Exam 2 solutions

1. In this problem you will be asked to prove that three languages are context-free. If you choose to do this by giving either a CFG or a PDA for each language, you do not need to give a formal proof of correctness for the CFG or PDA. However, if it is too difficult for me to verify the correctness of your CFG or PDA, then you may lose points—so please aim to make your CFGs and/or PDAs as simple and clear as possible.

   (a) Let $\Sigma = \{0,1\}$ and $\Gamma = \{0,1,\#\}$. Prove that the language

   $$A = \left\{ u\#v \,:\, u,v \in \Sigma^* \text{ and } u \text{ is a substring of } v^{\mathrm{R}} \right\}$$

   over the alphabet Γ is context-free.

   Solution. Define a context-free grammar G as follows:

   $$\begin{aligned} S &\to XY \\ X &\to 0X0 \,\big|\, 1X1 \,\big|\, \#Y \\ Y &\to 0Y \,\big|\, 1Y \,\big|\, \varepsilon. \end{aligned}$$

   It holds that $\mathrm{L}(G) = A$, and therefore A is context-free.

   (b) Again let $\Sigma = \{0,1\}$ and $\Gamma = \{0,1,\#\}$. Prove that the language

   $$B = \left\{ u\#v \,:\, u,v \in \Sigma^* \text{ and } u \neq v \right\}$$

   over the alphabet Γ is context-free.

   Solution. Define a context-free grammar G as follows:

   $$\begin{aligned} S &\to W_0 1Y \,\big|\, W_1 0Y \,\big|\, Z \\ W_0 &\to XW_0X \,\big|\, 0Y\# \\ W_1 &\to XW_1X \,\big|\, 1Y\# \\ Z &\to XZX \,\big|\, XY\# \,\big|\, \#XY \\ X &\to 0 \,\big|\, 1 \\ Y &\to XY \,\big|\, \varepsilon. \end{aligned}$$

   The idea behind this CFG is as follows. First, the variable Z generates strings of the form $u\#v$ where u and v have different lengths. The variable W_0 generates strings that look like this:

   $$\underbrace{\square\square\cdots\square}_{n \text{ bits}} 0 \underbrace{\square\square\cdots\square}_{m \text{ bits}} \# \underbrace{\square\square\cdots\square}_{n \text{ bits}}$$

(where $\square$ denotes either 0 or 1), so that $W_0 1 Y$ generates strings that look like this:

$$\underbrace{\square\square\cdots\square}_{n \text{ bits}} 0 \underbrace{\square\square\cdots\square}_{m \text{ bits}} \# \underbrace{\square\square\cdots\square}_{n \text{ bits}} 1 \underbrace{\square\square\cdots\square}_{k \text{ bits}}$$

(for any choice of $n, m, k \in \mathbb{N}$). Similarly, $W_1 0 Y$ generates strings that look like this:

$$\underbrace{\square\square\cdots\square}_{n \text{ bits}} 1 \underbrace{\square\square\cdots\square}_{m \text{ bits}} \# \underbrace{\square\square\cdots\square}_{n \text{ bits}} 0 \underbrace{\square\square\cdots\square}_{k \text{ bits}}$$

Taken together, these two possibilities generate $u\#v$ for all binary strings u and v that differ in at least one position (and that may or may not have the same length). The three options together cover all possible $u\#v$ for which u and v are non-equal binary strings.

As $\mathrm{L}(G) = B$, it holds that B is context-free.

(c) Once again let $\Sigma = \{0, 1\}$ and $\Gamma = \{0, 1, \#\}$, and suppose that $C \subseteq \Sigma^*$ is a given regular language. Prove that the language

$$D = \left\{ u\#v \,:\, u \in C, v \in \Sigma^*, \text{and } |u| = |v| \right\}$$

over the alphabet Γ is context-free.

Solution. Let $M = (Q, \Sigma, \delta, q_0, F)$ be a DFA such that $\mathrm{L}(M) = C$. Define a CFG G as follows:

- G will have one variable X_q for each state $q \in Q$, with X_{q_0} being the start variable of G.
- For each choice of states $p, q \in Q$ and a symbol $\sigma \in \Sigma$ satisfying $\delta(p, \sigma) = q$, the rules

  $$X_p \to \sigma X_q 0 \,\big|\, \sigma X_q 1$$

  are included as rules of G.
- For each accepting state $q \in F$, the rule

  $$X_q \to \#$$

  is included as a rule of G.

It holds that $\mathrm{L}(G) = D$, and therefore D is context-free.

2. In this problem you will be asked to prove that two languages are context-free. The same guidelines regarding CFGs and PDAs as in the previous problem should be assumed for this problem as well.

   For both cases, let $\Sigma = \{0, 1\}$ and let $C \subseteq \Sigma^*$ be a given context-free language.

   (a) Prove that the language

   $$A = \left\{ uxv \,:\, u, x, v \in \Sigma^* \text{ and } uv \in C \right\}$$

   is context-free. In words, A is the language of all strings that can be obtained by taking any string in C and inserting an arbitrary substring $x \in \Sigma^*$ into that string.

