Aalto University
Department of Computer Science

CS-C2150 Theoretical Computer Science
Example Problems, Spring 2019

1. **Problem:**

   Describe the following languages **both** in terms of regular expressions **and** in terms of deterministic finite automata:

   (a) $\{w \in \{0,1\}^* \mid w \text{ contains 101 as a substring}\}$,

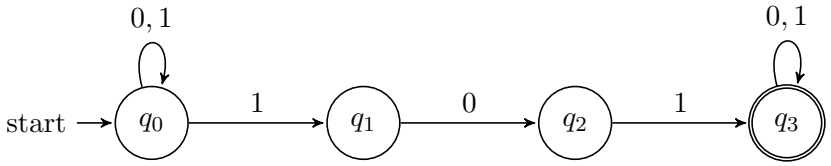
   (b) $\{w \in \{0,1\}^* \mid w \text{ does not contain 101 as a substring}\}$.

   Solution:

   (a) A regular expression for this language is easy to construct:

   $$(0|1)^*101(0|1)^*.$$

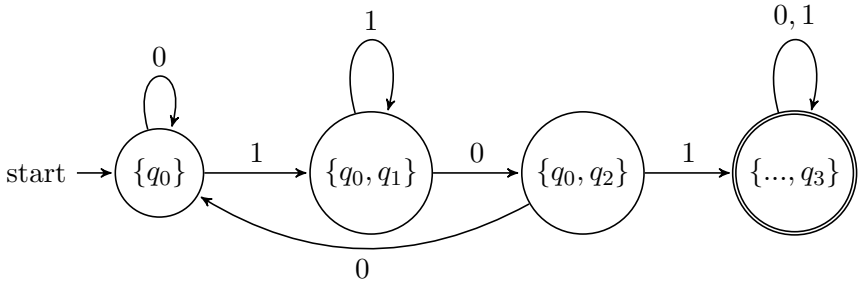
   A nondeterministic finite automaton is also easy to come up with:



   Let us then determinise this using the subset construction:

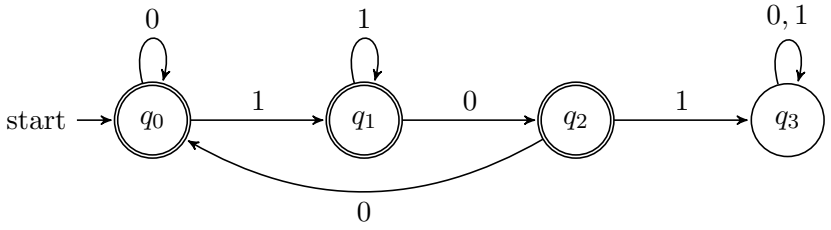
		0	1
$\rightarrow$	$\{q_0\}$	$\{q_0\}$	$\{q_0, q_1\}$
	$\{q_0, q_1\}$	$\{q_0, q_2\}$	$\{q_0, q_1\}$
	$\{q_0, q_2\}$	$\{q_0\}$	$\{q_0, q_1, q_3\}$
$\leftarrow$	$\{..., q_3\}$	$\{..., q_3\}$	$\{..., q_3\}$

   The last row of the table has been simplified based on the observation that from a set containing the accepting state q_3, one always moves again to some set containing q_3, and so all such states are equivalent. Thus, we obtain the following deterministic automaton:

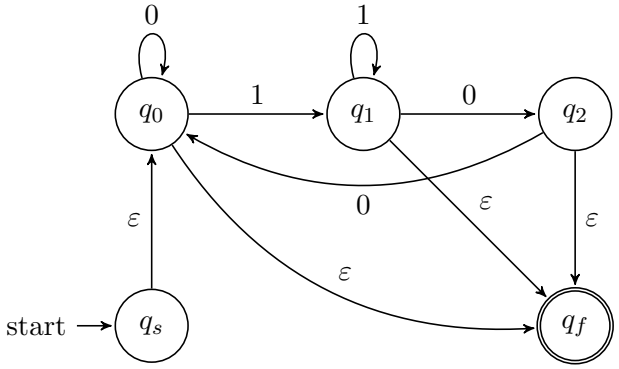


   (b) Observe that the language here is the complement of the language in part (a), and the DFA provided in part (a) is complete, i.e. all possible transitions are explicitly listed. Therefore, complementing the DFA from part (a) yields a DFA for this

