the same "type" as those originally operated on, i.e., regular.

We won't be using the closure properties extensively here; consequently we will state the theorems and give some examples. See the book for proofs of the theorems.

1. Closure under Union

   Let L and M be languages over $\Sigma$. Then $L \cup M$ is the language that contains all strings that are either in L or in M.

   We have already shown how to compute R+S, where R+S is the union of two regular languages. If L=L(R) and M=L(S) then L(R+S) is the same as $L \cup M$.

   Note that if $\Sigma$ is different for L and M, then $\Sigma$ for $(L \cup M)$ is the union of the alphabets for L and the alphabets for M.

2. Closure under Complementation

   Let L and M be languages over $\Sigma$. Then $\overline{L}$ is the complement of L, which is the set of strings of $\Sigma^*$ that are not in L.

   In other words, the complement of a language is everything that is not accepted by the language over our alphabet. Here is an argument as to why complementation is closed. To complement a language:

   - First construct a DFA for that language
   - Complement the accepting states of the DFA

   Note that this requires there be no missing transitions in the DFA. If the automaton "dies" on missing edges, these states will be missing from the complemented DFA (which should be accepting states).

3. Closure under Intersection

Let L and M be languages over $\Sigma$. Then $L \cap M$ is the language that contains all strings that are both in L and M.

To show closure under intersection, note DeMorgan's law which states:

$$L \cap M = \overline{\overline{L} \cup \overline{M}}$$

We have already shown closure under union and complementation, therefore we also have closure under intersection.

4. Closure under Difference

Let L and M be languages over $\Sigma$. Then L – M, the difference of L and M, is the set of strings that are in L but not in M.

To show closure, note that: $L - M = L \cap \overline{M}$ . Since we have shown closure under intersection and complementation, we also have closure under difference.

5. Closure under Reversal

The reversal of a string $a_1a_2\ldots a_n$ is the string written backwards, that is, $a_na_{n-1}a_{n-2}\ldots a_1$. We will use $w^R$ to denote the reversal of string w. For example, $1011^R$ is 1101.

The reversal of a language L, written $L^R$, is the language consisting of the reversals of all its strings. The reversal of a regular language produces another regular language. We can show this informally using an automaton A for L(A):

- Reverse all the edges in the transition diagram for A
- Make the start state of A be the only accepting state for the new automaton
- Create a new start state $p_0$ with transitions on ε to what were the original accepting states of A

This creates a reversed automaton for A, that accepts a string iff A accepts $w^R$.

6. Closure under Kleene Star

We have already shown this for regular expressions in the construction of an equivalent ε-NFA for the star operation.

7. Closure under concatenation

We have already shown this for regular expressions in the construction of an equivalent ε-NFA for concatenation.

8. Homomorphisms

We will show closure under homomorphisms, but first we should say what a homomorphism is.

Given a language L with alphabet $\sum 1$, A homomorphism h is defined from some alphabet $\sum 2$. For symbol(s) a $\in \sum 1$, h(a) is some string from $\sum 2$. We apply h to each symbol of a word w from L and concatenate the results in order to get a new string. h(L) is the homomorphism of every word in L.

Example: h is the homomorphism from the alphabet {0,1,2} to the alphabet {a,b} defined by h(0) = a, h(1) = ab, h(2) = ba.

h(0120) = aabbaa
h(21120) = baababbaa
$h(01^*2) = a(ab)^*ba$

Theorem: If L is a regular language over alphabet $\sum$, and h is a homomorphism on $\sum$, then h(L) is also regular.

Informally, if L is turned into a regular expression, we are substituting a regular expression for some other regular expression matching the homomorphism. The result is still a regular language.

Note: we can expand homomorphisms to more general substitution, where substituting a substring in L (instead of an individual symbol) with some new string also results in a language that is regular.

9. Inverse Homomorphism

A homomorphism may be applied backwards; i.e. given the h(L), determine what L is. This is denoted as $h^{-1}(L)$, or the inverse homomorphism of L, which results in the **set** of strings w in $\sum^*$ such that h(w) is in L.

Note that while applying the homomorphism to a string resulted in a single string, we may get a set of strings in the inverse.

