

CS334: Theory of Computation Notes

Steven DeFalco

Fall 2023

Contents

1	Introduction	3
1.1	Automata, Computability, and Complexity	3
1.2	Mathematical Notions and Terminology	3
1.2.1	Sets	3
1.2.2	Sequences and Tuples	4
1.2.3	Functions and Relations	4
1.2.4	Graphs	5
1.2.5	Strings and Languages	6
1.2.6	Boolean Logic	6
1.3	Definitions, Theorems, and Proofs	7
1.3.1	Finding Proofs	7
1.4	Types of Proofs	8
1.4.1	Proof by Construction	8
1.4.2	Proof by Contradiction	8
1.4.3	Proof by Induction	8
2	Regular Languages	8
2.1	Finite Automata	8
2.1.1	Formal Definition of Finite Automaton	9
2.1.2	Formal Definition of Computation	10
2.1.3	Designing Finite Automata	10
2.1.4	The Regular Operations	11
2.2	Nondeterminism	11
2.3	Equivalence of NFAs and DFAs	13
2.4	Closure under the Regular Operations	13
2.5	Regular Expressions	14
2.6	Equivalence with Finite Automata	15
2.7	Nonregular Languages	17

3	Context-Free Languages	17
3.1	Context-Free Grammars	18
3.1.1	Designing Context-Free Grammars	19
3.1.2	Ambiguity	19
3.1.3	Chomsky Normal Form	20

1 Introduction

1.1 Automata, Computability, and Complexity

The central question of **complexity theory** is *what makes some problems computationally hard and others easy?*; the answer is unknown. Cryptography is unique in that it specifically requires computational problems that are hard, rather than easy. In **complexity theory**, the objective is to classify problems as easy ones and hard ones; whereas in **computability theory**, the classification of problems is by those that are solvable and those that are not.

Automata theory deals with the definitions and properties of mathematical models of computations.

1.2 Mathematical Notions and Terminology

1.2.1 Sets

A **set** is a group of objects represented as a unit. Sets may contain any type of object, including numbers, symbols, and even other sets. The objects in a set are called its *elements* or *members*. One way to describe a set is to list the set's elements inside braces. Thus the set

$$S = \{7, 21, 57\}$$

contains the elements 7, 21, and 59. The symbols \in and \notin denote set membership and nonmembership. We say that A is a **subset** of B , written $A \subseteq B$, if every member of A is also a member of B . We say that A is a **proper subset** of B , written $A \subset B$, if A is a subset of B and not equal to B .

The order of describing a set doesn't matter, nor does repetition of its members. If we do want to take the number of occurrences of members into account, we call the group a **multiset** instead of a set. An **infinite set** contains infinitely many elements.

We write the set of **natural numbers** N as

$$\{1, 2, 3, \dots\}$$

. The set of **integers** Z is written as

$$\{\dots, -2, -1, 0, 1, 2, \dots\}$$

The set with zero members is called the **empty set** and is written \emptyset . A set with one member is sometimes called a **singleton set** and a set with two members is called an **unordered pair**.

When we want to describe a set containing elements according to some rule, we write $\{n | \text{rule about } n\}$.

If we have two sets A and B , the **union** of A and B , written $A \cup B$, is the set we get by combining all the elements in A and B into a single set. The **intersection** of A and B , written $A \cap B$, is the set of elements that are both A and B . The **complement** of A , written \overline{A} , is the set of elements under consideration that are *not* in A .

1.2.2 Sequences and Tuples

A **sequence** of objects is a list of these objects in some order. We usually designate a sequence by writing the list within parentheses. For example, the sequence 7,21,57 would be written

$$(7, 21, 57)$$

. The order does matter in a set.

Finite sequences often are called **tuples**. A sequence with k elements is a **k -tuple**. A 2-tuple is also called an **ordered pair**.

Sets and sequences may appear as elements of other sets and sequences. For example, the **power set** of A is the set of all subsets of A . If A is the set 0,1, the power set of A is the set $\{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$.

If A and B are two sets, the **Cartesian product** or defcross product of A and B , written $A \times B$, is the set of all ordered pairs wherein the first element is a member of A and the second element is a member of B .

1.2.3 Functions and Relations

A **function** is an object that sets up an input-output relationship. A function takes an input and produces an output. In every function, the same input always produces the same output.

A function is also called a **mapping**, and, if $f(a) = b$, we say that f maps a to b .

