Chapter 1

Finite Automata

Another way of characterizing regular languages are finite automata (FA). We will show that the languages of finite automata are exactly the regular languages. Furthermore, we will also derive a decision procedure for equivalence of regular expressions.

1.1 Definition

A finite automaton consists of

1. finite set of states Q,
2. an alphabet Σ,
3. a starting state $s \in Q$,
4. a set of final states $F \subseteq Q$
5. and a state-transition relation δ. **?**

We define a **run** of a word $w \in \Sigma^*$ on an automaton $A = (\Sigma, Q, s, F, \delta)$ as any sequence of states σ such that $\forall\, 0 <= i < |\sigma| - 2.\, (\sigma_i, w_i, \sigma_{i-1}) \in \delta$. A word w is **accepted** by A if and only if there exists a run σ of w on A such that $\sigma_0 = s \wedge \sigma_{|\sigma|-1} \in F$. The **language** of A is exactly the set of words accepted by A and is denoted $\mathcal{L}(A)$. It will later be useful to also have a acceptance criterion defined by runs starting in a given state $x \in Q$, for which we will denote the resulting language $\mathcal{L}_x(A)$.

1.1.1 Non-Deterministic Finite Automata

Finite automata can be **non-deterministic** (NFA) in the sense that there exist multiple distinct runs for a word. This is the case if and only if δ is not functional.

Listing 1.1: Non-Deterministic Finite Automata

```
Record nfa : Type :=
  {
    nfa_state :> finType;
    nfa_s : nfa_state ;
    nfa_fin : pred nfa_state ;
    nfa_step : nfa_state -> char -> pred nfa_state
    }.
Fixpoint nfa_accept (x: A) w :=
match w with
  | [::] => nfa_fin A x
  | a::w => existsb y, (nfa_step A x a y) && nfa_accept y w
end.
Definition nfa_lang := [ pred w | nfa_accept (nfa_s A) w].
```

The acceptance criterion given here avoids the matter of runs. In many cases, this will help us with proofs by in induction on the accepted word. However, we will need runs in some of the proofs. Due to the fact that runs are not unique on NFAs, we will speak of **labeled paths** in this context. Labeled paths are paths through the graph implied by the automaton's step function. They can be valid runs for a word, but they could also be invalid by themselves in that they contain a sub-sequence of states that are not connected by edges. Therefore, we give a predicate that decides if a labeled path is a valid run for w on A. We then show that the acceptance criterion given above corresponds to the mathematical definition.

```
Fixpoint nfa_lpath x (xs : seq A) (w: word) {struct xs} :=
match xs,w with
  | y :: xs', a::w' => nfa_step A x a y && nfa_lpath y xs' w'
  | [::]    , [::]  => true
  | _       , _     => false
end.
Lemma nfa_lpath_accept x xs w:
  nfa_lpath x xs w -> nfa_fin A (last x xs) -> nfa_accept x w.
Lemma nfa_accept_lpath x w:
  nfa_accept x w -> exists xs, nfa_lpath x xs w /\ nfa_fin A (last x xs).
```

1.1.2 Deterministic Finite Automata

For functional δ, we speak of **deterministic** finite automata (DFA). In this case, we write δ as a function in our Coq development.

Listing 1.2: Deterministic Finite Automata

```
Record dfa : Type :=
  {
    dfa_state :> finType;
    dfa_s : dfa_state ;
```

```
    dfa_fin : pred dfa_state ;
    dfa_step : dfa_state -> char -> dfa_state
    }.
Fixpoint dfa_accept x w :=
match w with
  | [::] => dfa_fin A x
  | a :: w => dfa_accept (dfa_step A x a) w
end.
Definition dfa_lang := [pred w | dfa_accept (dfa_s A) w].
```

