# CS5371 Theory of Computation

Homework 2 (Solution)

1. Let $G$ be a CFG in Chomsky normal form that contains $b$ variables. Show that, if $G$ generates some string with a derivation having at least $2^b$ steps, $L(G)$ is infinite.

   **Answer:** Since $G$ is a CFG in Chomsky normal form, every derivation can generate at most two non-terminals, so that in any parse tree using $G$, an internal node can have at most two children. This implies that every parse tree with height $k$ has at most $2^k - 1$ internal nodes.

   If $G$ generates some string with a derivation having at least $2^b$ steps, the parse tree of that string will have at least $2^b$ internal nodes. Based on the above argument, this parse tree has height is at least $b + 1$, so that there exists a path from root to leaf containing $b + 1$ variables. By pigeonhole principle, there is one variable occurring at least twice. So, we can use the technique in the proof of the pumping lemma to construct infinitely many strings which are all in $L(G)$.

2. Give a formal description and the corresponding state diagram of a PDA that recognizes the language $A = \{w \mid 2\#_{\mathtt{a}}(w) \neq 3\#_{\mathtt{b}}(w), w \in \{\mathtt{a}, \mathtt{b}\}^*\}$, where $\#_c(w)$ denotes the number of character $c$ occurring in the string $w$.

   **Answer:** We want to construct a PDA so that if the input string $w$ has $2\#_{\mathtt{a}}(w) = 3\#_{\mathtt{b}}(w)$, we will reject it; otherwise, we will accept $w$. The key idea of the PDA is to use the stack to keep track of how many as or bs are 'in extra' (for $2\#_{\mathtt{a}}(w)$ to match with $3\#_{\mathtt{b}}(w)$). To do so, we assign each a to worth 2 units, and each b to worth -3 units, and the stack keeps track of the 'net total' as we process the string. It is easy to check that $w$ is in $A$ if and only if the net total after processing $w$ is not 0 units.

   Precisely, in figure-pda.pdf, we give the PDA for recognizing $A$.

3. Let $C = \{xy \mid x, y \in \{\mathtt{0}, \mathtt{1}\}^*, |x| = |y|, \text{ and } x \neq y\}$. Show that $C$ is a context-free language.

   **Answer:** We observe that a string is in $C$ if and only if it can be written as $xy$ with $|x| = |y|$ such that for some $i$, the $i$th character of $x$ is different from the $i$th character of $y$. To obtain such a string, we start generating the corresponding $i$th characters, and fill up the remaining characters.

   Based on the above idea, we define the CFG for $C$ is as follows:

$$\begin{aligned} S &\to AB \mid BA \\ A &\to XAX \mid \mathtt{0} \\ B &\to XBX \mid \mathtt{1} \\ X &\to \mathtt{0} \mid \mathtt{1} \end{aligned}$$

4. Let $A = \left\{wtw^R \mid w, t \in \{\mathtt{0}, \mathtt{1}\}^* \text{ and } |w| = |t|\right\}$. Prove that $A$ is not a context-free language.

   **Answer:** Suppose on the contrary that $A$ is context-free. Then, let $p$ be the pumping length for $A$, such that any string in $A$ of length at least $p$ will satisfy the pumping lemma.

   Now, we select a string $s$ in $A$ with $s = \mathtt{0}^{2p}\mathtt{0}^p\mathtt{1}^p\mathtt{0}^{2p}$. For $s$ to satisfy the pumping lemma, there is a way that $s$ can be written as $uvxyz$, with $|vxy| \leq p$ and $|vy| \geq 1$, and for any $i$, $uv^ixy^iz$ is a string in $A$.

   There are only three cases to write $s$ with the above conditions:

**Case 1:** $vy$ contains only 0s and these 0s are chosen from the last $0^{2p}$ of $s$. Let $i$ be a number with $7p > |vy| \times (i+1) \geq 6p$. Then, either the length of $uv^ixy^iz$ is not a multiple of 3, or this string is of the form $wtw'$ such that $|w| = |t| = |w'|$ with $w'$ is all 0s and $w$ is not all 0s (this is, $w' \neq w^R$).