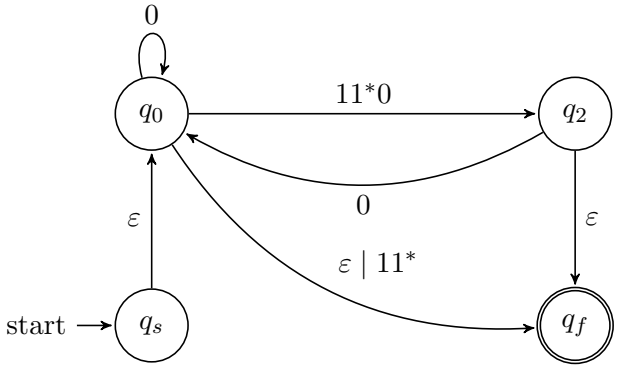
language:



All accepting states are inaccessible from state q_3, thus we may ignore it. To find a corresponding regular expression, we first add a new initial and final state and their connecting ε-transitions:

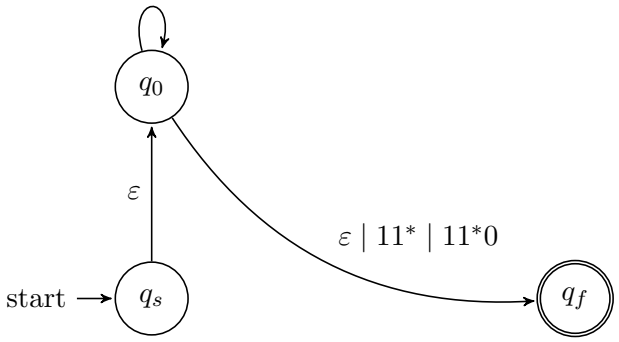


Remove state q_1 :

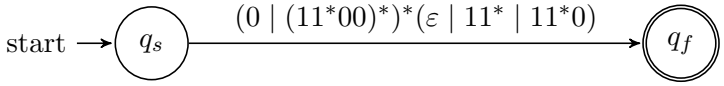


Remove state q_2:

$0 \mid (11^*00)^*$



Remove state q_0:



From the above, we may read off a regular expression describing the language:

$$(0 \mid (11^*00)^*)^*(\varepsilon \mid 11^* \mid 11^*0).$$

2. **Problem:**

Describe the following languages **both** in terms of regular expressions **and** in terms of deterministic finite automata:

(a) $\{w \in \{0,1\}^* \mid w \text{ contains an even number (possibly zero) of 1's}\}$

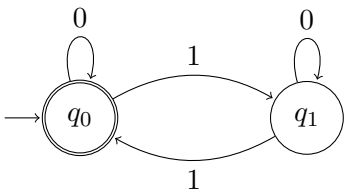
(b) $\{w \in \{0,1\}^* \mid w \text{ contains an odd number of 1's}\}$

(c) $\{wb \in \{0,1\}^* \mid \text{either } w \text{ contains an even number (possibly zero) of 1's and } b = 0, \text{ or } w \text{ contains an odd number of 1's and } b = 1\}$.

(*Hint:* In part (c) it may, depending on your solution method, be useful to first design a nondeterministic automaton.)

Solution:

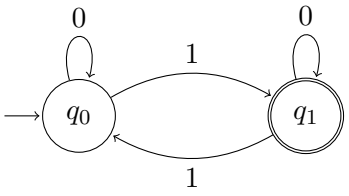
(a) Language is recognised by the deterministic finite automaton



and described by the following regular expression

$$(0^*10^*1)^*0^*$$

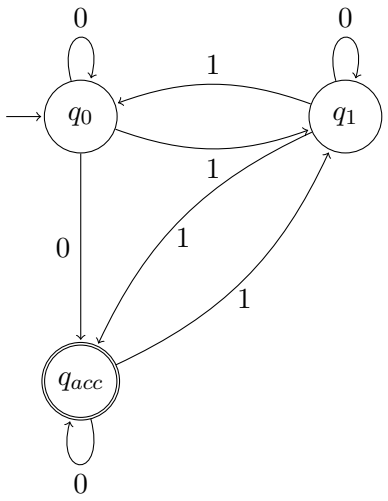
(b) This language is the complement of the one considered in part (a). We can use a lemma derived earlier in the course: the complement of a regular language can be arrived at by constructing a new DFA for which one simply changes the rejecting states of the original DFA to accepting states and vice versa.



