Formal Languages and Automata

Professor Frank Stajano

UNIVERSITY OF CAMBRIDGE

(Dr Tom Sterling)

5 lectures for
2022–2023 (3-2024) Computer Science Tripos
Part IA Discrete Mathematics

Originally written by **Professor Andrew Pitts**

Minor tweaks by Prof. Ian Leslie and Prof. Frank Stajano

Revision 2 of 2023-01-07 17:50:09 +0000 (Sat, 07 Jan 2023)

Contents and Syllabus

- Formal Languages 4
- Inductive Definitions and Rule Induction 11
- Regular expressions and Pattern Matching 24
- Finite Automata 41
- Regular Languages and Kleene's Theorem 63
- The Pumping Lemma 98

Common theme: mathematical techniques for defining formal languages and reasoning about their properties.

Key concepts: inductive definitions, automata

Relevant to:

Part IB	Compiler Construction, Computation Theory, Complexity Theory, Semantics of Programming Languages
Part II	Natural Language Processing, Optimising Compilers, Denotational Semantics, Hoare Logic and Model Checking

Formal Languages

Alphabets

An **alphabet** is specified by giving a finite set, Σ, whose elements are called **symbols**. For us, any set qualifies as a possible alphabet, so long as it is finite.

Examples:

- $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, **10**-element set of decimal digits.
- $\{a, b, c, \ldots, x, y, z\}$, **26**-element set of lower-case characters of the English language.
- $\{S \mid S \subseteq \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}\}$, $\mathbf{2^{10}}$-element set of all subsets of the alphabet of decimal digits.

Non-example:

- $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$, set of all non-negative whole numbers is not an alphabet, because it is infinite.

Strings over an alphabet

A **string of length** n (for $n = 0, 1, 2, \ldots$) over an alphabet Σ is just an ordered n-tuple of elements of Σ, written without punctuation.

Σ^* denotes the set of all strings over Σ of any finite length.

Examples:

notation for the string of length 0

- If $\Sigma = \{a, b, c\}$, then ε, a, ab, aac, and $bbac$ are strings over Σ of lengths zero, one, two, three and four respectively.
- If $\Sigma = \{a\}$, then Σ^* contains ε, a, aa, aaa, $aaaa$, etc.
- If $\Sigma = \emptyset$ (the empty set), then $\Sigma^* = \{\varepsilon\}$.

Notes

- There is a unique string of length zero over Σ, called the **null string** (or **empty string**) and denoted ε, no matter which alphabet Σ we are talking about.
- We make no notational distinction between a symbol $a \in \Sigma$ and the string of length 1 containing a. Thus we regard Σ as a subset of Σ^*.
- $\emptyset$, $\{\varepsilon\}$ and ε are three different things!
  - $\emptyset$ is the (unique) set with no elements,
  - $\{\varepsilon\}$ is a set with one element (the null string),
  - ε is the string of length 0.
- The length of a string $u \in \Sigma^*$ is denoted $|u|$.
- We are not concerned here with data structures and algorithms for implementing strings (so strings and finite lists are interchangeable concepts here).
- **Warning!** the symbol * is highly overloaded – it means different things in different contexts in this course. (The same comment applies to the symbol ε and, to a lesser extent, the symbol $\emptyset$.)

(abc)

(a) vs. a

Concatenation of strings

The **concatenation** of two strings u and v is the string uv obtained by joining the strings end-to-end. This generalises to the concatenation of three or more strings.

Examples:

If $\Sigma = \{a, b, c, \ldots, z\}$ and $u, v, w \in \Sigma^*$ are $u = ab$, $v = ra$ and $w = cad$, then

$$vu = raab$$
$$uu = abab$$
$$wv = cadra$$
$$uvwuv = abracadabra$$

Notes

- Concatenation satisfies:

$$u\varepsilon = u = \varepsilon u$$
$$(uv)w = uvw = u(vw)$$
(but in general $uv \neq vu$)

$|u| = \text{len}(u)$

- $|uv| = |u| + |v|$

Notation

If $u \in \Sigma^*$, then u^n denotes n copies of u concatenated together. By convention $u^0 = \varepsilon$.

$u^0 = \varepsilon$
$u^1 = u$
$u^2 = uu$ ...

Formal languages

An extensional view of what constitutes a formal language is that it is completely determined by the set of 'words in the dictionary':

Given an alphabet Σ, we call any subset of Σ^* a (formal) **language** over the alphabet Σ.

We will use inductive definitions to describe languages in terms of grammatical rules for generating subsets of Σ^*.

