Where Are We?

We have a definition of recognizable languages in terms of deterministic finite automata (DFAs).

There are two killer apps for recognizable languages:
- pattern matching and
decision algorithms.

In order to get a better understanding of recognizable languages, it turns out that other characterizations can be very useful, both from the theory perspective as well as for the construction of algorithms.

Language Operations

The key to many of these algorithms is the fact that recognizable languages are closed under a great many more operations. To begin with, there is closure under Boolean operations:
- union
- intersection
- complement

This is useful for pattern matching but also for decision algorithms (one can deal with propositional logic; in a while, we will see how to handle quantifiers).

More Operations

Here are some more operations on languages that do not affect recognizability:
- reversal
- concatenation
- Kleene star
- homomorphisms
- inverse homomorphisms

Alas, it is difficult to establish these properties within the framework of DFAs: the constructions of the corresponding machines become rather complicated.

One elegant way to avoid these problems is to generalize our machine model to allow for nondeterminism, and show that the general machines still only accept recognizable languages.

Effective Closure

Right now, our only hope to prove a closure result is to argue

Given DFAs $A_i$ for recognizable languages $L_i$, one can effectively construct a new DFA $A$ for $L_1 \cup \cdots \cup L_n$.

So we have effective closure: there are algorithms that compute the appropriate machines.

And, in many interesting cases, these algorithms for FSMs are in fact highly efficient.
**Why Not DFAs?**

DFAs are great when it comes to the recognition problem: one can check very easily whether a DFA accepts some input. Alas, this is useless for the logic applications.

The problem with DFAs is that they are overly constrained: it can be difficult to construct an appropriate DFA for a given task.

As it turns out, even when the construction is simple in principle, the machine may be exponentially large—and ruin any hope for a usable algorithm. It is often better to deal with nondeterministic machines.

**Nondeterministic FSMs**

Here is a straightforward generalization of DFAs that allows for nondeterministic behavior. Recall that transition systems may well be nondeterministic.

**Definition**

A nondeterministic finite automaton (NFA) is a structure 

\[ A = (Q, \Sigma, \tau, I, F) \]

where \( (Q, \Sigma) \) is a transition system and the acceptance condition is given by \( I, F \subseteq Q \), the initial and final states, respectively.

So in general there is no unique next state in an NFA: there may be no next state, or there may be many. Of course, we can think of a DFA as a special type of NFA.

Some authors insist that \( I = \{ q_0 \} \). This makes no sense.

**Traces and Runs**

It is straightforward to lift the definition of acceptance from DFAs to NFAs (it all comes down to path existence, anyway).

Recall that in any transition system \((Q, \Sigma, \tau)\) a run is an alternating sequence 

\[ \pi = p_0, a_1, p_1, \ldots, a_r, p_r \]

where \( p_i \in Q, a_i \in \Sigma \) and \( \tau(p_{i-1}, a_i, p_i) \) for all \( i = 1, \ldots, r \). \( p_0 \) is the source of the run and \( p_r \) its target. The length of \( \pi \) is \( r \).

The corresponding trace or label is the word \( a_1 a_2 \ldots a_r \).

**The Fatal Definition**

The acceptance condition is essentially the same as for DFAs, except that initial states are no longer unique (and even if they were, there could be multiple traces).

**Definition**

An NFA \( A = (Q, \Sigma, \tau, I, F) \) accepts a word \( w \in \Sigma^* \) if there is a run of \( A \) with label \( w \), source in \( I \) and target in \( F \). We write \( L(A) \) for the acceptance language of \( A \).

But note that now there may be exponentially many runs with the same label. In particular, some of the runs starting in \( I \) may end up in \( F \), others may not. There is a hidden existential quantifier here. Again: all that is needed for acceptance is one accepting run, there may be many runs that fail to lead to acceptance.

**Sources of Nondeterminism**

Note that nondeterminism can arise from two different sources:

- **Transition nondeterminism:** there are different transitions \( p \xrightarrow{a} q \) and \( p \xrightarrow{a} q' \).
- **Initial state nondeterminism:** there are multiple initial states.

In other words, even if the transition relation is deterministic we obtain a nondeterministic machine by allowing multiple initial states. Intuitively, this second type of nondeterminism is less wild.

**Autonomous Transitions Epsilon Moves**

While we are at it: there is yet another natural generalization beyond just nondeterminism: autonomous transitions, aka epsilon moves. These are transitions where no symbol is read, only the state changes. This is perfectly fine considering our Turing machines ancestors.

