CSE491/596 Wed. 9/20/23: Non-Regular Languages

Rather than the "Pumping Lemma", we will employ the **Myhill-Nerode Theorem** (MNT) to prove non-regularity of certain languages. Although it was proved in Chicago in 1957-58 where John Myhill and Anil Nerode were students, we can claim it as Western NY heritage: Myhill was a professor at UB until his death in 1987, and Anil Nerode still teaches at Cornell past age 90(!) Nerode was my supervisor when I had a postdoc at Cornell.

Building up to the Proof

Given a DFA $M = (Q, \Sigma, \delta, s, F)$ and two strings $x, y \in \Sigma^*$, suppose $\delta^*(s, x)$ and $\delta^*(s, y)$ both give the same state q. Then for any further string $z \in \Sigma^*$, the computations on the strings xz and yz go through the same states after q. In particular, they end at the same state r.

- If $r \in F$, then $xz \in L$ and $yz \in L$, where $L = L(M)$.
- If $r \notin F$, then $xz \notin L$ and $yz \notin L$.
- Either way, $L(xz) = L(yz)$, for all z.

Suppose, on the other hand, we have strings x, y for which there exists a string z such that

$$L(xz) \neq L(yz).$$

Then M cannot process x and y to the same state. Moreover, this goes for *any* DFA M such that $L(M) = L$. In particular, every such DFA must at least *have* two states.

Now let us build some definitions around these ideas. Given any language L (not necessarily regular) and strings x, y "over" the alphabet Σ that L is "over", define:

- x and y are *L-equivalent*, written $x \sim_L y$, if **for all** $z \in \Sigma^*$, $L(xz) = L(yz)$.
- x and y are *distinctive for* L, written $x \nsim_L y$, if there exists $z \in \Sigma^*$ s.t. $L(xz) \neq L(yz)$.

Lemma 1. The relation $\sim_L$ is an equivalence relation.

Proof: We need to show that it is
- Reflexive: $x \sim_L x$ is obvious.
- Symmetric: indeed, $y \sim_L x$ immediately means the same as $x \sim_L y$.
- Transitive: Suppose $w \sim_L x$ and $x \sim_L y$. This means:
  - for all $v \in \Sigma^*$, $L(wv) = L(xv)$ and
  - for all $z \in \Sigma^*$, $L(xz) = L(yz)$.

    Because v and z range over the same span of strings, it *follows* that
  - for all $z \in \Sigma^*$, $L(wz) = L(xz)$ and $L(xz) = L(yz)$.

Hence we get:

– for all $z \in \Sigma^*$, $L(wz) = L(yz)$.

So $w \sim_L y$.

This ends the proof. ☒

Any equivalence relation on a set such as Σ^* partitions that set into disjoint *equivalence classes*. So $x \nsim_L y$ is the same as saying x and y belong to different equivalence classes.

Now say that a set S of strings is ***Pairwise Distinctive for L*** if all of its strings belong to separate equivalence classes under the relation $\sim_L$. Other names we will use are "distinctive set" and "**PD set**" for L. This is the same as saying:

- for all $x, y \in S$, $x \neq y$, there exists $z \in \Sigma^*$ such that $L(xz) \neq L(yz)$.

Thus we can re-state something we said above as:

Lemma 2. If L has a PD set S of size 2, then any DFA M such that $L(M) = L$ must process the two strings in S to different states, so M must have at least 2 states.

Note: "L has" does not mean S must be a subset of L, it just means "has by association." Now we can take this logic further:

Lemma k. If L has a PD set S of size k, then any DFA M such that $L(M) = L$ must process the k strings in S to different states, so M must have at least k states.

I've worded this to try to make it as "obvious" as possible, but actually it needs proof: Suppose we have a DFA M with $k - 1$ or fewer states such that $L(M) = L$. Then there must be (at least) two strings in S that M processes to the same state. This follows by the **Pigeonhole Principle**. [story from GLL blog]

Lemma ∞. If L has a PD set S of size ∞, then any DFA M such that $L(M) = L$ must process the strings in S to different states, so M must have at least ∞ states...but then M is not a *finite* automaton. So L is not accepted by any finite automaton...which means **L is not a regular language**. ☒

Myhill-Nerode Theorem, first half: If L has an infinite PD set, then L is not regular.