1) Σ^* is a language
2) $\emptyset$ is a language

Inductive Definitions

Axioms and rules

for inductively defining a subset of a given set U (Σ^*)

- **axioms** $\dfrac{}{a}$ are specified by giving an element a of U
- **rules** $\dfrac{h_1 \; h_2 \; \cdots \; h_n}{c}$

  are specified by giving a finite subset $\{h_1, h_2, \ldots, h_n\}$ of U (the **hypotheses** of the rule) and an element c of U (the **conclusion** of the rule)

$\mathcal{L}$

$\Sigma = \{a, b, c\}$

$$\frac{u \in \mathcal{L} \quad v \in \mathcal{L}}{u\,b\,v \in \mathcal{L}}$$

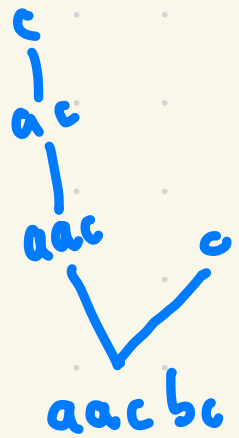
$$\frac{}{c \in \mathcal{L}} \qquad \frac{u \in \mathcal{L}}{a\,u \in \mathcal{L}}$$

$$\frac{}{\mathcal{L} \ni c,\ ac,\ aac,\ \ldots}$$

e.g.

$$\cfrac{\cfrac{c \in \mathcal{L}}{a\,c \in \mathcal{L}}}{a\,a\,c \in \mathcal{L}}$$

$$\frac{\vdots}{aac \in \mathcal{L}} \qquad \frac{}{c \in \mathcal{L}}$$
$$\overline{\qquad\qquad aacbc}$$



Derivations

Given a set of axioms and rules for inductively defining a subset of a given set U, a **derivation** (or proof) that a particular element $u \in U$ is in the subset is by definition

a finite tree with vertices labelled by elements of U and such that:

- the root of the tree is u (the conclusion of the whole derivation),
- each vertex of the tree is the conclusion of a rule whose hypotheses are the children of the node,
- each leaf of the tree is an axiom.

Example

$U = \{a, b\}^*$

axiom: $\dfrac{}{\varepsilon}$

rules: $\dfrac{u}{aub}$ $\dfrac{u}{bua}$ $\dfrac{u \quad v}{uv}$ (for all $u, v \in U$)

Example derivations:

$$\dfrac{\dfrac{\dfrac{\dfrac{\varepsilon}{\varepsilon}}{ab} \quad \dfrac{\dfrac{\varepsilon}{ab}}{aabb}}{abaabb} \qquad \dfrac{\dfrac{\dfrac{\varepsilon}{ba} \quad \dfrac{\varepsilon}{ab}}{baab}}{abaabb}}{}$$

aεb ∼

Inductively defined subsets

Given a set of axioms and rules over a set U, the subset of U **inductively defined** by the axioms and rules consists of all and only the elements $u \in U$ for which there is a derivation with conclusion u.

For example, for the axioms and rules on Slide 14

- $abaabb$ is in the subset they inductively define (as witnessed by either derivation on that slide)
- $abaab$ is not in that subset (there is no derivation with that conclusion – why?)

(In fact $u \in \{a,b\}^*$ is in the subset iff it contains the same number of a and b symbols.)

Notes

- Axioms are special cases of rules – the ones where $n = 0$, i.e. the set of hypotheses is empty.
- We are generally interested in inductive definitions of subsets that are infinite. An inductive definition with only finitely many axioms and rules defines a finite subset. (Why?) So we usually have to consider infinite sets of axioms and rules. However, those sets are usually specified *schematically*: an axiom scheme, or a rule scheme is a template involving variables that can be instantiated to get a whole family of actual axioms or rules.

  For example, on Slide 14, we used the rule scheme $\dfrac{u}{aub}$ where u is meant to be instantiated with any string over the alphabet $\{a, b\}$. Thus this rule scheme stands for the infinite collection of rules $\dfrac{\varepsilon}{ab}$, $\dfrac{a}{aab}$, $\dfrac{b}{abb}$, $\dfrac{aa}{aaab}$, etc.
- It is sometimes convenient to flatten derivations into finite lists, because they are easier to fit on a page. The last element of the list is the conclusion of the derivation. Every element of the list is either an axiom, or the conclusion of a rule all of whose hypotheses occur earlier in the list.
- The fact that an element is in an inductively defined subset may be witnessed by more than one derivation (see the example on Slide 14).
- In general, there is no sure-fire, algorithmic method for showing that an element is *not* in a particular inductively defined subset.

Example: transitive closure

Given a binary relation $R \subseteq X \times X$ on a set X, its **transitive closure** R^+ is the smallest (for subset inclusion) binary relation on X which contains R and which is **transitive** ($\forall x, y, z \in X.\ (x, y) \in R^+ \ \&\ (y, z) \in R^+ \Rightarrow (x, z) \in R^+$).

R^+ is equal to the subset of $X \times X$ inductively defined by

axioms $\dfrac{}{(x, y)}$ (for all $(x, y) \in R$)

rules $\dfrac{(x, y) \qquad (y, z)}{(x, z)}$ (for all $x, y, z \in X$)

Reflexive-transitive

$\dfrac{}{(x,x)}$ (for $x \in X$)

Example: reflexive-transitive closure

Given a binary relation $R \subseteq X \times X$ on a set X, its **reflexive-transitive closure** R^* is defined to be the smallest binary relation on X which contains R, is both transitive and **reflexive** ($\forall x \in X.\ (x,x) \in R^*$).