**Definition**

A nondeterministic finite automaton with \( \varepsilon \)-moves (NFAE) is defined like an NFA, except that the transition relation has the format \( \tau \subseteq Q \times (\Sigma \cup \{ \varepsilon \}) \times Q \).

Thus, an NFAE may perform several transitions without scanning a symbol.

Hence a trace may now be longer than the corresponding input word. Other than that, the acceptance condition is the same as for NFAs: there has to be run from an initial state to a final state.
We will encounter several occasions where it is convenient to "enlarge" the alphabet \( \Sigma \) by adding the empty word \( \varepsilon \):
\[
\Sigma_{\varepsilon} = \Sigma \cup \{ \varepsilon \}
\]
Of course, \( \varepsilon \) is not a new alphabet symbol. What’s really going on? \( \Sigma \) freely generates the monoid \( \Sigma^* \), and \( \varepsilon \) is the unit element of this monoid. We can add the unit element to the generators without changing the monoid.

Exercise

Explain why this makes no difference as far as languages are concerned.

---

Here is a perfect example of an operation that preserves recognizability, but is difficult to capture within the confines of DFAs.

Let
\[
L^{op} = \{ x^{op} | x \in L \}
\]
be the reversal of a language, \((x_1x_2 \ldots x_{n-1}x_n)^{op} = x_nx_{n-1} \ldots x_2x_1\).

The direction in which we read a string should be of supreme irrelevance, so for recognizable languages to form a reasonable class they should be closed under reversal.

Suppose \(L\) is recognizable. How would we go about constructing a machine for \(L^{op}\)?

---

It is very easy to build a DFA for \(L_{a,3} = \{ x | x_3 = a \}\).

We omit the sink to keep the diagram simple.

But \(L_{a,3}^{op} = \{ x | x_{-3} = a \} = L_{a,-3}\) is hard for DFAs: we don’t know how far from the end we are.

By flipping transitions and interchanging initial and final states we obtain a machine that looks like so:

It is clear that the new machine accepts \(L_{a,-3}\).

Of course, it’s no longer a DFA—but it is a perfectly legitimate NFA.

---

Our first order of business is to show that NFAs and NFAEs are no more powerful than DFAs in the sense that they only accept recognizable languages. Note, though, that the size of the machines may change in the conversion process, so one needs to be a bit careful.

The transformation is effective: the key algorithms are

- **Epsilon Elimination** Convert an NFAE into an equivalent NFA.
- **Determinization** Convert an NFA into an equivalent DFA.
For the time being, we will refer to DFAs, NFAs and NFAEs simply as finite automata.

Strictly speaking, all three types are distinct, but there are natural inclusions

\[ \text{DFA} \subseteq \text{NFA} \subseteq \text{NFAE} \subseteq \text{GFA} \]

The heart of the OO fanatic now beats faster . . .

Epsilon elimination is quite straightforward and can easily be handled in polynomial time:

- introduce new ordinary transitions that have the same effect as chains of \( \varepsilon \) transitions, and
- remove all \( \varepsilon \)-transitions.

Since there may be chains of \( \varepsilon \)-transitions this is in essence a transitive closure problem. Hence part I of the algorithm can be handled with the usual graph techniques.

A transitive closure problem: we have to replace chains of transitions

\[
\begin{array}{c}
\text{a} \\
\varepsilon \\
\varepsilon \\
\varepsilon \\
\varepsilon
\end{array}
\]

by new transitions

\[
\begin{array}{c}
\text{a} \\
\text{a} \\
\text{a} \\
\text{a}
\end{array}
\]

Theorem
For every NFAE there is an equivalent NFA.

Proof. This requires no new states, only a change in transitions. Suppose \( \mathcal{A} = (Q, \Sigma, \tau, I, F) \) is an NFAE for \( L \). Let

\[
\mathcal{A}' = (Q, \Sigma, \tau', I', F)
\]

where \( \tau' \) is obtained from \( \tau \) as on the last slide. \( I' \) is the \( \varepsilon \)-closure of \( I \): all states reachable from \( I \) using only \( \varepsilon \)-transitions.

Again, there may be quadratic blow-up in the number of transitions and it may well be worth the effort to try to construct the NFAE in such a way that this blow-up does not occur.

In 1959, Rabin and Scott wrote the seminal paper in automata theory

M. Rabin, D. Scott

Finite Automata and Their Decision Problems

IBM Journal of Research and Development

Volume 3, Number 2, Page 114 (1959)

This paper introduces nondeterminism and the systematic study of decision problems associated with finite state machines. It’s a must-read classic.