We modify the regular expression of part (a) accordingly:

$$0^*1(0^*10^*1)^*0^*$$

(c) We use the DFA introduced in part (a) as a starting point, and add new complementary non-deterministic transitions from q_0 and q_1 to a new unique accepting state q_{acc}. However we also need to add transitions for the situations where the DFA have already transitioned to state q_{acc} but we still have more alphabet-symbols left in the string. As a result we get the following **NFA**:



We determinise the NFA by applying the subset construction for which the space of possible resulting states is the powerset of the number of original states in the NFA. In our case this yields a space of $2^3 = 8$ states for the resulting DFA.

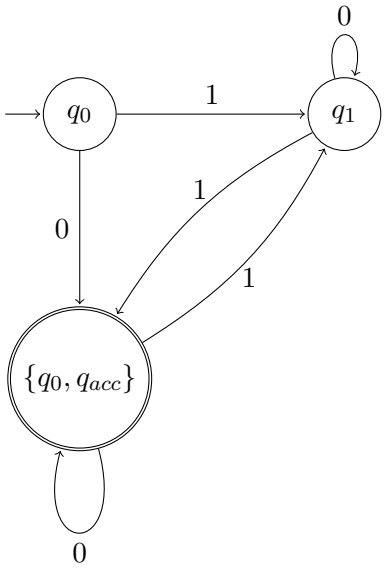
However when we derive new transitions for the DFA, we can simplify the procedure by only including those states that can be eventually reached from the starting state q_0. Transitions for the new transition function $\hat{\delta}$ are thus:

		0	1
$\rightarrow$	$\{q_0\}$	$\{q_0, q_{acc}\}$	$\{q_1\}$
	$\{q_1\}$	$\{q_1\}$	$\{q_0, q_{acc}\}$
$\leftarrow$	$\{q_0, q_{acc}\}$	$\{q_0, q_{acc}\}$	$\{q_1\}$

Resulting set of states in the determinised DFA is then

$$\hat{Q} = \{q_0, q_1, \{q_0, q_{acc}\}\}$$

Resulting **DFA**:



First we concatenate the parity bit at the end of the regular expressions from parts (a) and (b). Then by using the knowledge that regular expressions are closed under union, we derive the whole regular expression as a union of two parts:

$$(0^*10^*1)^*0^*0 \mid 0^*1(0^*10^*1)^*0^*1$$

3. **Problem:**

(a) Design a context-free grammar that generates the language

$$L = \{a^m b^n c^{m+n} \mid m, n \geq 0\}.$$

(b) Prove that the language L in part (a) is not regular.

Solution: Language L is generated by the following context-free grammar:

(a)

$$S \to aSc \mid A \mid \varepsilon$$
$$A \to bAc \mid \varepsilon$$

The grammar is split into two parts: variable S first generates matching pairs of a and c, and then variable A generates matching pairs of b and c. This arrangement guarantees that the total number of c's equals the sum of the numbers of a's and b's, and also that the symbols appear in the correct order.

(b) We will use the pumping lemma for regular languages to show that language L is not regular. For a contradiction, suppose that L is a regular language. Then by the pumping lemma there exists $p > 0$ such that every string $w \in L$ of length $|w| \geq p$ can be decomposed into three parts $w = xyz$ which satisfy the following conditions:

(i) $|xy| \leq p$,
(ii) $|y| \geq 1$,
(iii) for each $i \geq 0$, $xy^i z \in L$.