The set of possible inputs to the function is called its **domain**. The outputs of a function come from a set called the **range**. The notation for saying that f is a function with domain D and range R is

$$f : D \rightarrow R$$

When the domain of a function f is $A_1 \times \cdots \times A_k$ for some sets A_1, \dots, A_k , the input to f is a k -tuple and we call the a_i the **arguments** to f . A function with k arguments is called a **k -ary function**, and k is called the **arity** of the

function. If k is 1, f has a single argument and f called a **unary function**. If k is 2, f is a **binary function**. Certain familiar binary functions are written in a special **infix notation**, with the symbol for the function placed between its two arguments, rather than in **prefix notation**, with the symbol preceding.

A **predicate** or **property** is a function whose range is $\{\text{TRUE}, \text{FALSE}\}$. For example, let *even* be a property that is TRUE if its input is an even number and FALSE if its input is an odd number.

A property whose domain is a set of k -tuples $A \times \cdots \times A$ is called a **relation**, a **k -ary relation**, or a **k -ary relation on A** .

A special type of binary relation, called an **equivalence relation**, captures the notion of two objects being equal in some feature. A binary relation R is an equivalence relation if R satisfies three conditions:

1. R is **reflexive** if for every x , xRx ;
2. R is **symmetric** if for every x and y , xRy implies yRx ; and
3. R is **transitive** if for every x, y , and z , xRy and yRz implies xRz .

1.2.4 Graphs

An **undirected graph**, or simply a **graph**, is a set of points with lines connecting some of the points. The points are called **nodes** or **vertices**, and the lines are called **edges**.

The number of edges at a particular node is the **degree** of that node. No more than one edge is allowed between any two nodes. We may allow an edge from a node to itself, called a **self-loop**.

We say that graph G is a **subgraph** of graph H if the nodes of G are a subset of the nodes of H , and the edges of G are the edges of H on the corresponding nodes.

A **path** in a graph is a sequence of nodes connected by edges. A **simple path** is a path that doesn't repeat any nodes. A graph is **connected** if every two nodes have a path between them. A path is a **cycle** if it starts and ends in the same node. A **simple cycle** is one that contains at least three nodes and repeats only the first and last nodes. A graph is a **tree** if it is connected and has no simple cycles. A tree may contain a specially designated node called the **root**. The nodes of degree 1 in a tree, other than the root, are called the **leaves** of the tree.

A **directed graph** has arrows instead of lines. The number of arrows pointing from a particular node is the **outdegree** of that node, and the number of arrows

pointing to a particular node is the *indegree*.

A path in which all the arrows point in the same direction as its steps is called a *directed graph*. A directed graph is *strongly connected* if a directed path connects every two nodes.

1.2.5 Strings and Languages

Strings of characters are fundamental building blocks in computer science. The alphabet over which the strings are defined may vary with the application. For our purposes, we define an *alphabet* to be any nonempty finite set. The members of the alphabet are *symbols* of the alphabet. We generally use capital Greek letter Σ and Γ to designate alphabets.

A **string over an alphabet** is a finite sequence of symbols from that alphabet, usually written next to one another and not separated by commas. If w is a string over Σ , the *length* of w , written $|w|$ is the number of symbols that it contains. The string of length zero is called the *empty string* and is written ϵ . The *reverse* of w , written w^R , is the string obtained by writing w in the opposite order. String z is a *substring* of w if z appears consecutively within w .

If we have a string x of length m and string y of length n , the *concatenation* of x and y , written xy , is the string obtained by appending y to the end of x , as in $x_1 \cdots x_m y_1 \cdots y_n$.

The *lexicographic order* of strings is the same as the familiar dictionary order. We'll occasionally use a modified lexicographic order, called *shortlex order* or simply *string order*, that is identical to lexicographic order, except that shorter strings precede longer strings. Thus the string ordering of all strings over the alphabet $\{0, 1\}$ is

$$(\epsilon, 0, 1, 00, 01, 10, 11, 000, \dots)$$

Say that string x is a *prefix* of string y if a string z exists where $xz = y$, and that x is a *proper prefix* of y if in addition $x \neq y$. A *language* is a set of strings. A language is a *prefix-free* if no member is a proper prefix of another member.

1.2.6 Boolean Logic

Boolean logic is a mathematical system built around the two values TRUE and FALSE. The values TRUE and FALSE are called the *Boolean values* and are often represented by the values 1 and 0.

We can manipulate Boolean values with the *boolean operations*. The simplest boolean operation is the *negation* or *NOT* operation, designated with the symbol \neg . The negation of a Boolean value is the opposite value. We designate the *conjunction* or *AND* operation with the symbol \wedge . The conjunction

of two Boolean values is 1 if both of those values are 1. The **disjunction** or **OR** operation is designated with the symbol \vee . The disjunction of two Boolean values is 1 if either of those values is 1.