Again, we avoid runs in our formalization of the acceptance criterion. In this case, however, we can give a function that computes the unique run of a word on A. This allows us to give an alternative acceptance criterion that is closer to the mathematical definition. We also prove that both criteria are equivalent.

```
Fixpoint dfa_run' (x: A) (w: word) : seq A :=
match w with
  | [::] => [::]
  | a :: w => (dfa_step A x a) :: dfa_run' (dfa_step A x a) w
end.
Lemma dfa_run_accept x w: last x (dfa_run' x w) \in dfa_fin A = dfa_accept x w.
```

Equivalence between DFA and NFA

Deterministic and non-deterministic finite automata are equally expressive. One direction is trivial since every DFA can be seen as a NFA. We prove the other direction using the powerset construction. Given NFA A, we construct an equivalent DFA A_{det} in the following way:

$$\begin{aligned} Q_{det} &= \{q \mid q \subseteq Q\} \\ s_{det} &= \{s\} \\ F_{det} &= \{q \mid q \in Q_{det} \wedge q \cap F \neq \emptyset\} \\ \delta_{det} &= \{(p, a, \bigcup_{p \in P} \{q | (p, a, q) \in \delta\}) \mid p, q \in Q_{det}, a \in \Sigma\}. \\ A_{det} &= (Q_{det}, s_{det}, F_{det}, \delta_{det}). \end{aligned}$$

The formalization of A_{det} is straight-forward. We leave the set of states implicit.

```
Definition nfa_to_dfa :=
  {| dfa_s := set1 (nfa_s A);
    dfa_fin := [ pred X: {set A} | existsb x: A, (x \in X) && nfa_fin A x];
    dfa_step := [ fun X a => \bigcup_(x | x \in X) finset (nfa_step A x a) ]
   |}
.
```

Lemma 1.1.1. *For all powerset states X and for all states x with $x \in X$ we have that*

$$\mathcal{L}_x(A) \subseteq \mathcal{L}_X(A_{det}).$$

Proof. Let $w \in \mathcal{L}_x(A)$. We proof by induction on w that $w \in \mathcal{L}_X(A_{det})$.

- For $w = \varepsilon$ and $\varepsilon \in \mathcal{L}_x(A)$ we get $x \in F$ from $\varepsilon \in \mathcal{L}_x(A)$. From $x \in X$ we get $X \cap F \neq \emptyset$ and therefore $\varepsilon \in \mathcal{L}_X(A_{det})$.
- For $w = aw'$ and $aw' \in \mathcal{L}_x(A)$ we get y such that $w' \in \mathcal{L}_y(A)$ and $(x, a, y) \in \delta$. The latter gives us $y \in Y$ where Y is such that $(X, a, Y) \in \delta_{det}$. We apply the inductive hypothesis to y and Y and $w' \in \mathcal{L}_y(A)$, which gives us $w' \in \mathcal{L}_Y(A_{det})$. With $(X, a, Y) \in \delta_{det}$ we get $aw' \in \mathcal{L}_X(A_{det})$.

□

Lemma 1.1.2. *For all powerset states X and all words $w \in \mathcal{L}_X(A_{det})$ there exists a state x such that*

$$x \in X \wedge w \in \mathcal{L}_x(A).$$

Proof. We proof by induction on w that there exists x such that $x \in X \wedge w \in \mathcal{L}_x(A)$.

- For $w = \varepsilon$ and $\varepsilon \in \mathcal{L}_X(A_{det})$ we get $X \cap F \neq \emptyset$. Therefore, there exists x such that $x \in X$ and $x \in F$. Thus, we have $\varepsilon \in \mathcal{L}_x(A)$.
- For $w = aw'$ and $aw' \in \mathcal{L}_X(A_{det})$ we get Y such that $w' \in \mathcal{L}_Y(A_{det})$ and $(X, a, Y) \in \delta_{det}$. From the induction hypothesis we get y such that $y \in Y$ and $w' \in \mathcal{L}_y(A)$. From $y \in Y$ and $(X, a, Y) \in \delta_{det}$ we get x such that $x \in X$ and $(x, a, y) \in \delta$. Thus, $aw' \in \mathcal{L}_x(A)$.