**Case 2:** $vy$ does not contain any 0s in the last $0^2p$ of $s$. Then, either the length of $uv^2xy^2z$ is not a multiple of 3, or this string is of the form $wtw'$ such that $|w| = |t| = |w'|$ with $w$ is all 0s and $w'$ is not all 0s (that is, $w' \neq w^R$).

**Case 3:** $vy$ is not all 0s, and some 0s are from the last $0^2p$ of $s$. As $|vxy| \leq p$, $vxy$ in this case must be a substring in $1^p0^p$. Then, either the length of $uv^2xy^2z$ is not a multiple of 3, or this string is of the form $wtw'$ such that $|w| = |t| = |w'|$ with $w$ is all 0s and $w'$ is not all 0s (that is, $w' \neq w^R$).

In summary, we observe that there is no way $s$ can satisfy the pumping lemma. Thus, a contradiction occurs (where?), and we conclude that $A$ is not a context-free language.

5. **(Ogden's Lemma.)** There is a stronger version of the CFL pumping lemma known as *Ogden's lemma*. It differs from the pumping lemma by allowing us to focus on any $p$ "distinguished" positions of a string $z$ and guaranteeing that the strings to be pumped have between 1 and $p$ distinguished positions. The formal statement of Ogden's lemma is: Let $L$ be a context-free language. Then there is a constant $p$ such that for any string $z$ in $L$ with at least $p$ characters, we can mark any $p$ or more positions in $z$ to be distinguished, and then $z$ can be written as $z = uvwxy$, satisfying the following conditions:

   (i) $vwx$ has at most $p$ distinguished positions.

   (ii) $vx$ has at least one distinguished position.

   (iii) For all $i \geq$, $uv^iwx^iy$ is in $L$.

   Prove Ogden's lemma.

   **Answer:** Let $b$ be the maximum number of symbols in the right-hand side of a rule. For any parse tree $T$ of string $z$, and $z$ has at least $p$ marked positions. We say a leaf in $T$ is *marked* if its corresponding position in $z$ is marked. We say an internal node of $T$ is *marked* if the subtrees rooted at two or more of its children each contains a marked leaf.

   We claim that if every (root-to-leaf) path in $T$ contains at most $i$ marked internal nodes, $T$ has at most $b^i$ marked leaves. Assume that this claim is true (which will be proved shortly by induction). To prove Ogden's lemma, we set $p = b^{|V|+1}$. Then, the minimum marked internal nodes in a path is $|V| + 1$. By pigeonhole principle, there exists some variable appearing at least twice on that path. Then, by a similar way of proving the original pumping lemma, we can show that $z$ can be written as $uvwxy$ satisfying Ogden's lemma.

   We now go back to prove the claim. The claim is true for $i = 0$, since if every path in $T$ has at most 0 marked nodes, $T$ has no marked nodes. Thus, there must be at most $b^0 = 1$ marked leaves (why?). For $i \geq 1$, let $q$ be the unique marked internal node whose ancestors (if exist) are not marked. Then, the number of marked leaves in $T$ is equal to the total number of marked leaves under the children of $q$. Also, as $q$ is a marked node, in the subtree rooted at any child of $q$, every path has at most $i - 1$ marked nodes. By induction, every subtree has at most $b^{i-1}$ marked leaves. As $q$ has at most $b$ children, $T$ thus has at most $b^i$ marked leaves.

6. (Bonus Question) In this question, we apply Ogden's lemma and show that the language $L = \left\{\mathtt{a}^i\mathtt{b}^j\mathtt{c}^k \mid i = j \text{ or } j = k \text{ where } i, j, k \geq 0\right\}$ is inherently ambiguous.