The **exclusive or** or **XOR** operation is designated by the \oplus symbol and is 1 if either but not both of its two operands is 1. The **equality** operation, written \leftrightarrow , is 1 if both if its operandws have the same value. Finally, the **implication** operation is designated by the symbol \rightarrow and is 0 if its first operand is 1 and its second operand is 0; otherwise, \rightarrow is 1.

The **distributive law** for AND and OR comes in handy when we manipulate Boolean expression; it comes in two forms:

- $P \wedge (Q \vee R)$ equals $(P \wedge Q) \vee (P \wedge R)$, and its dual
- $P \vee (Q \wedge R)$ equals $(P \vee Q) \wedge (P \vee R)$

1.3 Definitions, Theorems, and Proofs

Definitions describe the objects and notions that we use. When defining some object, we must make clear what constitutes that object and what does not.

After we have defined various objects and notions, we usually make **mathematical statements** about them. Typically, a statement expresses that some object has a certain property.

A **proof** is a convincing logical argument that a statement is true.

A **theorem** is a mathematical statement proved true. Occasionally, we prove statements that are interesting only because they assist in the proof of another, more significant statement. Such statements are called **lemmas**. Occasionally a theorem or its proof may allow us to conclude easily that other, related statements are true. These statements are called **corollaries** of the theorem.

1.3.1 Finding Proofs

The only way to determine the truth or falsity of a mathematical statement is with a mathematical proof. Experimenting with examples is especially helpful. Thus if the statement says that all objects of a certain type have a particular property, pick a few objects of that type and observe that they actually do have that property. After doing so, try to find an object that fails to have the property, called a **counterexample**. If the statement actually is true, you will not be able to find a counterexample.

1.4 Types of Proofs

Several types of arguments arise frequently in mathematical proofs. Here, we describe a few that often occur in the theory of computation.

1.4.1 Proof by Construction

Many theorems state that a particular type of object exists. One way to prove such a theorem is by demonstrating how to construct the object. This technique is a *proof by construction*.

1.4.2 Proof by Contradiction

In one common form of argument for proving a theorem, we assume that the theorem is false and then show that this assumption leads to an obviously false consequence, called a contradiction.

1.4.3 Proof by Induction

Proof by induction is an advanced method used to show that all elements of an infinite set have a specified property. For example we may use a proof by induction to show that an arithmetic expression computes a desired quantity for every assignment to its variables, or that a program works correctly at all steps or for all inputs.

Every proof by induction consists of two parts, the *basis* and the *induction step*. Each part is an individual proof on its own. In the induction step, the assumption that $P(i)$ is true is called the *induction hypothesis*.

Basis: Prove that $P(i)$ is true. **Induction step:** For each $i \geq 1$, assume that $P(i)$ is true and use this assumption to show that $P(i + 1)$ is true.

2 Regular Languages

Real computers are quite complicated: too much so to allow us to set up a manageable mathematical theory of them directly. Instead, we use an idealized computer called a *computational model*. As with any model in science, a computational model may be accurate in some ways but perhaps not in others. Thus we will use several different computation models, depending on the features we want to focus on. We begin with the simplest model, called the *finite state machine* or *finite automaton*.

2.1 Finite Automata

Finite automata are good models for computers with an extremely limited amount of memory. Finite automata and their probabilistic counterpart *Markov*

chains are useful tools when we are attempting to recognize patterns in data.

Finite automata can be represented by a *state diagram*. The *start state* is indicated by the arrow point at it from nowhere. The *accept state* is the one with a double circle. The arrows going from one state to another are called *transitions*. When this such automation receives an input, it processes that and produces an output; the output is either *accept* or *reject*.

2.1.1 Formal Definition of Finite Automaton

A finite automaton has several parts. It has a set of states and rules for going from one state to another, depending on the input symbol. It has an input alphabet that indicates the allowed input symbols. It has a start state and a set of accept states. The formal definition says that a finite automaton is a list of those five objects: set of states, input alphabet, rules for moving, start state, and accept states. In mathematical language, a list of five elements is often called a 5-tuple. Hence we define a finite automaton to be a 5-tuple consisting of these five parts.

We use something called a *transition function*, frequently denoted δ , to define the rules for moving. If the finite automaton has an arrow from a state x to a state y labeled with the input symbol 1, that means that if the automaton is in state x when it reads a 1, it then moves to state y . We can indicate the same thing with the transition function by saying that $\delta(x, 1) = y$. This notation is a kind of mathematical shorthand.

Definition 2.1 (Finite automaton) A 5-tuple $(Q, \Sigma, \delta, q_0, F)$, where

1. Q is a finite set called the *states*,
2. Σ is a finite set called the *alphabet*,
3. $\delta: Q \times \Sigma \rightarrow Q$ is the *transition function*,
4. $q_0 \in Q$ is the *start state*, and
5. $F \subseteq Q$ is the *set of accept states*

If A is the set of all strings that machine M accepts, we say that A is the *language of machine M* and write $L(M) = A$. We say that M *recognizes A* or that M *accepts A* . Because the term *accept* has different meanings when we refer to machines accepting strings and machines accepting languages, we prefer the term *recognize* for languages in order to avoid confusion.