□

Theorem 1.1.1. *The powerset automaton A_{det} accepts the same language as A, i.e.*

$$\mathcal{L}(A) = \mathcal{L}(A_{det}).$$

Proof. "$\subseteq$" This follows directly from lemma **??** with $x = s$ and $X = s_{det}$.

"$\supseteq$" From lemma **??** with $X = s_{det} = \{s\}$ we get $\mathcal{L}_{s_{det}}(A_{det}) \subseteq \mathcal{L}_s(A)$, which proves the claim. □

The formalization of this proof is straight-forward and follows the plan laid out above. The corresponding Lemmas are:

```
Lemma nfa_to_dfa_complete (x: A) w (X: nfa_to_dfa ):
  x \in X -> nfa_accept A x w -> dfa_accept nfa_to_dfa X w.
Lemma nfa_to_dfa_sound (X: nfa_to_dfa) w:
  dfa_accept  nfa_to_dfa  X w -> existsb x, (x \in X) && nfa_accept A x w.
Lemma nfa_to_dfa_correct w : nfa_lang  A w = dfa_lang  nfa_to_dfa  w.
```

1.2 Connected Components

Finite automata can have isolated subsets of states that are not reachable from the starting state. These states can not contribute to the language of the automaton. It will later be useful to have automata that only contain reachable states. Therefore, we define a procedure to extract the connected component from a given automaton.

Definition 1.2.1. *Let* $A = (\Sigma, Q, s, F, \delta)$ *be a DFA. We define* reachable1 *such that for all* x *and* y, $y \in \text{reachable}(x) \iff \exists a, (x, a, y) \in \delta$. *We write* reachable *to denote the transitive closure of* reachable1(s).

$$\begin{aligned} Q_c &= Q \cap reachable \\ s_c &= s \\ F_c &= F \cap reachable \\ \delta_c &= \{(x, a, y) \mid (x, a, y) \in \delta \wedge x, y \in Q_c\} \\ A_c &= (Q_c, s_c, F_c, \delta_c). \end{aligned}$$

We make use of SSReflect's *connect* predicate to extract a sequence of all states reachable from s. From this, we construct a finite type and use that as the new set of states. These new states carry a proof of reachability (ssvalP). We also have to construct a new transition function that ensures transitions always end in reachable states.

```
Definition reachable1 := [ fun x y => existsb a, dfa_step A1 x a == y ].
Definition reachable := enum (connect reachable1 (dfa_s A1)).
Lemma reachable0 : dfa_s A1 \in reachable.
Lemma reachable_step x a: x \in reachable -> dfa_step A1 x a \in reachable.
Definition dfa_connected :=
  {|
    dfa_s := {|ssvalP := reachable0 |};
    dfa_fin := [fun x => match x with {|ssval := x |} => dfa_fin A1 x end];
    dfa_step := [fun x a => match x with
      | {|ssvalP := Hx|} => {| ssvalP := (reachable_step _ a Hx) |}
      end]
  |}.
```

Lemma 1.2.1. *For every state* $x \in$ reachable *we have that*

$$\mathcal{L}_x(A_c) = \mathcal{L}_x(A).$$

Proof. "$\subseteq$" Trivial. "$\supseteq$" We do an induction on w.

- For $w = \varepsilon$ we have $\varepsilon \in \mathcal{L}_x(A)$ and therefore $x \in F$. With $x \in reachable$ we get $x \in F_c$. Thus, $\varepsilon \in \mathcal{L}_x(A_c)$.

- For $w = aw'$ we have have $y \in Q$ such that $(x, a, y) \in \delta$ and $w' \in \mathcal{L}_y(A)$. From $x \in$ reachable we get $y \in$ reachable by transitivity. Therefore, $(x, a, y) \in \delta_c$. The induction hypothesis gives us $w' \in \mathcal{L}_y(A_c)$. Thus, $aw' \in \mathcal{L}_x(A_c)$.