In the realm of finite state machines, nondeterministic machines are no more powerful than deterministic ones (this is also true for Turing machines, but fails for pushdown automata).

Theorem (Rabin, Scott 1959)
For every NFA there is an equivalent DFA.

The idea is to keep track of the set of possible states the NFA could be in. This produces a DFA whose states are sets of states of the original machine.
The transition relation in an NFA has the form
\[ \tau \subseteq Q \times \Sigma \times Q \]
By GAN we can think of it as a function:
\[ \tau : Q \times \Sigma \rightarrow \mathcal{P}(Q) \]
and this function naturally extends to
\[ \tau : \mathcal{P}(Q) \times \Sigma \rightarrow \mathcal{P}(Q) \]
The latter function can be interpreted as the transition function of a DFA on \( \mathcal{P}(Q) \). Done.

By GAN we can think of it as a function:
\[ \tau : Q \times \Sigma \rightarrow \mathcal{P}(Q) \]
and this function naturally extends to
\[ \tau : \mathcal{P}(Q) \times \Sigma \rightarrow \mathcal{P}(Q) \]
The latter function can be interpreted as the transition function of a DFA on \( \mathcal{P}(Q) \). Done.

:-)
Think of \( G = \langle A; f_1, f_2, \ldots, f_k \rangle \) as a labeled digraph with edges \( p \rightarrow q \) for \( f_i(p) = q \), the virtual graph (or ambient graph) where we live.

We need to compute the reachable part of \( B \) in this graph \( G \). This can be done using standard algorithms such as Depth-First-Search or Breadth-First-Search. The only difference is that we are not given an adjacency list representation of \( G \): we compute edges on the fly. No problem at all.

This is very important when the ambient graph is huge: we may only need to touch a small part.

### Power Automaton Algorithm

Here is a slightly more hacky version of this construction.

```plaintext
active = QQ = {I};
while( active != empty )
  P = active.extract();
  foreach a in Sigma do
    compute R = tau(P,a)
    keep track of transition P to R
  if( R notin QQ ) then
    add R to QQ and active
```

Upon completion, \( QQ \subseteq \mathcal{P}(Q) \) is the state set of the accessible part of the full power automaton. We write \( \text{pow}(A) \) for this machine.

### Example: \( L_{a,-3} \)

Recall \( L_{a,k} = \{ x \in \{a,b\}^* \mid x_k = a \} \).

For negative \( k \) this means: \(-k\)th symbol from the end. It is trivial to construct an NFA for \( L_{a,-3} \):

![NFA for L_{a,-3}](image)

Note that the full power set has size 16, our construction only builds the accessible part (which happens to have size 8).

### Acceptance Testing

Recall one of the key applications of FSMs: acceptance testing is very fast and can be used to deal with pattern matching problems.

How much of a computational hit do we take when we switch to nondeterministic machines?

We can use the same approach as in determinization: instead of computing all possible sets of states reachable from \( I \), we only compute the ones that actually occur along a particular trace given by some input word.
Here is a natural modification of the DFA acceptance testing program.

```c
P = I;
while( a = x.next() ) // next input symbol
    P = τ_a(P);
return ( P intersect F != empty );
```

The update step uses the same maps $τ_a : Q \rightarrow Q$ as in the Rabin-Scott construction.

Exercise

Think of dirty tricks like hashing to speed things up.

---

**Running Time**

- The loop executes $|x|$ times, just as with DFAs.
- Unfortunately, the loop body is no longer constant time: we have to update a set of states $P \subseteq Q$.
- This can certainly be done in $O(|Q|^2)$ steps though smart data structures may sometimes give better performance.
- Actually, it seems that in practice (i.e., in NFAs that appear naturally in some application such as pattern matching) one often deals with overhead that is linear in $|Q|$ rather than quadratic.
- At any rate, we can check acceptance in an NFA in $O(|x||Q|^2)$ steps. For fixed machines this is still linear in $x$, but the hidden constant may be significant.

---

**A Better Mousetrap?**

Acceptance testing is slower, nondeterministic machines are not simply all-round superior to DFAs.

- **Advantages:**
  - Easier to construct and manipulate.
  - Sometimes exponentially smaller.
  - Sometimes algorithms much easier.

- **Drawbacks:**
  - Acceptance testing slower.
  - Sometimes algorithms more complicated.

