Sample solution to HW 2

(1) (i) \(a^3b^3\) is in \(L(G)\) with derivation tree:

(ii) \(a^2b^3\) is not in \(L(G)\). (See question (4).)

(iii) \(abba\) is not in \(L(G)\). (See question (4).)

(iv) \(a^2b^3a^3b^3\) is in \(L(G)\) with derivation tree:

(v) \(baababba\) is not in \(L(G)\). (See question (4).)

(2) (i) \(L_1 = \{a^mb^n \mid m > n\}\) can be generated by the CFG \(G\) with the set of variables \(V = \{S,T\}\), \(S\) is the start variable, and \(R\) contains the following rules:

\[
S \rightarrow aS \mid aT \\
T \rightarrow aTb \mid \epsilon
\]

(ii) \(L_2 = \{a^mb^n \mid n > m\}\) can be generated by the CFG \(G\) with the set of variables \(V = \{S,T\}\), \(S\) is the start variable, and \(R\) contains the following rules:

\[
S \rightarrow Sb \mid Tb \\
T \rightarrow aTb \mid \epsilon
\]

(iii) \(L_3 = \{a^{2n}b^n \mid n \geq 0\}\) can be generated by the CFG \(G\) with the set of variables \(V = \{S\}\), \(S\) is the start variable, and \(R\) contains the following rules:

\[
S \rightarrow aaSb \mid \epsilon
\]
(iv) \(L_4 = \{w^R \mid w \in \{a,b\}^*\}\) can be generated by the CFG \(G\) with the set of variables \(V = \{S, T\}\), \(S\) is the start variable, and \(R\) contains the following rules:

\[
\begin{align*}
S & \to T$ \\
T & \to aTa \mid bTb \mid $
\end{align*}
\]

(v) \(L_5\) is the complement of the language \(\{a^nb^n \mid n \geq 0\}\) over the alphabet \(\{a, b\}\). More formally, \(L_5 = \Sigma^* - \{a^nb^n \mid n \geq 0\}\), where \(\Sigma = \{a, b\}\).

A word \(w \in \Sigma^*\) is not in \(\{a^nb^n \mid n \geq 0\}\), if it satisfies one of the following conditions.

- In \(w\) some \(a\) appears after \(b\), and such a word can be generated by the following rules:

\[
\begin{align*}
A & \to ZbaZ \\
Z & \to aZ \mid bZ \mid \epsilon
\end{align*}
\]

Here the purpose of the variable \(Z\) is to generate arbitrary word.

- \(w\) is of the form: \(a^mb^n\), where \(m > n\), i.e., \(w \in L_1\) defined in (i) above.

Renaming the variables, we get the following rules to generate \(L_1\):

\[
\begin{align*}
B & \to aB \mid aC \\
C & \to aCb \mid \epsilon
\end{align*}
\]

- \(w\) is of the form: \(a^mb^n\), where \(m < n\), i.e., \(w \in L_2\) defined in (ii) above.

Renaming the variables, we get the following rules to generate \(L_2\):

\[
\begin{align*}
D & \to Db \mid Eb \\
E & \to aEb \mid \epsilon
\end{align*}
\]

We can combine the all the rules above to get the following grammar that generates the complement of \(\{a^nb^n \mid n \geq 0\}\):

- \(\Sigma = \{a, b\}\).
- \(V = \{S, A, B, C, D, E, Z\}\).
- \(S\) is the start variable.
- \(R\) consists of all the rules above, as well as the rule:

\[
\begin{align*}
S & \to A \mid B \mid D \\
A & \to ZbaZ \\
Z & \to aZ \mid bZ \mid \epsilon \\
B & \to aB \mid aC \\
C & \to aCb \mid \epsilon \\
D & \to Db \mid Eb \\
E & \to aEb \mid \epsilon
\end{align*}
\]

(3) Show that the following languages are not CFL.

(i) \(L_1 = \{a^kb^mc^n \mid k \leq m \leq n\}\) is not CFL.

The proof is via pumping lemma. Suppose to the contrary that \(L_1\) is CFL. Let \(\mathcal{G} = (\Sigma, V, R, S)\) be its CFG.