A machine may accept several strings, but it always recognizes only one language. If the machine accepts no strings, it still recognizes one language—namely, the empty language \emptyset .

2.1.2 Formal Definition of Computation

Let $M = (Q, \Sigma, \delta, q_0, F)$ be a finite automaton and let $w = w_1w_2 \cdots w_n$ be a string where each w_i is a member of the alphabet Σ . Then M **accepts** w if a sequence of states r_0, r_1, \dots, r_n in Q exists with three conditions:

1. $r_0 = q_0$,
2. $\delta(r_i, w_{i+1}) = r_{i+1}$, for $i = 0, \dots, n-1$, and
3. $r_n \in F$

Condition 1 says that the machine starts in the start state. Condition 2 says that the machine goes from state to state according to the transition function. Condition 3 says that the machine accepts its input if it ends up in an accept state. We say that M **recognizes language** A if $A = \{w \mid M \text{ accepts } w\}$.

Definition 2.2 In state q , if the next input symbol is a , the next state is $\delta(q, a)$.
Define: $\Delta : Q \times \Sigma^* \rightarrow Q$ recursively as follows:

$$\begin{aligned}\Delta(q, \epsilon) &= q \\ \Delta(q, ax) &= \Delta(\delta(q, a), x), a \in \Sigma, x \in \Sigma^*\end{aligned}$$

$M = (Q, \Sigma, \delta, q_0, F)$ **accepts** a string $w = w_1w_2 \cdots w_n$ iff $\Delta(q_0, w) \in F$

Definition 2.3 (Regular Language) A language that is recognized by some finite automaton

2.1.3 Designing Finite Automata

Try putting yourself in the place of the machine you are trying to design and then see how you would go about performing the machine's task. Suppose that you are given some language and want to design a finite automaton that recognizes it. Pretending to be the automaton, you receive an input string and must determine whether it is a member of the language the automaton is supposed to recognize. First, in order to make these decisions, you have to figure out what you need to remember about the string as you are reading it. For many languages, you don't need to remember the entire input. You need to remember only certain crucial information. Exactly which information is crucial depends on the particular language considered.

Once you have determined the necessary information to remember about the string as it is being read, you represent this information as a finite list of possibilities. Then you assign a state to each of the possibilities. Next, you assign the transitions by seeing how to go from one possibility to another upon reading a symbol. Next, you set the start state to be the state corresponding to the possibility associated with having seen 0 symbols so far. Last, set the accept states to be those corresponding to the possibilities where you want to accept the input string.

2.1.4 The Regular Operations

In the theory of computation, the objects are languages and the tools include operations specifically designed for manipulating them. We define three operations on languages, called the **regular operations**, and use them to study properties of the regular languages.

[Regular Operations] Let A and B be languages. We define the regular operations **union**, **concatenation**, and **star** as follows:

- **Union:** $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$
- **Concatenation:** $A \circ B = \{xy \mid x \in A \text{ and } y \in B\}$
- **Star:** $A^* = \{x_1x_2 \dots x_k \mid k \geq 0 \text{ and each } x_i \in A\}$

The union operation takes all the strings in both A and B and lumps them together into one language. The concatenation operation attaches a string from A in front of a string from B in all possible ways to get the strings in the new language. The star operation is **unary operation** instead of a **binary operation**. It works by attaching any number of strings (including 0) in A together to get a string in the new language. Even if A doesn't contain the empty string, A^* *does* contain the empty string.

Example 2.1 Let $A = \{\text{oreo}, \text{ginger}\}$ and $B = \{\text{cookie}, \text{icecream}\}$...
 $A \cup B = \{\text{oreo}, \text{ginger}, \text{cookie}, \text{icecream}\}$
 $A \circ B = \{\text{oreocookie}, \text{oreoicecream}, \text{gingercookie}, \text{gingericecream}\}$
 $A^* = \{\epsilon, \text{oreo}, \text{ginger}, \dots\}$