For example, if h is the homomorphism from the alphabet {0,1,2} to the alphabet {a,b} defined by h(0) = a, h(1) = ab, h(2) = ba.

Given L is the language {ababa} what is $h^{-1}(L)$ ?
We can construct three strings that map into ababa:
$h^{-1}$={ 022, 110, 102 }

The result of $h^{-1}$ is also a regular language (see proof in book).

**Decision Properties of Regular Languages**

Given a (representation, e.g., RE, FA, of a) regular language L, what can we tell about L?

Since there are algorithms to convert between any two representations, we can choose the representation that makes whatever test we are interested in easiest.

Membership: Is string w in regular language L?

- Choose DFA representation for L.
- Simulate the DFA on input w.

Emptiness: Is L = Ø?

- Use DFA representation for L
- Use a graph-reachability algorithm (e.g. Breadth-First-Search or Depth-First-Search or Shortest-Path) to test if at least one accepting state is reachable from the start state. If so, the language is not empty. If we can't reach a accepting state, then the language is empty.

Finiteness: Is L a finite language?

- Note that every finite language is regular (why?), but a regular language is not necessarily finite.
- DFA method:
    - Given a DFA for L, eliminate all states that are not reachable from the start state and all states that do not reach an accepting state.
    - Test if there are any cycles in the remaining DFA; if so, L is infinite, if not, then L is finite. To test for cycles, we can use a Depth-First-Search algorithm. If in the search we ever visit a state that we've already seen, then there is a cycle.

Equivalence: Do two descriptions of a language actually describe the same language? If so, the languages are called equivalent. For example, we've seen many different (and sometimes complex) regular expressions that can describe the same language. How can we tell if two such expressions are actually equivalent?

To test for equivalence our strategy will be as follows:

- Use the DFA representation for the two languages
- Minimize each DFA to the minimum number of needed states
- If equivalent, the minimized DFA's will be structurally identical (i.e. there may be different names for the states, but all transitions will go to identical counterparts in each DFA).

For this strategy to work, we need a technique to minimize a DFA.

We say that two states p and q in a DFA are **equivalent** if:

- For **all** input strings w, $\delta^{\wedge}(p, w)$ is an accepting state if and only if $\delta^{\wedge}(q, w)$ is an accepting state.

This is saying that if we have reached state p or q, then any other string we might get that will lead to an accepting state from p will also lead to an accepting state from q. Also any string we get that does not lead to an accepting state from p also does not lead to an accepting state from q. If this condition holds, then p and q are equivalent. If this condition does **not** hold, then p and q are **distinguishable**.

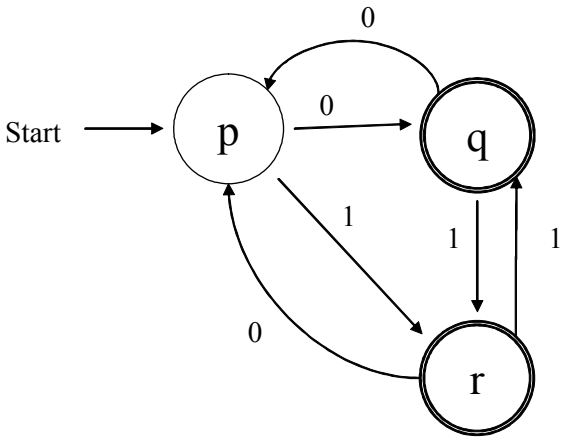
The simplest case of a distinguishable pair is an accepting state p and a non-accepting state q. In this case, $\delta^{\wedge}(p, \varepsilon)$ is an accepting state, but $\delta^{\wedge}(q, \in)$ is not an accepting state. Therefore, any pair of (accepting, non-accepting) states are distinguishable.

Here is a table-filling algorithm to recursively discover distinguishable pairs in a DFA A:

1. Given an automaton A that has states $q_0 \ldots q_n$ make a diagonal table with labels 1 to n on the rows and 0 to n-1 on the columns.
2. Initialize the table by placing X's for each pair that we know is distinguishable. Initially, this is any pair of accepting and non-accepting states.
3. Let p and q be states such that for some input symbol a, $r = \delta(p, a)$ and $s = \delta(q, a)$. If the pair (r,s) is known to be distinguishable (i.e. they have an X in their table entry) then the pair p and q are also distinguishable. Place an X for (p,q) in the table.
4. Repeat step 3 until all pairs have been examined.
5. Repeat steps 3-4 again for any empty entries in the table. If no new X's can be placed, then the algorithm is complete. Any entries in the table without an X are pairs of states that are equivalent.

