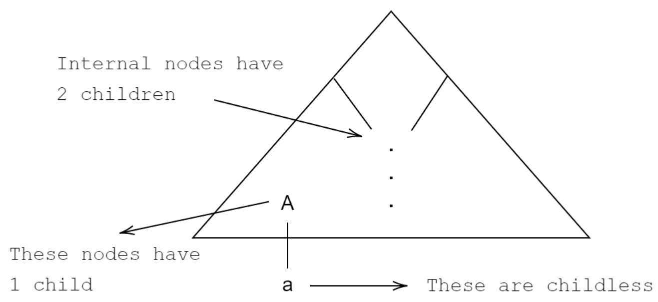
Proof pumped lemma for context-free languages

Nima Chitgar

A parse tree for any CFG is of the following form:

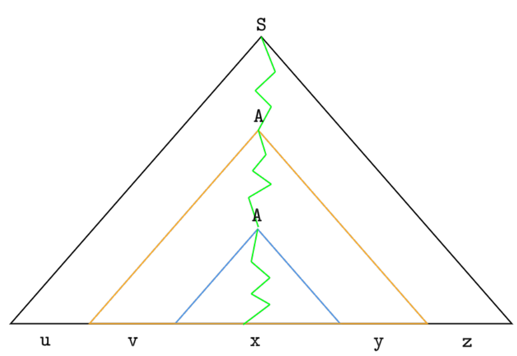
- Each internal node has 2 children. (Each variable produces two other variables)
- At the very bottom of the tree, each variable produces exactly one child and these children are childless.



For the first point, we desire to prove that at least 1 variable will repeat from top to bottom. If we can prove this for a path that goes from roof-to-leaf (e.g. green path), this will meet the requirement.

Assume that the height of the tree is #variables + 1. This will ensure that the variable A must repeat because of the pigeon-hole-principle. This means that there are more variables going from top-to-bottom than there actually are. (height is #variables + 1, but there are only #variables, so there must be a repetition)

Let A be the variable that repeats. Because we know that the parse tree is a binary tree, we can say that the length of the tree is at least $w \geq 2^{\#\text{var}+1}$. As A is a variable that repeats, we know that it can't be a start variable so it must produce a part of the string. Below you can see that we can split our length w into 5 pieces, namely $uvxyz$.



We can say that A will eventually produce x. We can write this as

$$A \Rightarrow^* x$$

and will also produce

$$A \Rightarrow^* vAy$$

This tells us something important about A: it will also produce vAy,$vvAyy$,$vvvAyyy$,...

Thus,

$$A \Rightarrow^* v^i A y^i \text{ for any } i \geq 0.$$

The start variable S will also produce

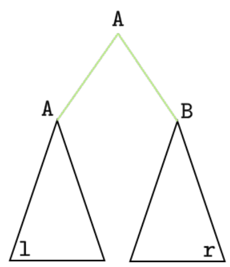
$$S \Rightarrow^* uAz$$

This will give us eventually

$$S \Rightarrow^* uv^i xy^i z \text{ for any } i \geq 0$$

and the first point of our definition is proved.

Now we need to prove that $|vy| > 0$. The shortest $|vy|$ possible is that when two A's are as close as possible to each other. You can see this on the figure below.



As B is not the start variable, it creates a non-empty string. Now, part r of the figure above corresponds with part y and part l corresponds with x. So in this 'worst-case' scenario, if y is non-empty, v is empty and vice-versa. So $|vy| \geq 1$ as both of them can't be simultaneously empty and point two of our definition is proved.

The last point is to prove that $|vxy| \leq p$. Imagine that our string is huge, if we assure that the orange region vxy is at least $2^{\#\text{var}+1}$, then we can guarantee that there must be a repetition within this region. For creating the longest string, the variables must be very far from the bottom, but even in this case $p \geq 2^{\#\text{var}+1} = |vxy|$.