Now let $p > 0$ be as stated above. Consider the word $w = a^p b^p c^{2p} \in L$ and decompose it into parts xyz as indicated. By condition (i) and the structure of the chosen w, it must be the case that both x and y contain only a's. Let thus $y = a^q$ for some $q \geq 1$ (condition (ii)). By condition (iii), the word $w' = xy^2z = a^{p+q}b^p c^{2p}$ should then also be in L. But w' is not in L, as the number of a's and b's in it together do not equal the number of c's ($2p + q \neq 2p$) which would be required for $w' \in L$. We thus arrive at a contradiction, and conclude that L cannot be regular.

4. **Problem:**

Prove that all regular languages are context-free, without appealing to the correspondence between context-free grammars and pushdown automata. (Using this correspondence would make the proof trivial, since finite state automata are a special case of pushdown automata.) Illustrate your proof with an example.

Solution:

In order to prove that every regular language is context-free, we show by construction how to transform a DFA accepting a regular language into a context-free grammar that generates the same language.

Let B be a regular language. By definition, there exists a DFA $M = (Q, \Sigma, \delta, q_0, F)$ such that $L(M) = B$. Based on M, we design a context-free grammar $G_B = (V, \Sigma, R, S)$ generating B in the following way:

- For each state $q \in Q$ introduce corresponding variable $R_q \in V$.
- For every transition rule $\delta(q_i, a) = (q_j)$, where $q_i, q_j \in Q$ and $a \in \Sigma$, add production $R_{q_i} \to aR_{q_j}$.
- For every accepting state $q \in F$, add production $R_q \to \varepsilon$.
- Define the start variable S as R_{q_0}.

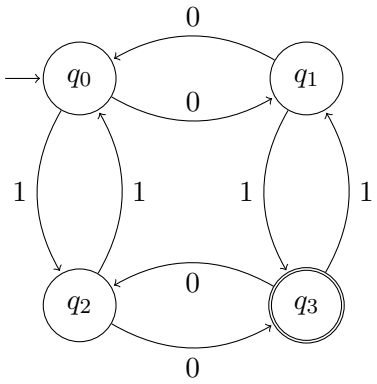
This "right-linear" context-free grammar G_B generates B, and thus B is a context-free language.

Let us demonstrate this method with the regular language

$$L = \{w \in \{0,1\}^* \mid \text{w contains odd number of 0's and odd number of 1's}\}.$$

We will show how to obtain a grammar for this language from its recognising DFA, and how to generate string $01001010 \in L$ using the grammar.

DFA:



We transform this DFA to a context-free grammar $G = (V, \Sigma, R, S)$:

- For set of states $Q = \{q_0, q_1, q_2, q_3\}$ we introduce set of variables $V = \{R_{q_0}, R_{q_1}, R_{q_2}, R_{q_3}\}$.

- After adding new rules and renaming R_{q_0} as the start variable S we have the following grammar:

$$\begin{aligned} S &\to 0R_{q_1} \mid 1R_{q_2} \\ R_{q_1} &\to 0S \quad \mid 1R_{q_3} \\ R_{q_2} &\to 0R_{q_3} \mid 1S \\ R_{q_3} &\to 0R_{q_2} \mid 1R_{q_1} \mid \varepsilon \end{aligned}$$

This grammar generates string 01001010 in the following way:

$$\underline{S} \Rightarrow 0\underline{R}_{q_1} \Rightarrow 01\underline{R}_{q_3} \Rightarrow 010\underline{R}_{q_2} \Rightarrow 0100\underline{R}_{q_3} \Rightarrow 01001\underline{R}_{q_1} \Rightarrow 010010\underline{R}_{q_0} \Rightarrow 0100101\underline{R}_{q_2}$$
$$\Rightarrow 01001010\underline{R}_{q_3} \Rightarrow 01001010\varepsilon = 01001010$$

5. **Problem:**

Consider the following context-free grammar G:

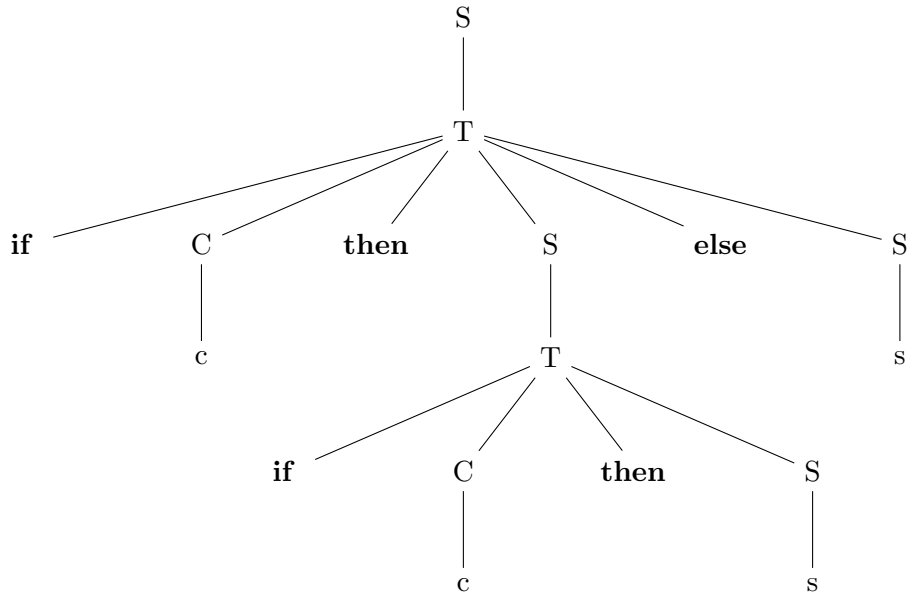
$$\begin{aligned} S &\rightarrow s \mid T \\ T &\rightarrow \textbf{if } C \textbf{ then } S \mid \textbf{ if } C \textbf{ then } S \textbf{ else } S \\ C &\rightarrow c \end{aligned}$$

(a) Give a parse tree for the string "**if** c **then if** c **then** s **else** s" in G.

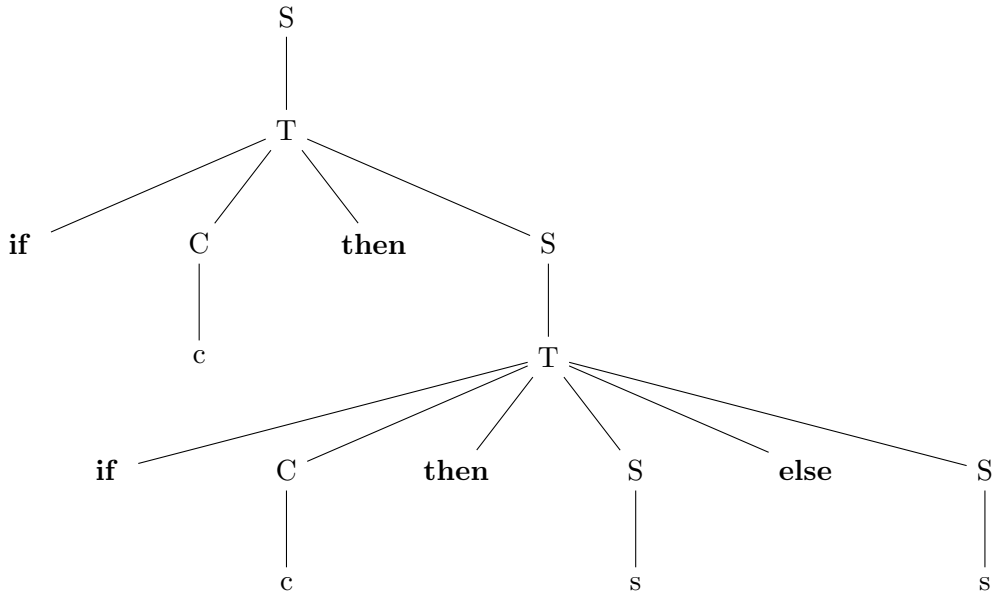
(b) Show that G is ambiguous.