Example 1: $L = \left\{a^n b^n : n \geq 0\right\}$. $\Sigma = \{a, b\}$. $S = \left\{a^n : n \geq 0\right\} = a^*$. Let any $x, y \in S$, $x \neq y$, be given. Then there are different numbers i and j such that $x = a^i$ and $y = a^j$. Take $z = b^i$. Then $xz = a^i b^i \in L$, but $yz = a^j b^i \notin L$, because $i \neq j$. Thus $L(xz) \neq L(yz)$. Thus for all $x, y \in S$ with $x \neq y$, there exists z such that $L(xz) \neq L(yz)$. Thus S is PD for L. Since S is infinite, L is not regular, by MNT. ☒

We have proved only one direction of the Myhill-Nerode Theorem: L has an infinite PD set $\implies$ L is nonregular, but this is the direction to apply for nonregularity proofs. Those proofs can all be made to follow a "script":

Take S = __________. [Observe S is infinite---this is usually immediately clear.]

Let any $x, y \in S$ $(x \neq y)$ **be given**. Then we can write x = _______ and y = _______ where ________ [and **without loss of generality**, _________].

Take z = _______.

Then $L(xz) \neq L(yz)$ because ________________________________ ____.

Because x, y are an arbitrary pair of strings in S, this shows that S is PD for L, and since S is infinite, it follows that L is nonregular by the Myhill-Nerode Theorem.

I have colored the words **take** and **let...be given** separately to show how they express the logical quantifiers in the formal statement of this direction of MNT:

If **there exists** an infinite set S such that **for all** distinct $x, y \in S$ **there exists** $z \in \Sigma^*$ such that $L(xz) \neq L(yz)$, **then** L is nonregular.

The difference is that *you* have control of choices in the existential parts, but in the "for-all" parts you have to be prepared for all possibilities. There is a habit to use "let" in both situations, but this can be confusing. [Give humorous story about how both "let" and "any" are self-contradictory words in English, but they are OK together with "...be given."] Now let's re-do Example 1 with the script:

Example 1. $L = \{a^n b^n : n \geq 0\}$.

Take S = ____a^*______. [Observe S is infinite---this is usually immediately clear.]

Let any $x, y \in S$ $(x \neq y)$ **be given**. Then we can write x = ___a^i____ and y = ___a^j____ where ___$i \neq j$ (and it is understood that $i, j \geq 0$)_____ [and **without loss of generality**, _________].

Take z = ___b^i____.

Then $L(xz) \neq L(yz)$ because ___$xz = a^i b^i$ which is in L since the counts are equal, but $yz = a^j b^i$ which is not in L because j is different from i. .

Because x, y are an arbitrary pair of strings in S, this shows that S is PD for L, and since S is infinite, it follows that L is nonregular by the Myhill-Nerode Theorem.

Thus to prove a given L nonregular we have to "act out" the proof---and the above is our script. The first example also illustrates the optional "**w.l.o.g.**" clause.

Example 2. $L = \left\{x \in \{s,d\}^* : \#s(x) \geq \#d(x)\right\}$.

Take $S =$ __s^*__. Clearly S is infinite.

Let any $x, y \in S$ $(x \neq y)$ **be given**. Then we can write $x =$ __s^i__ and $y =$ __s^j__ where ____$i \neq j$____ and **wlog.**, ____$j < i$____.

Take $z =$ __d^i__.

Then $L(xz) \neq L(yz)$ because ___$xz = s^i d^i \in L$___. Whereas $yz = s^j\, d^i$... is not in L because wlog. $j < i$.

Because x, y are an arbitrary pair of strings in S, this shows that S is PD for L, and since S is infinite, it follows that L is nonregular by the Myhill-Nerode Theorem.

Note that this L is not the same as the language of "spears-and-dragons with unlimited saving of spears" because e.g. the string "ds" belongs to this L despite the spear coming too late in the other. But the *proof* is exactly the same. The fun is that not only do these proofs become fairly automatic once you get comfortable with the script, they are often like re-usable code.

[Here and/or with *reductions*, I used to say for fun that this can be an exception to the university rule against recycling an old answer for a new assignment, even when it was your answer. I even used to sing a relevant section of the Tom Lehrer song "Lobachevsky" which you can find linked at https://gilkalai.wordpress.com/2020/08/29/to-cheer-you-up-in-difficult-times-11-immortal-songs-by-sabine-hossenfelder-and-by-tom-lehrer/. But an upsurge in academic integrity violations made this all stop being funny about 15 years ago...]