Proof. Proof: Let A, B be regular languages recognized by DFAs M_1 and M_2 . Let $M_1 = (Q_1, \Sigma, \delta_1, q_1, F_1)$ and $M_2 = (Q_2, \Sigma, \delta_2, q_2, F_2)$. Define a new DFA $M = (Q, \Sigma, \delta, q, F)$:
 $Q = Q_1 \times Q_2$ each state of M is a pair of states, one from M_1 , and the other from M_2
 $\delta((r_1, r_2), a) = (\delta_1(r_1, a), \delta_2(r_2, a))$ a transition in M tracks transitions in both M_1, M_2
 $q_0 = (q_1, q_2)$ start M in the start states of M_1, M_2
 $F = (F_1 \times Q_2) \cup (Q_1 \times F_2)$ accept if one of the machines end in an accept state
 \dots A formal proof of correctness proceeds by induction on the length of the input string... \square

2.2 Nondeterminism

When a machine is in a given state and reads the next input symbol, we know what the next state will be—it is determined. We call this **deterministic** computation. In a **nondeterministic** machine, several choices may exist for the next state at any point. Nondeterminism is a generalization of determinism, so every deterministic finite automaton is automatically a nondeterministic finite

automaton.

Machine is initially in start state. From the start state, clones appear in all states reachable using only ϵ transitions. Each clone in its current state.

1. Reads the next input symbol a
2. If there is no outgoing transition labeled a : the clone dies
3. For each outgoing transition labeled a : a clone appears at the next state
4. ...

Every state of a DFA always has exactly one exiting transition arrow for each symbol in the alphabet. NFAs may violate that rule. In an NFA, a state may have zero, one, or many exiting arrows for each alphabet symbol. Additionally, in a DFA, labels on the transition arrows are symbols from the alphabet. An NFA may have arrows labeled with members of the alphabet or ϵ . Zero, one, or many arrows may exit from each state with the label ϵ .

So how does an NFA compute... Suppose that we are running an NFA on an input string and come to a state with multiple ways to proceed. After reading said symbol, the machine splits into multiple copies of itself and follows *all* the possibilities in parallel. Each copy of the machine takes one of the possible ways to proceed and continues as before. If there are subsequent choices, the machine splits again. If the next input symbol doesn't appear on any of the arrows exiting the state occupied by a copy of the machine, that copy of the machine dies, along with the branch of the computation associated with it. Finally, if *any one* of these copies of the machine is in an accept state at the end of the input, the NFA accepts the input string.

If a state with an ϵ symbol on an exiting arrow is encountered, something similar happens. Without reading any input, the machine splits into multiple copies, one following each of the exiting ϵ -labeled arrows and one staying at the current state. Then the machine proceeds nondeterministically as before.

Definition 2.4 (Nondeterministic Finite State Automata (NFA)) a 5-tuple $(Q, \Sigma, \delta, q_0, F)$, where

1. Q is a finite set of states
2. Σ is a finite alphabet
3. $\delta : Q \times \Sigma_\epsilon \rightarrow P(Q)$ is the transition function
4. $q_0 \in Q$ is the start state
5. $F \subseteq Q$ is the set of accept states

For any set Q we write $P(Q)$ to be the collection of all subsets of Q . Here $P(Q)$ is called the **power set** of Q . For any alphabet Σ , we write Σ_ϵ to be $\Sigma \cup \{\epsilon\}$. Thus, we can write the formal description of the type of the transition function in an NFA as $\delta : Q \times \Sigma_\epsilon \rightarrow P(Q)$.

Definition 2.5 (Computation for an NFA) Let $N = (Q, \Sigma, \delta, q_0, F)$ be an NFA and w a string over the alphabet Σ . Then we say that N *accepts* w if we can write w as $w = y_1 y_2 \cdots y_m$, where each y_i is a member of Σ_ϵ and a sequence of states r_0, r_1, \dots, r_m exists in Q with three conditions.

1. $r_0 = q_0$
2. $r_{i+1} \in \delta(r_i, y_{i+1})$, for $i = 0, \dots, m-1$
3. $r_m \in F$

Condition 1 says that the machine starts out in the start state. Condition 2 says that state r_{i+1} is one of the allowable next states when N is in the state r_i and reading y_{i+1} . Observe that $\delta(r_i, y_{i+1})$ is the *set* of allowable next states and so we say that r_{i+1} is a member of that set. Finally, condition 3 says that the machine accepts its input if the last state is an accept state.

2.3 Equivalence of NFAs and DFAs

Deterministic and nondeterministic finite automata recognize the same class of languages. Say that two machines are **equivalent** if they recognize the same language.

Theorem 2.1 Every nondeterministic finite automaton has an equivalent deterministic finite automaton

Lemma 1 A language is regular if and only if some nondeterministic finite automaton recognizes it.

One part of the condition states that a language is regular if some NFA recognizes it. Consequently, if an NFA recognizes some language, so does some DFA, and hence the language is regular. The condition also states that a language is regular only if some NFA recognizes it. That is, if a language is regular, some NFA must be recognizing it.

For a NFA with n states, the equivalent DFA will need to have 2^n states.