Consider the word \(w = a^kb^kc^k\), where \(k \geq M|R| + 1\) and \(M\) is the maximal length of the rule in \(R\). By pumping lemma, there is a partition \(w = sx_0yzt\) such that \(|x| + |z| \geq 1\) and for each \(i \geq 0\), \(v sx_0yzt w \in L(\mathcal{G})\). There are a few cases.
(a) If either $x$ or $z$ consists of more than two symbols, then by pumping lemma, either some $a$’s will appear after some $b$’s or $c$’s, or some $c$’s will appear after some $b$’s or $a$’s. This violates the criteria for a word to be in $L_1$.

(b) If both $x$ and $z$ do not contain $c$, then the number of $c$’s in $sx^2yz^2t$ will be less than either the number of $a$’s or $b$’s. This violates the criteria for a word to be in $L_1$.

(c) If one of $x$ or $z$ contains $c$, then the number of $c$’s in $sx^0yz^0t$ will be less than either the number of $a$’s or $b$’s.

Again, this will violate the criteria for a word to be in $L_1$.

Therefore, we conclude that $L_1$ is not CFL.

(ii) $L_2 = \{a^n \mid n \text{ is a prime number} \}$

Again, the proof is via pumping lemma. Suppose to the contrary that $L_2$ is CFL. Let $G = \langle \Sigma, V, R, S \rangle$ be its CFG.

Consider the word $a^n$, where $M[R] + 1 \leq m \leq n$ and $M$ is the maximal length of the rule in $R$. By pumping lemma, there is a partition $sxyzt$ such that $|x| + |z| \geq 1$ and for each $i \geq 0$, $sx^iyyz^it \in L(G)$. Now, $|sx^iyyz^it| = |s| + |y| + |t| + i(|x| + |z|)$.

If $|s| + |y| + |t| = 0$, the length $|sx^iyyz^it|$ is $|x^iz^i| = i(|x| + |z|)$, which is not a prime number. So, suppose $|s| + |y| + |t| \neq 0$, in which case, if we take $i = |s| + |y| + |t|$, the length of the word $v sx^iyyz^it$ is $(|x| + |z| + 1)(|s| + |y| + |t|)$, which again is not a prime number. Thus, it contradicts the fact that $v sx^iyyz^it \in L(G)$, for each $i \geq 0$, and therefore, $L_2$ is not CFL.

(4) Consider the grammar defined in (1). Prove that $w \in L(G)$ if and only if every prefix of $w$ has at least as many $a$’s as $b$’s.

**Proof:** We first prove the “only if” direction. Suppose $w \in L(G)$. The proof is by induction on the length of the derivation of $w$. The base case is when the length is 1, which implies that $w = \epsilon$, which is trivial. For the induction hypothesis, we assume that it holds for the all words $w \in L(G)$ with the length of derivation $\leq m - 1$.

For the induction step, suppose $w \in L(G)$ with the following derivation of length $m$:

$$S \Rightarrow u_1 \Rightarrow \cdots \Rightarrow u_m, \quad \text{where} \quad u_m = w.$$  

There are two cases:

- The first rule applied is $S \rightarrow aS$, i.e., $u_1 = aS$.

  Then, $w = aw'$, and $w'$ has derivation with length $m - 1$. By the induction hypothesis, every prefix of $w'$ has at least as many $a$’s as $b$’s, and hence, so does every prefix of $w$.

- The first rule applied is $S \rightarrow aSbS$, i.e., $u_1 = aSbS$.

  Then, $w = aw_1bw_2$, and $w_1, w_2$ have derivations with length $\leq m - 1$. By the induction hypothesis, every prefix of $w_1$ and $w_2$ has at least as many $a$’s as $b$’s, and hence, so does every prefix of $w$.

This completes the proof of the “only if” direction.

Now we will prove the “if” direction by induction on the length of $w$. The base case is when $w = \epsilon$, which is trivial.

For the induction hypothesis, we assume that it holds for every word of length $\leq m - 1$. The induction step is as follows. Let $w$ be a word of length $m$ such every prefix of $w$ has at least as many $a$’s as $b$’s.
Let \( w = d_1d_2\cdots d_m \), i.e., the symbol in position \( i \) is denoted by \( d_i \). We define a function \( f_w : \{1, \ldots, m\} \rightarrow \{1, \ldots, m\} \) as follows.

\[
f_w(i) = (\text{the number of } a's \text{ in } d_1 \cdots d_i) - (\text{the number of } b's \text{ in } d_1 \cdots d_i)
\]

The following claim is straightforward:

**Claim 1** Every prefix of \( w \) has at least as many \( a's \) as \( b's \) if and only if \( f_w(i) \geq 0 \) for every \( i = 1, \ldots, m \).

