TOC PROBLEM SET-7

SIDDHANT CHAUDHARY
BMC201953

Consider the following grammar G over the alphabet $\{a, b\}$.

$$
\begin{aligned}
S &\to aBS \mid bAS \mid \epsilon \\
A &\to a \mid bAA \\
B &\to b \mid aBB
\end{aligned}
$$

The first 3 questions use this grammar.

1. Prove that every word generated by G has equal number of a's and b's.

Solution. Suppose $S \xrightarrow{*} \alpha$ where $\alpha \in (N \cup \Sigma)^*$, where N is the set of non-terminals of G. We will show that

$$|\alpha|_a + |\alpha|_A = |\alpha|_b + |\alpha|_B$$

by induction on the length of the derivation. For the base case, suppose the length of the derivation is 0. The only possibility is that $\alpha = S$, and clearly the base case is true, because the word S does not contain any terminal or the symbols A, B. So, suppose the statement is true for any derivation of length n, and let

$$S \xrightarrow{*} \alpha$$

be a derivation of length $n+1$, i.e suppose the derivation is

$$S \xrightarrow{*(n \text{ steps})} \alpha' \to \alpha$$

By the induction hypothesis, we know that

$$|\alpha'|_a + |\alpha'|_A = |\alpha'|_b + |\alpha'|_B$$

Now, suppose the step $\alpha' \to \alpha$ involves one of the production $S \to aBS$ or $S \to bAS$ or $S \to \epsilon$. Observe that, in either of these productions, we are always adding one of a, A and one of b, B, i.e in any of these productions, we have the following two equations:

$$
\begin{aligned}
|\alpha|_a + |\alpha|_A &= |\alpha'|_a + |\alpha'|_A + 1 \\
|\alpha|_b + |\alpha|_B &= |\alpha'|_b + |\alpha'|_B + 1
\end{aligned}
$$

and hence in any case we see that

$$|\alpha|_a + |\alpha|_A = |\alpha|_b + |\alpha|_B$$

Now suppose the step $\alpha' \to \alpha$ involves the production $A \to a$. In that case, we have

$$
\begin{aligned}
&|\alpha|_a + |\alpha|_A = (|\alpha'|_a + 1) + (|\alpha'|_A - 1) = |\alpha'|_a + |\alpha'|_A \\
&|\alpha|_b + |\alpha|_B = |\alpha'|_b + |\alpha'|_B
\end{aligned}
$$

Date: October 2020.

and once again we see that

$$|\alpha|_a + |\alpha|_A = |\alpha|_b + |\alpha|_B$$

Next, suppose the step $\alpha' \to \alpha$ involves the production $A \to bAA$. In that case, the equations we have are

$$|\alpha|_a + |\alpha|_A = |\alpha'|_a + |\alpha'|_A + 1$$
$$|\alpha|_b + |\alpha|_B = |\alpha'|_b + |\alpha'|_B + 1$$

and hence in this case as well we have

$$|\alpha|_a + |\alpha|_A = |\alpha|_b + |\alpha|_B$$

Finally, if the step $\alpha' \to \alpha$ involves one of the productions $B \to b$ or $B \to aBB$, then we can apply the same argument as above. Hence, the induction proof is complete. So, it follows that if the word $\alpha \in \Sigma^*$ is generated by G, it will be true that

$$|\alpha|_a + |\alpha|_A = |\alpha|_b + |\alpha|_B$$

which implies that $|\alpha|_a = |\alpha|_b$, because α cannot contain the non-terminals A, B. This proves the claim.

Before doing the next problem, I will prove the hint given in the footnotes, and I will do this in two steps.

Lemma 0.1. *Let $w \in \{a, b\}^*$ be a non-empty word such that for every non-empty prefix u of w,*

$$|u|_a > |u|_b \tag{†}$$

holds. Then, we show that there is a left-most derivation in G of the form

$$S \xrightarrow{*} w\alpha S$$

where α is a sequence of $|w|_a - |w_b|$ B's. (An analogous statement holds for the case $|u|_b > |u|_a$ for every non-empty prefix u of w.)

Proof. If $|w| = 1$, then clearly $w = a$. Now, the derivation

$$S \to aBS$$

does the job. So, we can assume that $|w| > 1$. Now, let u be a non-empty prefix of w. We will show that there is a derivation of the form

$$S \xrightarrow{*} u\alpha_u S$$

where α_u is a sequence of $|u|_a - |u|_b$ B's, and we will do so by induction on the length of the prefix $|u|$. For the base case, suppose $|u| = 1$, i.e u is the first letter of w. Clearly, it must be that $u = a$. In that case, consider the derivation

$$S \to aBS = u\alpha_u S$$

and clearly the base case is true. Now, let u be a prefix of w of length > 1. Suppose s is the last letter of u, so we can write $u = u's$, where u' is a non-empty prefix of w. By our induction hypothesis, there is a derivation

$$S \xrightarrow{*} u'\alpha_{u'} S$$

(and this is where we will use the condition †) By the condition †, we know that $\alpha_{u'}$ contains *atleast one* B, and hence there is a left-most B in $\alpha_{u'}$. If $s = a$, then we expand the left most B as aBB (by using the production $B \to aBB$), and hence we will get

$$S \xrightarrow{*} u'\alpha_{u'}S \xrightarrow{B\to aBB} u'aB\alpha_{u'}S = u\alpha_u S$$

