# Algebraic language theory

Mikołaj Bojańczyk

March 6, 2020
The latest version can be downloaded from:
https://www.mimuw.edu.pl/ bojan/2019-2020/algebraic-language-theory-2020

# Contents

# Preface

These are lecture notes on the algebraic approach to regular languages. The classical algebraic approach is for finite words; it uses semigroups instead of automata. However, the algebraic approach can be extended to structures beyond words, e.g. infinite words, or trees or graphs.

# PART ONE

## WORDS

# 1
# Semigroups

In this chapter, we define semigroups and monoids, and show how they can be used to recognise languages of finite words.

**Definition 1.1** (Semigroup)**.** A *semigroup* consists of an underlying set $S$ together with a binary product operation

$$(a, b) \mapsto ab,$$

that is associative in the sense that

$$a(bc) = (ab)c \qquad \text{for all } a, b, c \in S.$$

The definition says that the order of evaluation in a semigroup is not important, i.e. that different ways of parenthesising a sequence of elements in the monoid will yield the same result as far as far as the semigroup product is concerned. For example,

$$((ab)c)(d(ef)) = ((((ab)c)d)e)f.$$

Therefore, it makes sense to omit the parentheses and write simply

$$abcdef.$$

This means that the product operation in the semigroup can be seen as defined not just on pairs of semigroups elements, but also on all finite words consisting of semigroup elements.

A *semigroup homomorphism* if a function between semigroups that preserves the structure of semigroups, i.e. a function

$$h : \underbrace{S}_{\text{semigroup}} \to \underbrace{T}_{\text{semigroup}}$$

which is consistent with the product operation in the sense that

$$h(a \cdot b) = h(a) \cdot h(b),$$

where the semigroup product on the left is in $S$, and the semigroup product on the right is in $T$.

A *monoid* is the special case of a semigroup where there is an identity element, denoted by $1 \in S$, which satisfies

$$1a = a1 \qquad \text{for all } a \in S.$$

The identity element, if it exists, must be unique. This is because if there are two candidates for the identity, then taking their product reveals the true identity. A *monoid homomorphism* is a semigroup homomorphism that preserves the identity element.

**Example 1.2.** Here are some examples of monoids and semigroups.

(1) If $\Sigma$ is a set, then the set $\Sigma^+$ of nonempty words over $\Sigma$, equipped with concatenation, is a semigroup, called the *free*[1] *semigroup over generators* $\Sigma$. The *free monoid* is the set $\Sigma^*$ of possibly empty words.

(2) Every group is a monoid.

(3) For every set $Q$, the set of all functions $Q \to Q$, equipped with function composition, is a monoid. The monoid identity is the identity function.

(4) For every set $Q$, the set of all binary relations on $Q$ is a monoid, when equipped with relational composition

$$a \circ b = \{(p, q) : \text{there is some } r \in Q \text{ such that } (p, r) \in a \text{ and } (r, q) \in b\}.$$

The monoid identity is the identity function. The monoid from the previous item is a submonoid of this one, i.e. the inclusion map is a monoid homomorphism.

(5) Here are all semigroups of size two, up to semigroup isomorphism:

$$\underbrace{(\{0,1\},+)}_{\text{addition mod 2}} \qquad (\{0,1\},\times) \qquad \underbrace{(\{0,1\},\pi_1)}_{\text{product } ab \text{ is } a} \qquad \underbrace{(\{0,1\},\pi_2)}_{\text{product } ab \text{ is } b}$$

The first two are monoids.

[1] The reason for this name is the following universality property. The free semigroup is generated by $\Sigma$, and it is the biggest semigroup generated by $\Sigma$ in the following sense. For every semigroup $S$ that is generated by $\Sigma$, there exists a (unique) surjective semigroup homomorphism $h : \Sigma^+ \to S$ which is the identity on the $\Sigma$ generators.