Which type of machine to choose in a particular application can be a hard question, there is no easy general answer.

---

**Basic Operations**

Suppose we have two NFAs over $\Sigma$: $A_i = (Q_i, \tau_i; I_i, F_i)$. We may safely assume that the state sets are disjoint. There are two simple operations the combine the computations of both machines:

- **Sum Automaton**
  
  $$A_1 + A_2 = (Q_1 \cup Q_2, \tau_1 \cup \tau_2; I_1 \cup I_2, F_1 \cup F_2).$$

- **Cartesian Product Automaton**

  $$A_1 \times A_2 = (Q_1 \times Q_2, \tau_1 \times \tau_2; I_1 \times I_2, F_1 \times F_2)$$

In the product construction

$$(p, q), (p', q') \in \tau_1 \times \tau_2 \iff \tau_1(p, a, p') \land \tau_2(q, a, q')$$

---

**Computations**

Clearly, the computations of $A_1 + A_2$ are exactly the union of the computations of $A_1$ and $A_2$.

The size of the sum automaton is linear in the size of the components.

The computations of $A_1 \times A_2$ are the computations of $A_1$ combined with the computations of $A_2$ provided both have the same label: essentially, we are running both machines in parallel.

A real implementation will only construct the accessible part, but still, the size of $A_1 \times A_2$ is potentially quadratic in the sizes of $A_1$ and $A_2$. This causes problems if a product machine construction is used repeatedly.
By our choice of acceptance condition we have
\[ L(A_1 + A_2) = L_1 \cup L_2 \]
\[ L(A_1 \times A_2) = L_1 \cap L_2 \]
By changing the final states in the product, we can also get union:
\[ \text{union } F = F_1 \times Q_2 \cup Q_1 \times F_2 \]
\[ \text{intersection } F = F_1 \times F_2 \]
Why bother with a quadratic product for union when we can get it cheaper from a linear sum?

In the case where we are dealing with DFAs \(A_1\) and \(A_2\), the product is again a DFA: the new transition function \(\delta = \delta_1 \times \delta_2\) looks like
\[ \delta((p, q), a) = (\delta_1(p, a), \delta_2(q, a)) \]
More importantly, we can also construct a machine for the complement \(L(A_1) - L(A_2)\).
\[ \text{difference } F = F_1 \times (Q_2 - F_2) \]

Dire Warning: Determinism is essential here, we will see shortly that complementation for nondeterministic machines is much harder.

Exercise
Construct a counterexample that shows that the switch-states construction in general fails to produce complements in NFAs.

For the umpteenth time: a real algorithm for product machines should not construct the full Cartesian product. Instead, one should compute closures of the appropriate initial states under the transition function/relation of the product machine.

Exercise
Figure out in detail how to do these constructions producing the accessible part only.
Observe that we actually are solving two instances of a closely related problem here:

**Problem:** Inclusion
**Instance:** Two DFAs \( A_1 \) and \( A_2 \).
**Question:** Is \( L(A_1) \subseteq L(A_2) \)?

which problem can be handled by

**Lemma**
\[
L(A_1) \subseteq L(A_2) \iff L(A_1) - L(A_2) = \emptyset.
\]

Note that for any class of languages Equivalence is decidable when Inclusion is so decidable. However, the converse may be false – but it’s not so easy to come up with an example.

**Definition**
Given two languages \( L_1, L_2 \subseteq \Sigma^* \) their concatenation (or product) is defined by
\[
L_1 \cdot L_2 = \{ xy \mid x \in L_1, y \in L_2 \}.
\]

Let \( L \) be a language. The **powers** of \( L \) are the languages obtained by repeated concatenation:
\[
L^0 = \{ \epsilon \} \\
L^{k+1} = L^k \cdot L
\]

The **Kleene star** of \( L \) is the language
\[
L^* = L^0 \cup L^1 \cup L^2 \ldots \cup L^n \cup \ldots
\]

Kleene star corresponds roughly to a while-loop or iteration.

At any rate, we have established a critical closure property for recognizable languages:

**Lemma**
Recognizable languages form a Boolean algebra: they are closed under union, intersection and complement. Moreover, the closure is effective and even polynomial time for DFAs.

Effective here means that, given two machines \( A_1 \) and \( A_2 \), we can compute a new machine for \( L(A_1) \cap L(A_2) \), \( L(A_1) \cup L(A_2) \) and \( L(A_1) - L(A_2) \). If the machines are DFAs all operations are polynomial time (but complement may blow up for NFAs).