2.4 Closure under the Regular Operations

Theorem 2.2 The class of regular languages is closed under the union operation. In other words, if A, B are both regular languages then $A \cup B$ is regular as well.

Theorem 2.3 The class of regular languages is closed under the concatenation operation.

Theorem 2.4 The class of regular languages is closed under the star operation.

2.5 Regular Expressions

We can use the regular operations to build up expressions describing languages, which are called *regular expressions*. The value of a regular expression is a language.

Definition 2.6 (Formal Definition of a Regular Expression) Say that R is a *regular expression* if R is...

1. a for some a in the alphabet Σ
2. ϵ
3. \emptyset
4. $(R_1 \cup R_2)$, where R_1 and R_2 are regular expressions
5. $(R_1 \circ R_2)$, where R_1 and R_2 are regular expressions
6. (R_1^*) , where R_1 and R_2 are regular expressions

In items 1 and 2, the regular expressions a and ϵ represent the languages $\{a\}$ and $\{\epsilon\}$, respectively. In item 3, the regular expression \emptyset represents the empty language. In items 4, 5, and 6, the expressions represent the languages obtained by taking the union or concatenation of the languages R_1 and R_2 , or the star of the language R_1 , respectively.

Remark The expression ϵ represents the language containing a single string—namely, the empty string—whereas \emptyset represents the language that doesn't contain any strings.

Remark If we let R be any regular expression, we have the following identities.

- $R \cup \emptyset = R$
Adding the empty language to any other language will not change it
- $R \circ \epsilon = R$
Joining the empty string to any string will not change it

However, exchanging \emptyset and ϵ in the preceding identities may cause the equalities to fail.

- $R \cup \epsilon$ may not equal R .
For example if $R = \emptyset$, then $L(R) = \{\}$ but $L(R \cup \epsilon) = \{\epsilon\}$.
- $R \circ \emptyset$ may not equal R .
For example if $R = \emptyset$, then $L(R) = \{\}$ but $L(R \circ \emptyset) = \{\}$ (i.e. $R \cup \epsilon = R$ iff $\epsilon \in R$).
- $R \circ \emptyset = \emptyset$
- $R \circ \epsilon = R$

- $\emptyset^* = \{\epsilon\}$

Example 2.2

Let $\Sigma = \{0, 1\}$

- $L_1 = \{\text{3rd last bit is 1}\}$
 $(0 \cup 1)^* 1 (0 \cup 1) (0 \cup 1)$ or $\Sigma^* 1 \Sigma \Sigma$
- $L_2 = \{\text{strings in which every 0 is followed by at least one 1}\}$
 $(01 \cup 1)^* \text{ or } 1^* (011^*)^*$
- $L_3 = \{\text{strings starting with 0 and followed by any number of 1s OR strings with no 0}\}$
 $01^* \cup 1^* \text{ or } (0 \cup \epsilon) 1^*$
- $L_4 = \{\text{strings with lengths that are multiples of 3}\}$
 $((0 \cup 1)(0 \cup 1)(0 \cup 1))^* \text{ or } (\Sigma \Sigma \Sigma)^*$
- $L_5 = \{\text{strings that end with 01}\}$
 $\Sigma^* 01$

Elemental objects in a programming language, called *tokens*, such as the variable names and constants, may be described with regular expressions. Once the syntax of a programming language has been described with a regular expression in terms of its tokens, automatic systems can generate the *lexical analyzer*, the part of a compiler that initially process the input program.

2.6 Equivalence with Finite Automata

Regular expressions and finite automata are equivalent in their descriptive power. However, any regular expression can be converted into a finite automaton that recognizes the language it describes, and vice versa.

Theorem 2.5 A language is regular if and only if some regular expression describes it.

Lemma 2 If a language is described by a regular expression, then it is regular.

Lemma 3 If a language is regular, then it is described by a regular expression

Definition 2.7 (Generalized nondeterministic finite automaton) are nondeterministic finite automata wherein the transition arrows may have any regular expressions as labels, instead of only members of the alphabet or ϵ . The GFNA reads blocks of symbols from the input, not necessarily just one symbol at a time as in an ordinary NFA. The GFNA moves along a transition arrow connecting two states by reading a block of symbols from the input, which themselves constitute a string described by the regular expression on that arrow.

Like an NFA, but the *start state* has no incoming transition and has outgoing transitions to every state. The *accept state* has no outgoing transition, has incoming transitions from every state, and is distinct from the start state. All the *other states* have outgoing transitions to every state. Transitions are labeled with regular expressions.