Semigroup homomorphisms are closely related with functions that are compositional in the sense defined below. Let $S$ be a semigroup, and let $X$ be a set (without a semigroup structure). A function

$$h : S \to X$$

is called *compositional* if for every $a, b \in S$, the value $h(a \cdot b)$ is uniquely determined by the values $h(a)$ and $h(b)$. If $X$ has a semigroup structure, then every semigroup homomorphism $S \to X$ is a compositional function. The following lemma shows that the converse is also true for surjective functions.

**Lemma 1.3.** *Let $S$ be a semigroup, let $X$ be a set, and let $h : S \to X$ be a surjective compositional function. Then there exists (a unique) semigroup structure on $X$ which makes $h$ into a semigroup homomorphism.*

*Proof* Saying that $h(a \cdot b)$ is uniquely determined by $h(a)$ and $h(b)$, as in the definition of compositionality, means that there is a binary operation $\circ$ on $X$, which is not yet known to be associative, that satisfies

$$h(a \cdot b) = h(a) \circ h(b) \qquad \text{for all } a, b \in S. \tag{1.1}$$

The semigroup structure on $X$ uses $\circ$ as the semigroup operation. It remains to prove associativity of $\circ$. Consider three elements of $X$, which can be written as $h(a), h(b), h(c)$ thanks to the assumption on surjectivity of $h$. We have

$$(h(a) \circ h(b)) \circ h(c) \overset{(1.1)}{=} (h(ab)) \circ h(c) \overset{(1.1)}{=} h(abc).$$

The same reasoning shows that $h(a) \circ (h(b) \circ h(c))$ is equal to $h(abc)$, thus establishing associativity. □

**Exercise 1.** Show a function between two monoids that is a semigroup homomoprhism, but not a monoid homomorphism.

**Exercise 2.** Show that there are exponentially many semigroups of size $n$.

**Exercise 3.** Show that for every semigroup homomorphism $h : \Sigma^+ \to S$, with $S$ finite, there exists some $N \in \{1, 2, \ldots\}$ such that every word of length at least $N$ can be factorised as $w = w_1 w_2 w_3$ where $h(w_2)$ is an idempotent[2].

[2] This exercise can be seen as the semigroup version of the pumping lemma.

## Recognising languages

In this book, we are interested in monoids and semigroups as an alternative to finite automata for the purpose of recognising languages. Since languages are usually defined for possibly empty words, we use monoids and not semigroups when recognising languages.

**Definition 1.4.** Let $\Sigma$ be a finite alphabet. A language $L \subseteq \Sigma^*$ is *recognised* by a monoid homomorphism

$$h : \Sigma^* \to M$$

if membership in $w \in L$ is determined uniquely by $h(w)$. In other words, there is a subset $F \subseteq M$ such that

$$w \in L \quad \text{iff} \quad h(w) \in F \qquad \text{for every } w \in \Sigma^*.$$

We say that a language is recognised by a monoid if it is recognised by some monoid homomorphism into that monoid. The following theorem shows that, for the purpose of recognising languages, finite monoids and finite automata are equivalent.

**Theorem 1.5.** *The following conditions are equivalent for every $L \subseteq \Sigma^*$:*

(1) *$L$ is recognised by a finite nondeterministic automaton;*
(2) *$L$ is recognised by a finite monoid.*

*Proof*

**2 ⇒ 1** From a monoid homomorphism one creates a deterministic automaton, whose states are elements of the monoid, the initial state is the identity, and the transition function is

$$(m, a) \mapsto m \cdot (\text{homomorphic image of } a).$$

After reading an input word, the state of the automaton is its homomorphic image, and therefore the accepting state from the monoid homomorphisms can be used. This automaton computes the monoid product according to the choice of parentheses illustrated in this example:

$$(((((ab)c)d)e)f)g.$$

**1 ⇒ 2** Let $Q$ be the states of the nondeterministic automaton recognising $L$. Define a function[3]

$$\delta : \Sigma^* \to \text{monoid of binary relations on } Q$$

[3] This transformation from a nondeterministic (or deterministic) finite automaton to a monoid incurs an exponential blow-up, which is unavoidable in the worst case.

which sends a word $w$ to the binary relation

$$\{(p, q) \in Q^2 : \text{some run over } w \text{ goes from } p \text{ to } q\}.$$

This is a monoid homomorphism. It recognises the language: a word is in the language if and only if its image under the homomorphism contains at least one (initial, accepting) pair.