Solution:

(a) A parse tree is drawn below:



(b) A grammar is ambiguous if there are two parse trees for some word in it. Therefore it suffices to present a second parse for the string in part (a):



6. **Problem:**

(a) Design a context-free grammar for the language

$$L = \{ucvcw \mid u, v, w \in \{0,1\}^*,\ v = u^R \text{ or } v = w^R \text{ (or both)}\}.$$

(Notation x^R denotes the reverse of string x, i.e. string x written backwards.)

(b) Show that the grammar you gave in part (a) is ambiguous.

(c) Prove (precisely!) that the language in part (a) is not regular. (*Hint:* Consider e.g. strings of the form $0^n c 0^n c 1^n$.)

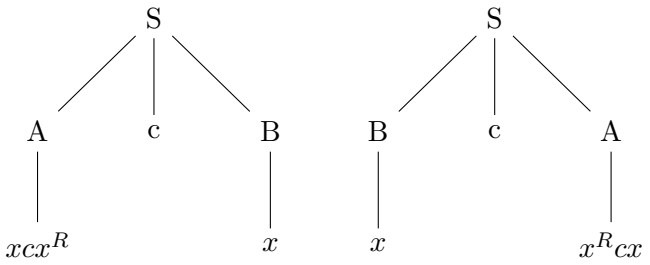
Solution:

(a) A context-free grammar is given below:

$$\begin{aligned} S &\to AcB \mid BcA \\ A &\to 0A0 \mid 1A1 \mid c \\ B &\to 0B \mid 1B \mid \varepsilon \end{aligned}$$

The variable A generates all the words of form xcx^R, and the variable B arbitrary words in $\{0,1\}^*$. The two productions for the start variable S correspond to the cases where $v = u^R$ ($S \to AcB$) and $v = w^R$ ($S \to BcA$).

(b) A context-free grammar G is ambiguous if there exists a word $w \in L(G)$ such that there are two different parse trees for w. In this particular case, every word of form xcx^Rcx, where $x \in \{0,1\}^*$, has two parse trees, as $x \in L(B)$ and $xcx^R, x^Rcx \in L(A)$:



Therefore, we conclude that the grammar is ambiguous.

(c) Let us show that the language L is not regular by using the pumping lemma. Recall the statement of this result:

If L is a regular language, there exists $p > 0$ such that every $w \in L$ of length $|w| \geq p$ can be decomposed into $w = xyz$ such that $|xy| \leq p, |y| \geq 1$ and for any $n \in \mathbb{N}$ (including $n = 0$), we also have $xy^nz \in L$.

For a contradiction, suppose the language L is regular. Let $p > 0$ be as in the pumping lemma for the language L, and consider the word $w = 0^p c 0^p c 1^p \in L$. Decompose $w = xyz$ as in the pumping lemma. Since xy and 0^p are both prefixes of w, and $|xy| \leq p = |0^p|$, it must be the case that xy is a prefix of 0^p. Since furthermore $|y| \geq 1$, we obtain that $xy^0z = xz = 0^q c 0^p c 1^p$ for some $q < p$. Since neither 0^q nor 1^p is the reverse of 0^p, it is clear that $xz \notin L$. However, according to the pumping lemma it should be the case that $xz \in L$, a contradiction. Therefore, we conclude that L is not a regular language.

7. **Problem:**

   (a) Prove that if the languages $L \subseteq \{0, 1, \#\}^*$ and $L' \subseteq \{0, 1\}^*$ are context-free, then so is the language $L'' = L[L'] \subseteq \{0, 1\}^*$, whose words are obtained from the words in L by replacing each #-symbol by some word in L' (not necessarily always the same).

   (b) The same problem as in part (a), but with respect to semi-decidable (Turing-recognisable) rather than context-free languages.

   Solution:

   (a) Let $L \subseteq \{0, 1, \#\}$ and $L' \subseteq \{0, 1\}$ be context-free, and let G and G' be context-free grammars generating the languages L and L', respectively. Without loss of generality, we may assume that all the variables in G and G' are distinct. Let the start symbol of G be S and the start symbol of G' be $S' \neq S$. Using G and G', we can then construct a grammar, call it $G[G']$, that generates the language $L[L']$ as follows:

   - The start symbol of $G[G']$ is the start symbol S of G.
   - For each production $A \to \omega$ of G, replace each occurence of the terminal # in ω by the start symbol S' of G'. Add the modified productions to $G[G']$.
   - Add every production $A \to \omega$ of G' to $G[G']$.