□

Theorem 1.2.1. *The language of the connected automaton A_c is identical to that of the original automaton A, i.e.*

$$\mathcal{L}(A) = \mathcal{L}(A_c).$$

Proof. By reflexivity, we have $s \in$ reachable. We use lemma **??** with $x = s$ to prove the claim. □

The formalization of lemma **??** and theorem **??** is straight-forward.

```
Lemma dfa_connected_correct' x (Hx: x \in reachable) :
  dfa_accept dfa_connected {|ssvalP := Hx|} =1 dfa_accept A1 x.
Lemma dfa_connected_correct: dfa_lang dfa_connected =1 dfa_lang A1.
```

To make use of the fact that A_c is fully connected, we will proof the characteristic property of A_c.

Definition 1.2.2. *A **representative** of a state x is a word w such that the unique run of w on A_c ends in x.*

Sigma?

Lemma 1.2.2. *We can give a representative for every state $x \in Q_c$.*

Proof. x carries a proof of reachability. This is equivalent to a path through the graph implied by reachable1 that ends ind x. We build the representative by extracting the edges taken in the path and building a word from those. □

The formalization of theorem **??** consists of a more general version of the theorem, which facilitates the proof by induction over the path, and the theorem itself.

```
Lemma dfa_connected_repr' (x y: dfa_connected):
  connect reachable1_connected y x ->
  exists w, last y (dfa_run' dfa_connected y w) = x.
Lemma dfa_connected_repr x :
  exists w, last (dfa_s dfa_connected) (dfa_run dfa_connected w) = x.
```

1.3 Emptiness

Given an automaton A, we can check if $\mathcal{L}(A) = \emptyset$. We simply obtain the connected automaton of A and check if there are any final states left.

Lemma 1.3.1. *The language of the connected automaton A_c is empty if and only if its set of final states F_c is empty, i.e.*

$$\mathcal{L}(A_c) = \emptyset \iff F_c = \emptyset.$$

Proof. "$\Leftarrow$" We have $\mathcal{L}(A_c) = \emptyset$ and have to show that for all $x \in Q_c$, $x \notin F_c$. Let $x \in Q_c$. By lemma **??** we get w such that the unique run of w on A_c ends in x. We use $\mathcal{L}(A_c) = \emptyset$ to get $w \notin \mathcal{L}(A_c)$, which implies that the run of w on A_c ends in a non-final state. By substituting the last state of the run by x we get $x \notin F_c$. "$\Rightarrow$" We have $F_c = \emptyset$ and have to show that for all words w, $w \notin \mathcal{L}(A_c)$. We use $F_c = \emptyset$ to show that the last state of the run of w on A_c is non-final. Thus, $w \notin \mathcal{L}(A_c)$. □

Theorem 1.3.1. *We can decide emptiness of $\mathcal{L}(A)$ by computing the cardinality of A_c's set of final states, i.e.*

$$F_c = \emptyset \iff \mathcal{L}(A) = \emptyset.$$

Proof. We use theorem **??** to substitute $\mathcal{L}(A)$ by $\mathcal{L}(A_c)$. We then use lemma **??** to prove the claim. □

The formalization of lemma **??** is split in two parts to facilitate its application.

Definition dfa_lang_empty := #|dfa_fin dfa_connected| == 0.
Lemma dfa_lang_empty_complete: dfa_lang dfa_connected =1 pred0 -> dfa_lang_empty.
Lemma dfa_lang_empty_sound: dfa_lang_empty -> dfa_lang dfa_connected =1 pred0.
Lemma dfa_lang_empty_correct: dfa_lang_empty <-> dfa_lang A1 =1 pred0.