Once the equivalent states have been identified, they can be combined into a single state to make a new, minimized automaton. DFA's have unique minimum-state equivalents.

Simple example: Minimize the following automaton.




Create a table:

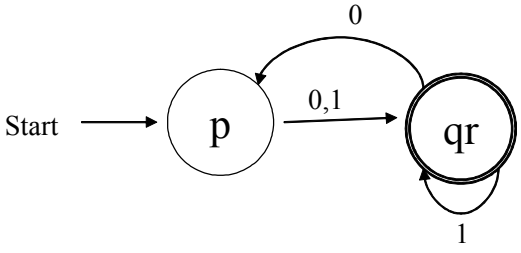
| | | |
|---|---|---|
| q | | |
| r | | |
| | p | q |

Fill an X in for states we know are distinguishable. Initially this is the final state and a non-final state:

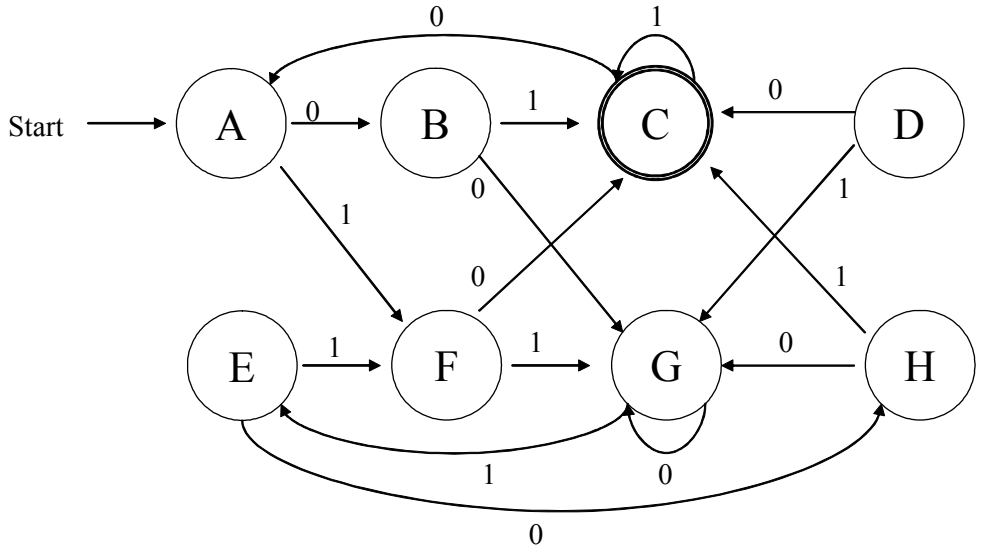
| | | |
|---|---|---|
| q | X | |
| r | X | |
| | p | q |

Now check to see if (q,r) is distinguishable. On input symbol 0, both go to state p, so they are not distinguishable on 0. On input symbol 1, r goes to q and q goes to r, giving us the pair (q,r). Looking in the table, (q,r) has no X so we don't put in X in this box. Indeed, we will never put a X in (q,r) because on input symbol 1 we will also refer to itself and it is not distinguishable.

This means that states q and r are equivalent and can be combined. This results in the following, minimized DFA:



Here is the example from the textbook:




Construct the table and initially we can populate X's to with any pair of states that contains C, since C is the sole accepting state.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| B | | | | | | | |
| C | X | X | | | | | |
| D | | | X | | | | |
| E | | | X | | | | |
| F | | | X | | | | |
| G | | | X | | | | |
| H | | | X | | | | |
| | A | B | C | D | E | F | G |

Let's say we look at pair (A,G) first. On input symbol 0, A goes to B and G goes to G. The pair (B,G) is not distinguishable so far, so we have no conclusion. On input symbol 1, A goes to F and G goes to E. (A,F) is also not distinguishable so far, so we have no conclusion and leave (A,G) blank for now.