Definition 2.8 (DFA to GFNA Conversion) 1. Convert the DFA into a generalized NFA (DFA $M \rightarrow$ GNFA G)

- Create a new start state with outgoing ϵ -transition to the start state of M
- Create a new accept state with incoming ϵ -transitions from every accept state of M
- If there are multiple labels on any transition, replace it with the union of the labels

2. Derive the regular expression from the GNFA.

The idea is to manipulate labels so that the label on (s, a) describes all strings accepted by the GNFA. Eliminate states one-at-a-time updating labels at each step. /

Definition 2.9 (Formal Definition of GFNA) A *generalized nondeterministic finite automaton* is a 5-tuple, $(Q, \Sigma, \delta, q_{\text{start}}, q_{\text{accept}})$, where

1. Q is the finite set of states
2. Σ is the input alphabet
3. $\delta : (Q - \{q_{\text{accept}}\}) \times (Q - \{q_{\text{start}}\}) \rightarrow R$ is transition function
4. q_{start} is the start state
5. q_{accept} is the accept state

A GFNA accepts a string w in Σ^* if $w = w_1w_2 \dots w_k$, where each w_i is in Σ^* and a sequence of states q_0, q_1, \dots, q_k exists such that

1. $q_0 = q_{\text{start}}$ is the start state,
2. $q_k = q_{\text{accept}}$ is the accept state
3. for each i , we have $w_i \in L(R_i)$, where $R_i = \delta(q_{i-1}, q_i)$; in other words, R_i is the expression on the arrow from q_{i-1} to q_i .

Theorem 2.6 For every regular language L , there exists a *unique* minimal DFA M such that $L = L(M)$. However, this is *not true* for NFAs; there is not necessarily a unique minimal DFA.

Definition 2.10 (Indistinguishable States) States p and q are indistinguishable: $p \sim q$ if for every string w , the computation starting in state p reaches an accept state if and only if the computation from q reaches an accept state.

Lemma 4 The indistinguishability relation, \sim , is an equivalence relation. **Proof...**

- **Reflexive:** $p \sim p$
- **Symmetric:** $p \sim q \leftrightarrow q \sim p$

- **Transitive:** $(p \sim q \cap q \sim r) \rightarrow p \sim r$

Lemma 5 (The Distinguishability Lemma) If for some symbol $\delta(p, a) \not\sim \delta(q, a)$ then $p \not\sim q$

Definition 2.11 (Finding Distinguishable Pairs of States) .

Initialize: Create an empty set D which will store pairs of distinguishable states

Step 0:

Add all pairs (p, a) where exactly one is an accept state to D . While True do: Scan over all pairs (p, q) of states not yet distinguished. For each alphabet symbol a . If the pair $(\delta(p, a), \delta(q, a))$ is distinguished, add (p, q) to D . If no new pair of states is added to $D \dots$ break. Return D

2.7 Nonregular Languages

The **pumping lemma** states that all regular languages have a special property. If we can show that a language does not have this property, we are guaranteed that it is not regular. The property states that all strings in the language can be "pumped" if they are at least as long as a certain special value called the **pumping length**. That means each such string contains a section that can be repeated any number of times with the resulting string remaining in the language.

Theorem 2.7 (Pumping Lemma) If A is a regular language, then there is a number p (the pumping length) where if s is any string in A of length at least p , then s may be divided into three pieces, $s = xyz$, satisfying the following conditions:

1. for each $i \geq 0$, $xy^iz \in A$,
2. $|y| > 0$,
3. $|xy| \leq p$

Recall the notation where $|s|$ represents the length of the string s , y^i means that i copies of y are concatenated together, and y^0 equals ϵ .

When s is divided into xyz , either x or z may be ϵ , but condition 2 says that $y \neq \epsilon$. Observe that without condition 2 the theorem would be trivially true. Condition 3 states that the pieces x and y together have length at most p . It is an exact technical condition that we occasionally find useful when proving certain languages to be nonregular.

3 Context-Free Languages

In this chapter, we present **context-free grammars**, a more powerful method of describing languages. Such grammars can describe features that have a recursive structure, which makes them useful in a variety of applications.

The collection of languages associated with context-free grammars are called the **context-free languages**. They include all the regular languages of context-free grammars and study the properties of context-free languages. We also introduce **pushdown automata**, a class of machines recognizing the context-free languages.

3.1 Context-Free Grammars

A grammar consists of a collection of **substitution rules**, also called **productions**. Each rule appears as a line in the grammar, comprising a symbol and a string separated by an arrow. The symbol is called a **variable**. The string consists of variables and other symbols called **terminals**. The variable symbols often are represented by capital letters. The terminals are analogous to the input alphabet and often are represented by lowercase letters, numbers, or special symbols. One variable is designated as the **start variable**. It usually occurs on the left-hand side of the topmost rule.