1.4 Deciding Equivalence of Finite Automata

Definition 1.4.1. *We say that two automata are **equivalent** if and only if their languages are equal.*

Given finite automata A_1 and A_2, we construct DFA A such that the language of A is the symmetric difference of the languages of A_1 and A_2, i.e.,

$$\mathcal{L}(A) = \mathcal{L}(A_1) \ominus \mathcal{L}(A_2) = \mathcal{L}(A_1) \cap \neg\mathcal{L}(A_2) \cup \mathcal{L}(A_2) \cap \neg\mathcal{L}(A_1).$$

Theorem 1.4.1. *The equivalence of A_1 and A_2 is decidable, i.e.*

$$\mathcal{L}(A_1) = \mathcal{L}(A_2) \text{ if and only if } \mathcal{L}(A) \text{ is empty.}$$

Proof. The correctness of this procedure follows from the properties of the symmetric difference operator, i.e.

$$\mathcal{L}(A_1) \ominus \mathcal{L}(A_2) = \emptyset \Leftrightarrow \mathcal{L}(A_1) = \mathcal{L}(A_2).$$

The decidability of this procedure follows directly from theorem **??**. □

Listing 1.3: Formalization of theorem **??**

```
Definition dfa_sym_diff :=
    dfa_disj (dfa_conj A1 (dfa_compl A2)) (dfa_conj A2 (dfa_compl A1)).
Lemma dfa_sym_diff_correct:
    dfa_lang dfa_sym_diff =1 pred0 <-> dfa_lang A1 =1 dfa_lang A2.
```

1.5 Regular Expressions and Finite Automata

We prove that there is a finite automaton for every extended regular expression and vice versa. In fact, we can give a standard regular expression for every finite automaton. With this, we will prove that extended regular expressions are equivalent to standard regular expressions, thereby proving closure under intersection and negation.

1.5.1 Regular Expressions to Finite Automata

We prove that there exists an equivalent automaton for every extended regular expressions. The structure of this proof is given by the inductive nature of regular expressions. For every constructor, we provide an equivalent automaton.

Include all proofs?

1.5.2 Deciding Equivalence of Regular Expressions

Based on our procedure to construct an equivalent automaton from a regular expression, we can decide equivalence of regular expressions. Given r_1 and r_2, we construct equivalent DFA A_1 and A_2 as above.

1.5.3 Finite Automata to Regular Expressions

We prove that there is an equivalent standard regular expression for every finite automaton.

Since we are given an automaton it is not obvious how to partition our proof obligations into smaller parts. We use Kleene's original proof, the transitive closure method. This method recursively constructs a regular expression that is equivalent to the given automaton. Given a DFA A, we first assign some ordering to its states. We then define $R^k_{i,j}$ such that $\mathcal{L}(R^k_{i,j})$ is the set of all words that have a run on A starting in state i that ends in state j without ever leaving a state smaller than k. The base case $R^0_{i,j}$ is the set of all singleton words that are edges between state i and j, and ε if $i = j$. Given $R^k_{i,j}$ we can easily define $R^{k+1}_{i,j}$ based on the observation that only one new state has to be considered:

Insert complete formal definition

$$R^{k+1}_{i,j} = R^k_{i,k} \cdot (R^k_{k,k})^* \cdot R^k_{k,j} + R^k_{i,j}.$$

We make use of SSReflect's ordinals to get an ordering on states. We chose to employ ordinals for i and j, but not for k. This simplifies the inductive definitions on k. It does, however, lead to explicit conversions when k is used in place of i or j. In fact, i and j are states in our Coq implementation. We only rely on ordinals for comparison to k.

Add implementation of R

Furthermore, we define $L^k_{i,j} \subseteq \mathcal{L}(A)$ in terms of runs on the automaton. The relation of $L^k_{i,j}$ to $\mathcal{L}(A)$ can be proven very easily. We will also prove it equivalent to $R^k_{i,j}$. This allows us to connect $R^k_{i,j}$ to $\mathcal{L}(A)$.

```
Definition allbutlast : pred (seq X) :=
  fun xs => all p (belast xs).
Definition L :=
  [fun X: {set A} => [fun x y: A =>
      [pred w |
       (last x (dfa_run' A x w) == y)
         && allbutlast (mem X) (dfa_run' A x w)
      ]
    ]
  ].
```

Theorem 1.5.1. *We can express $\mathcal{L}(A)$ in terms of L. L is equivalent to R.*

$$\mathcal{L}(A) = \bigcup_{f \in F} L^{|Q|}_{s,f} = \mathcal{L}(\sum_{f \in F} R^{|Q|}_{s,f}).$$

Proof. By definition, every $w \in \mathcal{L}(A)$ has a run that ends in some $f \in F$. Then, by definition, $w \in L^{|Q|}_{s,f}$.

