Infinite Regular Languages and Nonregular Subsets

I want to show that every infinite regular language has a nonregular subset. (An infinite language is a language with infinitely many strings in it. $\{\mathtt{a}^n \mid n \geq 0\}$, $\{\mathtt{a}^m\mathtt{b}^n \mid m, n \geq 0\}$, and $\{a, b\}^*$ are all infinite regular languages.)

First, note that this can only be true for infinite regular languages. (Why?)

Second, infinity is big. Way big. The following lemma gives a sense how big:[1]

Lemma 1. *If A is an infinite language, then for every natural number $n \geq 0$, there exists a string $w \in A$ such that $|w| > n$.*

Proof. Let A be an infinite language over finite alphabet Σ. Let $n \geq 0$. We argue by contradiction. Assume, by way of contradiction, that there is no string $w \in A$ such that $|w| > n$. Thus, for all strings w in A, $|w| \leq n$. But there are at most $1 + |\Sigma| + |\Sigma| + |\Sigma|^2 + \cdots + |\Sigma|^n = \sum_{i=0}^{n} |\Sigma|^i$ strings of length up to n over alphabet Σ, and so $|A| \leq \sum_{i=0}^{n} |\Sigma|^i$, which contradicts A being infinite. □

It pays off to look at an example when considering whether an infinite regular language has a nonregular subset. Consider $A = \{\mathtt{a}^n \mid n \geq 0\}$. If what I claim is true, then it has a nonregular subset. But the only thing that differentiates strings in A are their length. You may think that's a big restriction, but in fact, it provides the key insight for the proof.

It turns out the pumping lemma offers some guarantees about the length of strings that must appear in a regular language. It's perhaps a non-intuitive consequence of the pumping lemma, but once you look at the pumping lemma from the right perspective, it makes sense.

I'm going to use the "simple" version of the pumping lemma, which is not the one that is given in Kozen, but that I gave in class. (The argument can be pushed through with Kozen's version of the pumping lemma, but this one is simpler to use for this result.)

Lemma 2 (Pumping Lemma, simple version). *If A is regular, then there exists a $p \geq 1$ such that for all string $w \in A$ with $|w| \geq p$, we can find x, y, z such that $w = xyz$, $|y| \geq 1$, $|xy| \leq p$, and for all $i \geq 0$, $xy^iz \in A$.*

We're in fact going to use this pumping lemma in its contrapositive form:

Lemma 3 (Pumping Lemma, simple version, contrapositive form). *Let A be a language. If for all $p \geq 1$ there is a string $w_p \in A$ with $w_p \geq p$ with the property that:*

> *for all x, y, z such that $w_p = xyz$, $|y| \geq 1$, and $|xy| \leq p$ we can find $i \geq 0$ such that $xy^iz \notin A$*

then A is nonregular.

[1] When X is a set, $|X|$ is the size of X; when w is a string, $|w|$ is the length of w.

Now let's prove the claim.

Theorem 4. *Every infinite regular language has a nonregular subset.*

Proof. Let A be an infinite regular language.

We are going to construct a subset N of A that we will prove nonregular.

We construct N by the following process: we pick w_0 to be an arbitrary string of A. We pick w_1 an arbitrary string of A such that $|w_1| > 4|w_0|$. (We know at least one such string exists by Lemma 1.) We pick w_2 an arbitrary string of A such that $|w_2 > 4|w_1|$. Generally, having picked $w_0, w_1, w_2, \ldots, w_n$, we pick w_{n+1} an arbitrary string of A such that $|w_{n+1}| > 4|w_n|$. Let $N = \{w_0, w_1, \ldots, w_n, \ldots\}$ be the set of all the strings picked in this way.[2]

Clearly, $N \subseteq A$. I claim N is nonregular.

I'm going to show this using the contrapositive form of the pumping lemma above. Let $p \geq 1$. I'm going to take w_p to be the shortest string w in N such that $w \geq p$. By Lemma 3, if I can show that no matter what x, y, z we choose such that $w = xyz$, $|y| \geq 1$, and $|xy| \leq p$, I can find $i \geq 0$ such that $xy^2z \notin N$, then I can conclude that N is nonregular.

So let x, y, z be such that $w = xyz$, $|y| \geq 1$, and $|xy| \leq p$. I'm going to show that $xy^2z \notin N$.

Note that $|w| = |xyz| < |xy^2z| = |x| + 2|y| + |z| \leq |w| + 2|w| + |w| = 4|w|$.

Thus, $|w| < |xy^2z| \leq 4|w|$.

But by construction, the shortest string v in N such that $|v| > |w|$ is such that $|v| > 4|w|$.

Thus, $|v| > 4|w| \geq |xy^2z| > |w|$. Since v is the shortest string in N of length greater than that of w, xy^2z (which is longer than w but shorter than v) cannot be in N.

Therefore, N is nonregular. □

Applying the construction in the proof to the language $\{\mathtt{a}^n \mid n \geq 0\}$ and taking w_0 to be $\mathtt{a}$, we can construct

$$N = \{\mathtt{a}, \mathtt{a}^5, \mathtt{a}^{21}, \mathtt{a}^{85}, \ldots\}$$

or more generally, $N = \{\mathtt{a}^{\sum_{i=0}^{n} 4^i} \mid n \geq 0\}$, a nonregular subset of $\{\mathtt{a}^n \mid n \geq 0\}$.

[2] More formally, define the infinite family of sets $N_0, N_1, \ldots$ by taking

$$\begin{aligned} N_0 &= \{w_0\} && \text{for some } w_0 \in A \\ N_{i+1} &= N_i \cup \{w_{i+1}\} && \text{for some } w_{i+1} \in A \text{ with } |w_{i+1}| > 4|w| \text{ for all } w \in N_i \end{aligned}$$

and define $N = \bigcup_{i \geq 0} N_i$.