If $s = b$, then we expand the left most B as b (by using the production $B \to b$), and hence we will get

$$S \xrightarrow{*} u'\alpha_{u'}S \xrightarrow{B\to b} u'b\alpha_u S = u\alpha_u S$$

and hence,in any case, the required derivation has been found. So, by induction, we see that there is a left-most derivation of the form

$$S \xrightarrow{*} w\alpha S$$

completing the proof of the claim. ■

Next, we will prove the hint in the footnotes.

Lemma 0.2. *Let w be any word in $\{a,b\}^*$. Then, there is a left-most derivation in G of the form*

$$S \xrightarrow{*} w\alpha S$$

where α is as defined in ***Lemma** 0.1*.

Proof. We will prove this by induction on the length of w. For the base case, $|w| = 1$. If $w = a$, then we have

$$S \to aBS$$

and if $w = b$, we have

$$S \to bAS$$

and hence the base $|w| = 1$ is true. Now, suppose the statement holds for all words of length atmost n, and let w be a word of length $n+1$. First, suppose $|w|_a = |w|_b$, and hence α will be an empty word in this case. Let s be the last letter of w, and let $w = w's$, where w' is non-empty. First, if $s = a$, then we have $w = w'a$. Moreover, it must be true that $|w'|_b = |w'|_a + 1$, and clearly w' is a word of length atmost n. So, by the inductive hypothesis, we know that there is a derivation

$$S \xrightarrow{*} w'AS$$

So, we can just do

$$S \xrightarrow{*} w'AS \xrightarrow{A\to a} w'aS = wS = w\alpha S$$

and clearly, the case when $s = b$ is analogous to this.

Next, suppose $|w|_a > |w|_b$ (the case $|w|_b > |w|_a$) has an analogous proof). There are three cases here, which we handle below.

(1) In the first case, suppose there is some non-empty prefix u of w with $|u|_b > |u|_a$. Since $|w|_a > |w|_b$, it follows that there is some non-empty prefix u' of w with $|u'|_a = |u'|_b$. In that case, we can write $w = u'v$, where v satisfies

$$|v|_a > |v|_b$$

and clearly, both u', v have length atmost n. So, by induction hypothesis, there are derivations $S \xrightarrow{*} u'S$ and $S \xrightarrow{*} v\alpha_v S$, where α_v is a sequence of $|v|_a - |v|_b$ B's. Combining these, we have a derivation

$$S \xrightarrow{*} u'S \xrightarrow{*} u'v\alpha_v S = w\alpha S$$

because clearly, $\alpha_v = \alpha$, and this case is handled.

(2) In the second case, there is some non-empty prefix u of w with $|u|_a = |u|_b$. This case can be handled the same way as case (1).
(3) In this case, *all* non-empty prefixes u of w satisfy $|u|_a > |u|_b$, and this case reduces to **Lemma** 0.1. So, all the cases have been handled.

So, by induction, the lemma has been proven. ■

2. Prove that every word with equal number of a's and b's is derivable from S to conclude that this is yet another grammar for the language of words with equal number of a's and b's.

Solution. By **Lemma** 0.2, if w is a word with $|w|_a = |w|_b$, then there is a left-most derivation of the form

$$S \xrightarrow{*} wS$$

Then, we can just do

$$S \xrightarrow{*} wS \to w$$

and hence every such word w is derivable in G. So, by problems **1.** and **2.** we conclude that G is a grammar accepting all words with equal number of a's and b's.

3. For the following, you need not prove the correctness of G.
(a) Modify G to a grammar over $\{a, b, c\}$ that accepts the language $\{w \mid vc$ prefix w then $|v|_a = |v|_b\}$.

Solution. It can be proven that any word w generated by $B \to b \mid aBB$ has the property

$$|w|_b = |w|_a + 1$$

and an analogous statement hold for every word generated by $A \to a \mid bAA$. So, we modify the grammar as follows.

$$\begin{aligned}
&S \to aBS \mid bAS \mid R \mid \epsilon \mid cS \\
&A \to a \mid bAA \\
&B \to b \mid aBB \\
&R \to aR \mid bR \mid \epsilon
\end{aligned}$$

where we have introduced the non-terminal R to handle the case where if there are no c's in the word, then it can be any word over $\{a, b\}^*$.

(b) Modify G to a grammar over $\{a, b, c\}$ that accepts the language $\{w \mid vc$ prefix w then $|v|_a \neq |v|_b\}$.