   The construction ensures that every word of L can be also derived from $G[G']$ with each occurence of # replaced by S', and vice versa. From each S', any word in G' can be derived, therefore $G[G']$ generates the language $L[L']$

   (b) Let T and T' be Turing machines recognising the languages L and L', respectively. We may construct a non-deterministic Turing machine $T[T']$, which replaces non-deterministically arbitrary substrings w' (possibly $w' = \varepsilon$) of the input w by the character # if $w' \in L'$ (i.e. w' is accepted by T') to get a new string $\widetilde{w}$, and then simulates T with input $\widetilde{w}$. By construction, $T[T']$ recognises $L[L']$, and since deterministic and non-deterministic Turing-machines recognise the same languages, we conclude that also $L[L']$ is Turing-recognisable i.e. semi-decidable.

8. **Problem:**

A language class C is *closed under complement*, if for every $L \in C$ also $\bar{L} \in C$.

(a) Show that the class of regular languages is closed under complement.

(b) Show that the class of context-free languages is not closed under complement. (*Hint:* The language $L = \{a^n b^n c^n \mid n \geq 0\}$ is not context-free.)

(c) Show that the class of decidable languages is closed under complement.

Solution:

(a) We want to prove that if a language L is regular, then $\bar{L} = \Sigma^* \setminus L$ is also regular. Assume then that L is regular. This means there is a finite automaton M that recognises L. Based on this machine we can construct a machine M' that recognises $\bar{L}$, thereby proving that $\bar{L}$ is regular. The construction is quite simple: M' is identical to M expect that its accept states are normal states, and its normal states accept states. Consider how M' behaves on some input x. If $x \in L$ then M accepts, meaning that the computation ends in an accepting state. But this state is not accepting in M' and therefore M' rejects as it should since $x \in L$ is equivalent to $x \notin \bar{L}$. If $x \notin L$ then M rejects, meaning that the computation ends in a state that is not accepting. But such a state is accepting in M' and therefore M' accepts, as it should since $x \notin L$ is equivalent to $x \in \bar{L}$.

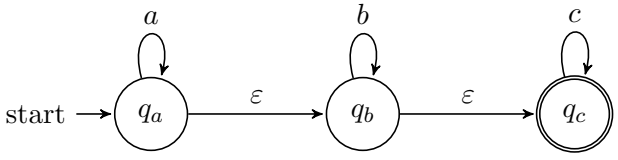
(b) We prove the statement by counterexample: we show that the complement $\bar{L}$ of the non-contextfree language $L = \{a^n b^n c^n \mid n \geq 0\}$ is in fact context-free. Hence there is a context-free language ($\bar{L}$) whose complement ($\bar{\bar{L}} = L$) is not context-free. Consider the language $\bar{L}$. It contains all strings that don't have the same amount of a's b's and c's appearing in order. We proceed to separate the language into parts that we can show are context-free, eventually showing the entire language is context-free.

Strings in $\bar{L}$ can be of two types: they can either have differing numbers of a's, b's and c's, or the letters can be in the wrong order. In other words $\bar{L} = \bar{L}_{\text{ord}} \cup L_{\text{num}}$ where $L_{\text{ord}} = \{a^i b^j c^k \mid i, j, k \in \mathbb{N}\}$ and $L_{\text{num}} = \{a^i b^j c^k \mid i \neq j \text{ or } j \neq k\}$. We can further separate $L_{\text{num}} = L_{a \neq b} \cup L_{b \neq c}$, where $L_{x \neq y}$ means the language where the numbers of letters x and y differ. We have then that

$$\bar{L} = \bar{L}_{\text{ord}} \cup L_{a \neq b} \cup L_{b \neq c}.$$

We will show that these component languages are context-free one by one, and then show that the class of context-free languages is closed under union, thus concluding that $\bar{L}$ is context-free.