We will have more to say about the complexity of the corresponding algorithms; this is critical for applications.
Star Examples 55

Example
\{a, b\}'\*: all words over \{a, b\}

Example
\{a, b\}'\* \{a\} \{a, b\}'\*: all words over \{a, b\} containing at least two a's

Example
\{e, a, aa\} \{b, ba, baa\}'\*: all words over \{a, b\} not containing a subword aaa

Example
\{0, 1\}'\* \{0\}: all numbers in binary, no leading 0's

Example
\{1\}'\* \{0\}: all numbers in binary, with leading 0's

Concatenation is Hard 57

Now suppose we have two DFAs \(A_1\) and \(A_2\) for \(L_1\) and \(L_2\).
We can get a DFA from the previous machine, but can we build a DFA for \(L_1 \cdot L_2\) directly?

The problem is that given a word \(w\) we need to split it as \(w = xy\) and then feed \(x\) to \(A_1\) and \(y\) to \(A_2\). But there are \(|w| + 1\) many ways to do the split, and we have a priori no idea where the break should be.

One can also think of this as a guess and verify problem: guess \(x\) and \(y\), and then check that indeed \(A_1\) accepts \(x\), and \(A_2\) accepts \(y\).

Of course, there is a slight problem: DFAs don’t know how to guess.

Pebbles 60

Another way of thinking about this is to place pebbles on the states.

- Initially, each state in \(I\) has a pebble.
- Under input \(a\), a pebble on \(p\) multiplies and moves to all \(q\) such that \(p \xrightarrow{a} q\).
- Multiple pebbles on a state are condensed into a single one.
- We accept whenever a pebble appears in \(F\).

Pebbles are very helpful in particular in the direct construction of DFAs: the movement of the set of all pebbles is perfectly deterministic.
We start with one copy of DFA $A_1$, the master, and one copy of DFA $A_2$, the slave.

- Place one pebble on the initial state of the master machine.
- Move this and all other pebbles along transitions according to standard rules.
- Whenever the master pebble reaches a final state, place a new pebble on the initial state of the slave automaton.
- The composite machine accepts if a pebble sits on final state in the slave machine.

So the number of states is bounded by $|A_1|^{2|A_2|}$: the $A_1$ part is deterministic but the $A_2$ part is not.

First, the machine just described is deterministic: we place and remove a collection of pebbles according to an entirely deterministic rule.

The states are of the form $(p, P)$ where $p \in Q_1$ and $P \subseteq Q_2$, corresponding to a complete record of the positions of all the pebbles.

Now let $n_i$ be the state complexity of $A_i$. Then the number of states is at most

$$n_1 2^{n_2}$$

Of course, the accessible part may well be smaller.

**Exercises**

**Exercise**
There are several gaps and inaccuracies in the outline above, fix them all.

**Exercise**
Carry out this construction for the languages $E_a =$ even number of $a$’s and $E_b =$ even number of $b$’s and run some examples.

**Exercise**
Explain why the pebbling construction really defines a DFA.

**Exercise**
Carry out a pebbling construction for Kleene star.

More generally, suppose we have DFAs $A_i$ of size $n_i$, respectively.

Then the full product machine $\mathcal{A} = A_1 \times A_2 \times \ldots \times A_{n-1} \times A_n$ has $n = n_1 n_2 \ldots n_n$ states.

- The full product machine grows exponentially, but its accessible part may be much smaller.
- Alas, there are cases where exponential blow-up cannot be avoided.

**Aside: Complicated Intersections**

Product constructions are important even for relatively simple languages, it can be quite difficult to build automata for, say, the intersection of two recognizable languages directly by hand.

Here is an example: build a DFA for the language of all words that contain the scattered subword (not factor) $ab$ 3 times, and a multiple-of-3 number of $a$’s.

Building the two component machines and taking their product we get

```
1 -> 2 -> 3 -> 4 -> 5
6 -> 7 -> 8 -> 9
10
```

Here is the Emptiness Problem for a list of DFAs rather than just a single machine:

Problem: **DFA Intersection**
Instance: A list $A_1, \ldots, A_n$ of DFAs
Question: Is $\bigcap L(A_i)$ empty?

This is easily decidable: we can check Emptiness on the product machine $\mathcal{A} = \prod A_i$. The Emptiness algorithm is linear, but it is linear in the size of $\mathcal{A}$, which is itself exponential. And, there is no universal fix for this:

**Theorem**
The DFA Intersection Problem is PSPACE-hard.