Solution. The modified grammar is the following.

$$
\begin{aligned}
S &\to S_1 \mid S_2S \mid R \\
S_1 &\to \ aCB_1 \mid bCA_1 \\
A_1 &\to Ca \mid CbCA_1CA_1 \mid \epsilon \\
B_1 &\to Cb \mid CaCB_1CB_1 \mid \epsilon \\
S_2 &\to \ aCB_2S_2 \mid bCA_2S_2 \mid \epsilon \\
A_2 &\to Ca \mid CbCA_2CA_2 \\
B_2 &\to Cb \mid CaCB_2CB_2 \\
C &\to c \mid cC \mid \epsilon \\
R &\to aR \mid bR \mid \epsilon
\end{aligned}
$$

Now I will explain the reasoning behind this grammar. The start non-terminal is S, and R is used for generating any random word over $\{a,b\}^*$. Any word in $L(G)$ must of of one of the following forms.

(1) The first form is this: let w be any word with equal a's and b's, and let w_{coll} be w with some c's *embedded* in w such that w_{coll} lies in $L(G)$, and also some letters of w collapsed to ϵ. For instance, let $w = aaabbb$, and let $w_{coll} = aaac$, where all the $b's$ are collapsed. Words like this where collapsing occurs are generated by using S_1, which is a copy of the original grammar with some modifications. Note that, A_1, B_1 are allowed to go to ϵ, because collapsing is allowed in this case.

(2) The second form is $w_{ncoll}w'$, where $w' \in L(G)$, and we explain w_{ncoll}. Let w be any word with equal number of a's and b's, and let w_{ncoll} be w with some $c's$ *embedded* with *no collapsing*. For example, if we use $w = aaabbb$, then w_{ncoll} can be $aaacbbb$. Words of the form w_{ncoll} are generated using S_2, where S_2 is another copy of S with some modifications. Note that A_2, B_2 are not allowed to go to ϵ here, because no collapsing is allowed in this case.

The non-terminal C is used to embed c's wherever it can be embedded (I couldn't figure out a simpler way of explaining this construction. Apologies for that.)

4. Describe a procedure that takes a context-free grammar G as input and checks whether $\epsilon \in L(G)$.

Solution. Let $N_0 = \epsilon$, and we inductively define sets N_i for $i \in \mathbb{N}$ as follows. Suppose N is the set of all non-terminals of G. Define

$$N_i = \{x \in N \mid x \to \alpha, \alpha \in \{N_0 \cup N_1 \cup ... \cup N_{i-1}\}^*\}$$

We show that

$$(*) \qquad X \xrightarrow{*} \epsilon \iff X \in N_i \text{ , for some } i \geq 1$$

Suppose $X \in N_1$, and clearly by the definition of N_1, it is clear that $X \xrightarrow{*} \epsilon$. Now suppose $X \in N_i$ implies $X \xrightarrow{*} \epsilon$ for every $1 \leq i \leq n$, and let $X \in N_{n+1}$. By the definition of N_{n+1}, we know that $X \to \alpha$, for some $\alpha \in \{N_0 \cup N_1 \cup ... \cup N_n\}^*$. If $\alpha = \epsilon$, then we are done, i.e $X \xrightarrow{*} \epsilon$. Otherwise, we know that each letter of α is a non-terminal belonging to N_n (because $N_1 \subseteq N_2 \subseteq ... \subseteq N_n$, and that $N_0 = \epsilon$). By our inductive hypothesis, we know that each non-terminal in N_n can generate ϵ, and hence it follows that α can generate ϵ. So again, $X \xrightarrow{*} \epsilon$, and so by induction, one direction of $(*)$ is proven. To prove the other direction, suppose $X \xrightarrow{*} \epsilon$ for some $X \in N$, and let T be the derivation tree, and clearly, every leaf of T is ϵ, so that every leaf is in N_0. Remove all leaves from T to get a new tree T'. By definition of N_1, we see that every leaf of T' is in N_1. We continue to remove leaves this way, and hence it must be true that the root of T, which is X, is in some X_i, for some $i \in \mathbb{N}$. This completes the proof of $(*)$.

Finally, as we mentioned above, it is clear that $N_1 \subseteq N_2 \subseteq$ ..., i.e the sets N_i form an increasing chain of sets under inclusion. Moreover, the chain must be eventually constant, because there are only finitely many non-terminals, i.e $N_{|N|}$ is the largest set in this chain. So, it follows that $\epsilon \in L(G)$ if and only if $S \in N_{|N|}$, where S is the starting non-terminal of the grammar G.

5. Describe a procedure that takes a context-free grammar G and a letter a and checks whether $a \in L(G)$.

Solution. To be completed.

6. Describe a procedure that takes a context-free grammar G and a letter a and checks if there is a word in $L(G)$ that contains the letter a.

Solution. To be completed.