You use a grammar to describe a language by generating each string of that language in the following manner.

1. Write down the start variable. It is the variable on the left-hand side of the top rule, unless specified otherwise.
2. Find a variable that is written down and a rule that starts with that variable. Replace the written down variable with the right-hand side of that rule.
3. Repeat step 2 until no variables remain.

For example, grammar G_1 generates the string 000#111. The sequence of substitutions to obtain a string is called a **derivation**. A derivation of string 000#111 in grammar G_1 is

$$A \rightarrow 0A1 \rightarrow 00A11 \rightarrow 000A111 \rightarrow 000B111 \rightarrow 000\#111$$

You may also represent the same information pictorially with a **parse tree**.

Definition 3.1 (Formal Definition of a Context-Free Grammar) A **context-free grammar** is a 4-tuple (V, Σ, R, S) , where

1. V is a finite set called the **variables**
2. Σ is a finite set, disjoint from V , called the **terminals**
3. R is a finite set of **rules**, with each rule being a variable and a string of variables and terminals
4. $S \in V$ is the start variable.

If u , v , and w are strings of variables and terminals, and $A \rightarrow w$ is a rule of the grammar, we say that uAv **yields** uwv , written $uAv \rightarrow uwv$. Say that u **derives** v , written $u \rightarrow v$, if $u = v$ or if a sequence u_1, u_2, \dots, u_k exists for $k \geq 0$ and

$$u \rightarrow u_1 \rightarrow u_2 \rightarrow \dots \rightarrow u_k \rightarrow v$$

The **language of the grammar** is $\{w \in \Sigma^* \mid S \rightarrow w\}$.

3.1.1 Designing Context-Free Grammars

First, many CFLs are the union of simple CFLs. If you must construct a CFG for a CFL that you can break into simpler pieces, do so and then construct individual grammars for each piece. These individual grammars can be easily merged into a grammar for the original language by combining their rules and then adding the new rule $S \rightarrow S_1 S_2 \dots S_k$, where the variables S_i are the start variables for the individual grammars.

Second, constructing a CFG for a language that happens to be regular is easy if you can first construct a DFA for that language. You can convert any DFA into an equivalence CFG as follows. Make a variable R_i for each state q_i of the DFA. Add the rule $R_i \rightarrow \epsilon$ if q_i is an accept state of the DFA. Make R_0 the start variable of the grammar where q_0 is the start state of the machine. Verify on your own that the resulting CFG generates the same language that the DFA recognizes.

Third, certain context-free languages contain strings with two substrings that are "linked" in the sense that a machine for such a language would need to remember an unbounded amount of information about one of the substrings to verify that it corresponds properly to the other substring. This situation occurs in the language $\{0^n 1^n \mid n \geq 0\}$ because a machine would need to remember the number of 0s in order to verify that it equals the number of 1s. You can construct a CFG to handle this situation by using a rule of the form $R \rightarrow uRv$, which generates strings wherein the portion containing the u 's corresponds to the portion containing the v 's.

Finally, in more complex languages, the strings may contain certain structures that appear recursively as part of other (or the same) structures. Any time the symbol a appears, an entire parenthesized expression might appear recursively instead. To achieve this effect, place the variable symbol generating the structure in the location of the rules corresponding to where that structure may recursively appear.

3.1.2 Ambiguity

Sometimes a grammar can generate the same string in several different ways. Such a string will have several different parse trees and thus several different

meanings. This result may be undesirable for certain applications, such as programming languages, where a program should have a unique interpretation. If a grammar generates the same string in several different ways, we say that the string is derived *ambiguously* in that grammar. If a grammar generates some string ambiguously, we say that the grammar is *ambiguous*.

Definition 3.2 A string w is derived *ambiguously* in context-free grammar G if it has two or more different leftmost derivations. Grammar G is *ambiguous* if it generates some string ambiguously.

Sometimes when we have an ambiguous grammar we can find an unambiguous grammar that generates the same language. Some context-free languages, however, can be generated only by ambiguous grammars. Such languages are called *inherently ambiguous*.

3.1.3 Chomsky Normal Form

When working with context-free grammars, it is often convenient to have them in simplified form. One of the simplest and most useful forms is called the Chomsky normal form. Chomsky normal form is useful in giving algorithms for working with context-free grammars.

Definition 3.3 (Chomsky Normal Form) A context-free grammar is in *Chomsky normal form* if every rule is of the form

$$A \rightarrow BC$$

$$A \rightarrow a$$

where a is any terminal and A , B , and C are any variables—except that B and C may not be the start variable. In addition, we permit the rule $S \rightarrow \epsilon$, where S is the start variable.

Theorem 3.1 Any context-free language is generated by a context-free grammar in Chomsky normal form.