   (a) Suppose that $G = (V, T, \Sigma, S)$ is a CFG for $L$, and let $p$ be the constant specified for $G$ in Ogden's lemma. Assume that $p > 3$.[†] Consider the string $z = \mathtt{a}^p\mathtt{b}^p\mathtt{c}^{p+p!}$ in $L$. Suppose we mark all the positions of '$\mathtt{a}$' as distinguished. Let $u, v, w, x, y$ be the five parts of $z$ as specified in the Ogden's lemma. Show that $v = a^t$ and $x = b^t$ for some $t$.

   (b) Following the above step, show that there exists a variable $A$ in $V$ such that

   $$S \overset{*}{\Rightarrow} uAy \overset{*}{\Rightarrow} uvAxy \overset{*}{\Rightarrow} \mathtt{a}^{p+p!}\mathtt{b}^{p+p!}\mathtt{c}^{p+p!}.$$

   (c) In a similar manner, find another derivation for $\mathtt{a}^{p+p!}\mathtt{b}^{p+p!}\mathtt{c}^{p+p!}$, and show that this derivation corresponds to a distinct parse tree from the derivation in Part (b). Conclude that $L$ is inherently ambiguous.

**Answer:**

(a) We know that $v$ and $x$ can hold only one type of $\{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$, or otherwise $uv^2wx^2y$ is not in the form of $\mathtt{a}^i\mathtt{b}^j\mathtt{c}^k$. Let us condier string $s = uv^2wx^2y$. Since $z$ contains $p$ $\mathtt{b}$s, so that $s$ contains less than $p + p!$ $\mathtt{b}$s. Thus, $s$ is of the form $a^ib^jc^k$ with $i = j$. As all $\mathtt{a}$s in $z$ are marked, there is at least $p + 1$ $\mathtt{a}$s in $s$ (by condition 2 of Ogden's lemma). Hence, $v = \mathtt{a}^t$ and $x = \mathtt{b}^t$ for some $t$, with $1 \leq t \leq p$.

(b) From the proof of Ogden's lemma, we know that there must be a non-terminal $A$ such that $S \overset{*}{\Rightarrow} uAy \overset{*}{\Rightarrow} uvAxy \overset{*}{\Rightarrow} uvwxy$, with $v = \mathtt{a}^t$ and $x = \mathtt{b}^t$. Let $n = p!/t + 1$. By pumping $s$ $n$ times, we get $s' = uv^nwx^ny = \mathtt{a}^{p+p!}\mathtt{b}^{p+p!}\mathtt{c}^{p+p!}$, and we know that it can be derived from $S$ as follows: $S \overset{*}{\Rightarrow} uAy \overset{*}{\Rightarrow} uvAxy \overset{*}{\Rightarrow} uvvAxxy \overset{*}{\Rightarrow} \cdots \overset{*}{\Rightarrow} uv^nwx^ny = \mathtt{a}^{p+p!}\mathtt{b}^{p+p!}\mathtt{c}^{p+p!}$.

(c) Let us consider another string $z' = \mathtt{a}^{p+p!}\mathtt{b}^p\mathtt{c}^p$. This time, we mark all the positions of '$\mathtt{c}$' as distinguished. Using Ogden's lemma as before, we know that there is a non-terminal $B$ such that $S \overset{*}{\Rightarrow} uBy \overset{*}{\Rightarrow} uvBxy \overset{*}{\Rightarrow} uvvBxxy \cdots \overset{*}{\Rightarrow} uv^{n'}wx^{n'}y = \mathtt{a}^{p+p!}\mathtt{b}^{p+p!}\mathtt{c}^{p+p!}$, with $v = \mathtt{b}^{t'}$, $x = \mathtt{c}^{t'}$, and $n' = p!/t' + 1$.

Note that the variables $B$ and $A$ must be different (Otherwise, we will be able to generate a string not of the form $\mathtt{a}^i\mathtt{b}^j\mathtt{c}^k$ with the grammar $G$). Thus, we have shown that in any grammar $G$ for $L$, there is some string having at least two distinct derivations. By definition, $L$ is inherently ambiguous.

[†] We can have this assumption, because Ogden's lemma holds for $p$ implies that it holds for any $q$ with $q > p$.