$\bar{L}_{\text{ord}}$ We show that L_{ord} is a regular language, so its complement $\bar{L}_{\text{ord}}$ is also regular by part (a). Since the class of regular languages is a subset of the class of context-free languages, we can then conclude that $\bar{L}_{\text{ord}}$ is context-free. We show that L_{ord} is regular by describing a nondeterministic finite automaton that recognises it. The state diagram for such a machine is shown below.



$L_{a \neq b}$ We can describe this language with a context-free grammar. We can think of the language as composed of two cases: either there are more a's than b's or less a's than b's. The first rule of the grammar below expresses this idea.

$$\begin{aligned} S &\to MC \mid LC \\ M &\to aM \mid aE \\ L &\to Lb \mid Eb \\ E &\to aEb \mid \varepsilon \\ C &\to cC \mid \varepsilon \end{aligned}$$

If we use the production $S \to MC$ then the M first adds one or more a's to the beginning of the string and then an equal number of a's and b's, leading to an excess of a's. Then the C adds some number of c's at the end. In the production $S \to LC$ some b's are added at first, leading there to be fewer a's than b's.

$L_{b \neq c}$ The idea here is very similar to the previous part, the grammar is below

$$\begin{aligned} S &\to AM \mid AL \\ M &\to bM \mid bE \\ L &\to Lc \mid Ec \\ E &\to aEb \mid \varepsilon \\ A &\to aA \mid \varepsilon \end{aligned}$$

Finally to show that the class of context-free languages is closed under union, let L_1 and L_2 be two context-free languages, described by context-free grammars G_1 and G_2. A context-free grammar for $L_1 \cup L_2$ can then be constructed by combining all productions from G_1 and G_2 (renaming variables if necessary), renaming their start variables as S_1 and S_2 and adding a new start production $S \to S_1 \mid S_2$.

(c) This argument is very similar to that for regular languages. Assume that L is decidable, meaning that there is some Turing machine M that recognises it and halts on all inputs. We can construct a Turing machine M' that is identical to M expect that the accepting and rejecting states are swapped. Machine M' then also halts on every input, and gives the opposite answer from M. Clearly M' recognises the language $\bar{L}$.

9. **Problem:**

   Show that if a language $L \subseteq \Sigma^*$ is semi-decidable but not decidable, then its complement language $\bar{L} = \Sigma^* - L$ is not semi-decidable. (You may assume as known any auxiliary results related to this claim that have been presented during the course.)

   Solution: Let L be a language that is semi-decidable but not decidable, and suppose that its complement $\bar{L}$ is also semi-decidable. We will show that then L would in fact be decidable, a contradiction.

   Since L is semi-decidable, there is by definition a Turing machine M_1 that recognises it, i.e. on any input $x \in L$, M_1 halts and accepts. Similarily, there is a machine M_2 that recognises $\bar{L}$. We construct from these two machines a Turing machine M that recognises L correctly and halts on all inputs, establishing that L is decidable, contrary to our assumption.

   We will describe M on a general level, but an exact definition could be written based on this description. The machine M simulates the machines M_1 and M_2 in parallel on two independent tapes. Given an input x, M runs both M_1 and M_2 with input x and accepts if M_1 halts and accepts, and rejects if M_2 halts and accepts. Since an input x either is in the language L or it is not, machine M always halts and gives the correct answer on whether an input x is in language L or not. Hence L is decidable.

10. **Problem:**

Give a brief but precise justification, based on results presented on the course, for each of the following statements: (i) all regular languages are context-free, (ii) all context-free languages are decidable.

Solution:

(i) Any regular language L can, by definition, be recognised by some finite automaton. Each finite automaton can be transformed into a (left- or right-recursive) context-free grammar recognising the same language. This means that the language L is also context-free.

(ii) Any context-free language L is, by definition, generated by some context-free grammar G. Grammar G can be effectively transformed into an equivalent Chomsky normal form grammar G', which can then be used in the CYK parsing algorithm to decide whether a given input word is in the language L. The CYK-algorithm can be implemented for example in C, which by the Church-Turing thesis implies that the language L is decidable.