Coming back to the proof, there are two cases.

- There is \( 1 \leq k \leq m \) such that \( f_w(k) = 0 \).
  Let \( w_1 = d_1 \cdots d_k \) and \( w_2 = d_{k+1} \cdots d_m \). This implies that:
  - \( f_{w_1}(i) \geq 0 \), for every \( i = 1, \ldots, |w_1| \).
  - \( f_{w_2}(i) \geq 0 \), for every \( i = 1, \ldots, |w_2| \).
  By Claim 1, every prefix of the words \( w_1 \) and \( w_2 \) has at least as many \( a's \) as \( b's \). By the induction hypothesis, \( w_1, w_2 \in L(\mathcal{G}) \). Thus, we have derivations
    
    \[
    S \Rightarrow^* w_1 \\
    S \Rightarrow^* w_2
    \]

Now, we claim that \( w_1 \) starts with \( a \) and ends with \( b \). That it starts with \( a \) is obvious. If it ends with \( a \), we have \( f_w(k) = f_w(k-1) + 1 \). Since \( f_w(k) = 0 \), we will have \( f_w(k-1) = -1 \), which contradicts the assumption about \( w \). So, it has to end with \( b \). Now, let \( w_1 = aub \), hence, every prefix of \( u \) has at least as many \( a's \) as \( b's \). Therefore,

\[
S \Rightarrow^* u
\]

Combining \( S \Rightarrow^* w_1 \), \( S \Rightarrow^* w_2 \) and \( S \Rightarrow^* u \), we have:

\[
S \Rightarrow aSbS \Rightarrow^* aubS \Rightarrow^* aubw_2 = w_1w_2 = w.
\]

- There is no \( 1 \leq i \leq m \) such that \( f_w(i) = 0 \). Equivalently:
  \[
f_w(i) > 0 \quad \text{for every } 1 \leq i \leq m \quad (1)
\]
  Let \( w = aw_1 \) (since \( w \) must start with \( a \)). Now, \( f_{w_1}(i) = f_w(i) - 1 \), for every \( i = 1, \ldots, |w_1| \). By \( f_{w_1} \geq 0 \). By the induction hypothesis, \( S \Rightarrow^* w_1 \). Therefore, \( S \Rightarrow aS \Rightarrow^* aw_1 = w \).

Note: By the proof above, we can indeed simplify the grammar \( \mathcal{G} \) to be:

\[
S \rightarrow aS \mid aSb \mid SS \mid \epsilon
\]

(5) The statement is true.

Let \( A_1 = \langle \Sigma, \Gamma, Q_1, q_{01}, F_1, \delta_1 \rangle \) be a PDA and \( A_2 = \langle \Sigma, Q_2, q_{02}, F_2, \delta_2 \rangle \) be an NFA.

Construct the following PDA \( A = \langle \Sigma, \Gamma, Q, q_0, F, \delta \rangle \) that simulates both \( A_1 \) and \( A_2 \) simultaneously.

- \( Q = Q_1 \times Q_2 \).
\[ q_0 = (q_{01}, q_{02}) \cdot \]
\[ F = F_1 \times F_2 \cdot \]
\[ \delta \text{ is defined as follows.} \]
\[ \quad - \text{For every } (p_1, x, \text{pop}(y) \rightarrow (q_1, \text{push}(z)) \in \delta_1, \text{ where } x \neq \epsilon \text{ and } (p_2, x, q_2) \in \delta_2, \text{ the following transition is in } \delta:} \]
\[ ((p_1, p_2), x, \text{pop}(y) \rightarrow ((q_1, q_2), \text{push}(z)) \cdot \]
\[ \quad - \text{For every } (p_1, x, \text{pop}(y) \rightarrow (q_1, \text{push}(z)) \in \delta_1, \text{ where } x = \epsilon, \text{ for every } p_2 \in Q_2, \text{ the following transition is in } \delta:} \]
\[ ((p_1, p_2), x, \text{pop}(y) \rightarrow ((q_1, p_2), \text{push}(z)) \cdot \]

That \( \mathcal{A} \) accepts precisely \( L(\mathcal{A}_1) \cap L(\mathcal{A}_2) \) can be proved in a similar manner as the fact that regular languages are closed under intersection.