Solution. Because the language C is context-free, there must exist a CFG G in Chomsky normal form that generates C. Let us define a CFG H as follows:

- Let W be a variable that is not used in G, and include these rules in H:

  $$W \to 0W \,\big|\, 1W \,\big|\, \varepsilon.$$

  The variable W generates an arbitrary binary string.
- For every rule of the form $X \to YZ$ in G, include these rules in H:

  $$X \to YZ$$
  $$X_0 \to Y_0Z \,\big|\, YZ_0$$

- For every rule of the form $X \to \sigma$ in G, include these rules in H:

  $$X \to \sigma$$
  $$X_0 \to W\sigma \,\big|\, \sigma W.$$

- If the rule $S \to \varepsilon$ appears in G, include this rule in H:

  $$S_0 \to W.$$

- The start variable of H is taken to be S_0.

It holds that $\mathrm{L}(H) = A$, and therefore A is context-free.

(b) Prove that the language

$$B = \big\{uv \,:\, u, x, v \in \Sigma^* \text{ and } uxv \in C\big\}$$

is context-free. In words, B is the language of all strings that can be obtained by taking any string in C and *deleting* an arbitrary substring of that string.

Solution. As in the previous solution, we will assume that G is a CFG in Chomsky normal form that generates C. Let us define a CFG H as follows:

- For every rule of the form $X \to YZ$ in G, include these rules in H:

  $$X \to YZ$$
  $$X_0 \to YZ_0 \,\big|\, Y_0$$
  $$X_1 \to Y_1Z \,\big|\, Z_1$$
  $$X_{01} \to YZ_{01} \,\big|\, Y_{01}Z \,\big|\, Y_0Z_1$$

- For every rule of the form $X \to \sigma$ in G, include these rules in H:

  $$X \to \sigma$$
  $$X_0 \to \sigma \,\big|\, \varepsilon$$
  $$X_1 \to \sigma \,\big|\, \varepsilon$$
  $$X_{01} \to \sigma \,\big|\, \varepsilon$$

- If the rule $S \to \varepsilon$ appears in G, include the rule $S_{01} \to \varepsilon$ in H.
- The start variable of H is taken to be S_{01}.

An explanation of the meaning of the variables in this CFG is as follows: each variable X generates the same strings in H that it does in G, each variable X_0 generates prefixes of strings generated by X, each variable X_1 generates suffixes of strings generated by X, and each variable X_{01} generates the strings obtained by removing any substring from a string generated by X.

It holds that $\mathrm{L}(H) = A$, and therefore A is context-free.

3. Prove that the following language is decidable by giving a high-level description of a DTM that decides it:

$$\big\{\langle G, k\rangle \,:\, G \text{ is a CFG, } k \in \mathbb{N}, \text{ and } |\mathrm{L}(G)| \geq k\big\}.$$

As usual, you should assume that $\langle G, k\rangle$ refers to the encoding of a CFG G together with a nonnegative integer k, with respect to some reasonable way of encoding these objects as strings over a fixed alphabet Σ.

Note that if $\mathrm{L}(G)$ is infinite, then the inequality $|\mathrm{L}(G)| \geq k$ is indeed satisfied for all $k \in \mathbb{N}$.

Finally, you may wish to make use of the fact that the following two languages were already proved to be decidable in lecture:

$$\mathrm{A}_{\mathrm{CFG}} = \big\{\langle G, w\rangle \,:\, G \text{ is a CFG and } w \in \mathrm{L}(G)\big\},$$
$$\mathrm{E}_{\mathrm{CFG}} = \big\{\langle G\rangle \,:\, G \text{ is a CFG with } \mathrm{L}(G) = \varnothing\big\}.$$

Solution. Consider a DTM M defined as follows:

On input $\langle G, k\rangle$, where G is CFG and $k \in \mathbb{N}$:

1. If $k = 0$ then *accept*.
2. If $\langle G\rangle \in \mathrm{E}_{\mathrm{CFG}}$ then *reject*.
3. Find the first string x generated by G:
4. Set $x \leftarrow \varepsilon$.
5. If $\langle G, x\rangle \notin \mathrm{A}_{\mathrm{CFG}}$ then increment x with respect to the lexicographic ordering of Γ^* (where Γ is the alphabet of G).
6. Construct a CFG H for the language $\mathrm{L}(G) \cap (\Gamma^* \setminus \{x\})$. (This is possible because $\Gamma^* \setminus \{x\}$ is a regular language.)
7. Set $G \leftarrow H$ and $k \leftarrow k - 1$, and goto 1.

The DTM M decides the language described in the problem, and therefore that language is decidable.

Note that this is just one way to decide the given language. An alternative approach is to first check if $\mathrm{L}(G)$ is infinite or finite (as was done on Assignment 3)—and if it is finite to compute $|\mathrm{L}(G)|$ in some way. A natural first step when using this approach is to convert G to Chomsky normal form, let n be the number of variables, and use the fact that if $\mathrm{L}(G)$ is finite, it cannot contain any strings of length 2^n or greater.