R^* is equal to the subset of $X \times X$ inductively defined by

axioms $\dfrac{}{(x,y)}$ (for all $(x,y) \in R$) $\qquad \dfrac{}{(x,x)}$ (for all $x \in X$)

rules $\dfrac{(x,y) \qquad (y,z)}{(x,z)}$ (for all $x,y,z \in X$)

we can use Rule Induction (Slide 19) to prove this

Rule Induction

Theorem. The subset $I \subseteq U$ inductively defined by a collection of axioms and rules is closed under them and is the least such subset: if $S \subseteq U$ is also closed under the axioms and rules, then $I \subseteq S$.

Given axioms and rules for inductively defining a subset of a set U, we say that a subset $S \subseteq U$ is **closed under the axioms and rules** if

- for every axiom $\dfrac{}{a}$, it is the case that $a \in S$
- for every rule $\dfrac{h_1 \; h_2 \; \cdots \; h_n}{c}$, if $h_1, h_2, \ldots, h_n \in S$, then $c \in S$.

Rule Induction

Theorem. The subset $I \subseteq U$ inductively defined by a collection of axioms and rules is closed under them and is the least such subset: if $S \subseteq U$ is also closed under the axioms and rules, then $I \subseteq S$.

We use a similar approach as method of proof: given a property $P(u)$ of elements of U, to prove $\forall u \in I.\, P(u)$ it suffices to show

- **base cases:** $P(a)$ holds for each axiom $\dfrac{}{a}$
- **induction steps:** $P(h_1)\ \&\ P(h_2)\ \&\ \cdots\ \&\ P(h_n) \Rightarrow P(c)$ holds for each rule $\dfrac{h_1\ h_2\ \cdots\ h_n}{c}$

(To see this, apply the theorem with $S = \{u \in U \mid P(u)\}$.)

Proof of the Theorem on Slide 19

I is closed under any of the axioms $\dfrac{}{a}$, because a is a derivation of length 1 showing that $a \in I$. I is closed under any of the rules $\dfrac{h_1 \;\cdots\; h_n}{c}$, because if each h_i is in I, there is a derivation D_i with conclusion h_i; and then $\dfrac{D_1 \;\cdots\; D_n}{c}$ is a derivation (why?) with conclusion $c \in I$.

Now suppose $S \subseteq U$ is some subset closed under the axioms and rules. We can use mathematical induction to prove

$\forall n.$ all derivations of height $\leq n$ have their conclusion in S $\quad (*)$

Hence all derivations have their conclusions in S; and therefore $I \subseteq S$, as required. □

[Proof of $(*)$ by mathematical induction:

Base case $n = 0$: trivial, because there are no derivations of height 0.

Induction step for $n+1$: suppose D is a derivation of height $\leq n+1$, with conclusion c – say $D = \dfrac{D_1 \;\cdots\; D_m}{c}$ (some $m \geq 0$). We have to show $c \in S$. Note that each D_i is a derivation of height $\leq n$ and so by induction hypothesis its conclusion, c_i say, is in S. Since D is a well-formed derivation, $\dfrac{c_1 \;\cdots\; c_m}{c}$ has to be a rule (or $m = 0$ and it is an axiom). Since S is closed under the axioms and rules and each c_i is in S, we conclude that $c \in S$. □]

Example: reflexive-transitive closure

Given a binary relation $R \subseteq X \times X$ on a set X, its **reflexive-transitive closure** R^* is defined to be the smallest binary relation on X which contains R, is both transitive and **reflexive** ($\forall x \in X.\ (x,x) \in R^*$).

R^* is equal to the subset of $X \times X$ inductively defined by

axioms $\dfrac{}{(x,y)}$ (for all $(x,y) \in R$) $\qquad \dfrac{}{(x,x)}$ (for all $x \in X$)

rules $\dfrac{(x,y) \qquad (y,z)}{(x,z)}$ (for all $x,y,z \in X$)

we can use Rule Induction (Slide 19) to prove this, since $S \subseteq X \times X$ being closed under the axioms & rules is the same as it containing R, being reflexive and being transitive.

Example using rule induction

Let I be the subset of $\{a,b\}^*$ inductively defined by the axioms and rules on Slide 14.

For $u \in \{a,b\}^*$, let $P(u)$ be the property

u contains the same number of a and b symbols

We can prove $\forall u \in I.\, P(u)$ by rule induction:

- **base case:** $P(\varepsilon)$ is true (the number of as and bs is zero!)
- **induction steps:** if $P(u)$ and $P(v)$ hold, then clearly so do $P(aub)$, $P(bua)$ and $P(uv)$.

(It's not so easy to show $\forall u \in \{a,b\}^*.\, P(u) \Rightarrow u \in I$ – rule induction for I is not much help for that.)

NOTE:
In lecture, I claimed that the converse
to the theorem of slide 23 holds. This is true, but it does not seem to follow directly by induction on the length of the string. If you are bored, try and find a way to prove this converse! — JMS