It remains to show that $\mathcal{L}(R^k_{i,j}) = L^k_{i.j}$. This claim can be proven by induction over k. We begin with the inclusion of $\mathcal{L}(R^k_{i,j})$ in $L^k_{i,j}$. For $k = 0$,

we do a case distinction on $i == j$ and unfold R. The resulting three cases ($i == j \wedge w = \varepsilon$, $i == j \wedge |w| = 1$, $i <> j \wedge |w| = 1$) are easily closed.

The inductive step has two cases: A triple concatenation and a simple recursion. The second case is solved by the inductive hypothesis. In the firs case, we split up the concatenation such that

$$w = w_1 \cdot w_2 \cdot w_3 \wedge w_1 \in \mathcal{L}(R^k_{i,k}) \wedge w_2 \in \mathcal{L}((R^k_{k,k})^*) \wedge w_3 \in \mathcal{L}(R^k_{k,j}).$$

The induction hypothesis is applied to w_1 and w_3 to get $w_1 \in L^k_{i,k}$ and $w_3 \in L^k_{k,j}$. We use a lemma by Coquand and Siles that splits w_2 into a sequence of words from $\mathcal{L}(R^k_{k,k})$ to which we can apply the induction hypothesis. Two concatenation lemmas for L are used to merge the sequence of words proven to be in $L^k_{k,k}$, w_1 and w_3. This shows $\mathcal{L}(R^k_{i,j}) \subseteq L^k_{i,j}$.

Next, we show the inclusion of $L^k_{i,j}$ in $\mathcal{L}(R^k_{i,j})$, again by induction over k. The base case is solved by case distinction on $i == j$. The inductive step requires a **splitting lemma** for L which shows that every non-empty word in $L^{k+1}_{i,j}$ is either in $L^k_{i,j}$ or has a non-empty prefix in $L^k_{i,k}$ and a corresponding suffix in $L^{k+1}_{k,j}$. The In the first case, we can apply the induction hypothesis. In the second case, we use size induction on the word, apply the original induction hypothesis to the prefix and the size induction hypothesis to the suffix. We use two concatenation lemmas for R to merge the sub-expression. This finishes the proof. □

Formalizing theorem **??** requires infrastructure to deal with *allbutlast*. Once this is in place, we can formalize the concatenation lemmas for R and L. These are required later to connect sub-results.

```
Lemma L_catL X x y z w1 w2:
      w1 \in L^X x z ->
      w2 \in L^(z |: X) z y ->
      w1++w2 \in L^(z |: X) x y.
```

```
Lemma L_catL X x y z w1 w2:
      w1 \in L^X x z ->
      w2 \in L^(z |: X) z y ->
      w1++w2 \in L^(z |: X) x y.
```

We also need the splitting lemma mentioned earlier. This is quite intricate. We could split right after the first character and thereby simplify the lemma. However, the current form has the advantage of requiring simple concatenation lemmas.

```
Lemma L_split X x y z a w:
      (a::w) \in L^(z |: X) x y ->
```

```
(a :: w) \in L^X x y \/
exists w1, exists w2,
  a :: w = w1 ++ w2 /\
  w1 != [::] /\
  w1 \in L^X x z /\
  w2 \in L^(z |: X) z y.
```

These lemmas suffice to show the claim of theorem **??**.

Lemma L_R n (X: {set A}) x y: #|X| = n -> L^X x y =i R^X x y.

Fix this mess