Let's say we look at (A,H) next. On input symbol 1, H goes to C and A goes to F. The pair (C,F) is known to be distinguishable, so that means (A,H) is also distinguishable and we place an X in the box.

Let's look at (A,F) next. On input symbol 0, F goes to C and A goes to F. (F,C) is known to be distinguishable, so (A,F) is also distinguishable and we place an X in the box.

If we continue this way for all states, we can almost fill in the entire table on the first pass. This is because almost all of the states go to C, which we know to be distinguishable with everything else.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| B | X | | | | | | |
| C | X | X | | | | | |
| D | X | X | X | | | | |
| E | | X | X | X | | | |
| F | X | X | X | | X | | |
| G | | X | X | X | X | X | |
| H | X | | X | X | X | X | X |
| | A | B | C | D | E | F | G |

We can't quit yet though, we have to continue iterating until no more X's can be found. This is because when we made a check to see if a pair is distinguishable, we might have done that check before we placed an X in the box.

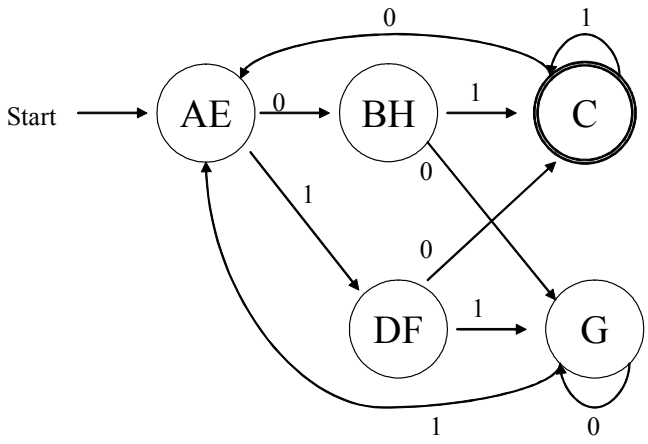
Let's examine (A,G) first. input symbol 0, A goes to B and G goes to G. The pair (B,G) we now know to be distinguishable. That means that (A,G) is also distinguishable and we can place an X in the box.

Let's examine (A,E) next. On input symbol 1, both A and E go to F so that is not distinguishable. On input symbol 0, A goes to B and E goes to H. (B,H) is not known to be distinguishable either, so we can't place an X in the box for (A,E).

If we continue to check all the unmarked states, no other states can be distinguished and we will finally be complete on the next iteration:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| B | X | | | | | | |
| C | X | X | | | | | |
| D | X | X | X | | | | |
| E | | X | X | X | | | |
| F | X | X | X | | X | | |
| G | X | X | X | X | X | X | |
| H | X | | X | X | X | X | X |
| | A | B | C | D | E | F | G |

This means that states (B,H), (A,E) and (D,E) are equivalent. These states can then be combined to produce a minimal DFA.




While we can perform the minimization technique to reduce and simplify a DFA, the reason we ended up here in the first place was to use minimization to show two DFA are equivalent. If we minimize two different DFA's and both are structurally identical, then the two DFA's (and the language they represent) are equivalent.

**Why the Minimized DFA Can't Be Beaten**

We took a DFA A and minimized it to produce an equivalent, minimal, DFA M. Could there be another DFA, N, that is unrelated to A that accepts the same language as A but yet has fewer states than M?

We can prove by contradiction that such a DFA N does not exist.

- First, run the state-distinguishability process on the states of M and N together, as if they were one DFA but with no interaction.
- The start states of M and N are indistinguishable since L(M) = L(N).
- If (p,q) are indistinguishable, then their successors on any one input symbol are also indistinguishable. This is because if we could distinguish the successors, then we could have distinguished p from q.
- M must not have an inaccessible state, or else we would have eliminated it when we produced M in the first place.
- We are assuming that N has fewer states than M. If this is so, then there are two states of M that are indistinguishable from the same state of N, and therefore indistinguishable from each other.
- But M was designed so that all states are distinguishable from each other. We have a contradiction, therefore our assumption that N has fewer states than M must be false.

The end result is that other DFA exists with fewer states that is equivalent to the minimized DFA. We can get an even stronger result, that there is a 1:1 correspondence between the states of N and M meaning that the minimized DFA is unique.