□

**Exercise 4.** Show that the translation from deterministic finite automata to monoids is exponential in the worst case.

**Exercise 5.** Show that the translation from (left-to-right) deterministic finite automata to monoids is exponential in the worst case, even if there is a right-to-left deterministic automaton of same size.

**Exercise 6.** Which languages are recognised by finite commutative monoids?

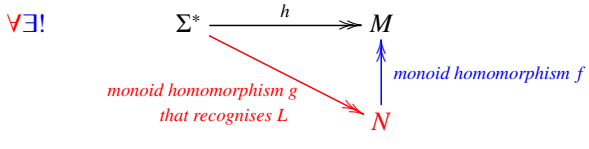
## The syntactic monoid of a language

Deterministic finite automata have minimisation, i.e. for every language there is a minimal deterministic automaton, which can be found inside every other deterministic automaton that recognises the language. The same is true for monoids, as proved in the following theorem.

**Theorem 1.6.** *For every language*[4] $L \subseteq \Sigma^*$ *there is a surjective monoid homomorphism*

$$h : \Sigma^* \to M,$$

*called the syntactic homomorphism of L, which recognises it and is minimal in the sense explained in the following quantified diagram*[5]




[4] The language need not be regular, and the alphabet need not be finite.

[5] Here is how to read the diagram. For every red extension of the black diagram there exists a unique blue extension which makes the diagram commute. Double headed arrows denote surjective homomorphisms, which means that ∀ quantifies over surjective homomorphisms, and the same is true for ∃!.

*Proof* The proof is the same as for the Myhill-Nerode theorem about minimal automata, except that the corresponding congruence is two-sided. Define the *syntactic congruence* of $L$ to be the equivalence relation $\sim$ on $\Sigma^*$ which identifies two words $w, w' \in \Sigma^*$ if

$$uwv \in L \quad \text{iff} \quad uw'v \in L \qquad \text{for all } u, v \in \Sigma^*.$$

Define $h$ to be the function that maps a word to its equivalence class under syntactic congruence. It is not hard to see that $h$ is compositional, and therefore by (the monoid version of) Lemma 1.3, one can equip the set of equivalence classes of syntactic congruences with a monoid structure – call $M$ the resulting monoid – which turns $h$ into a monoid homomorphism.

It remains to show minimality of $h$, as expressed by the diagram in the lemma. Let then $g$ be as in the diagram. Because $g$ recognises the language $L$, we have

$$g(w) = g(w') \quad \text{implies} \quad w \sim w',$$

which, thanks to surjectivity of $g$, yields some function $f$ from $N$ to $M$, which makes the diagram commute, i.e. $h = f \circ g$. Furthermore, $f$ must be a monoid homomorphism, because

$$\begin{aligned}
f(a_1 \cdot a_2) = &\quad \text{(by surjectivity of } g\text{, each } a_i \text{ can be presented as } g(w_i) \text{ for some } w_i) \\
f(g(w_1) \cdot g(w_2)) = &\quad (g \text{ is a monoid homomorphism}) \\
f(g(w_1 w_2)) = &\quad \text{(the diagram commutes)} \\
h(w_1 w_2) = &\quad (h \text{ is a monoid homomorphism}) \\
h(w_1) \cdot h(w_2) = &\quad \text{(the diagram commutes)} \\
f(g(w_1)) \cdot f(g(w_2)) &= \\
f(a_1) \cdot f(a_2). &
\end{aligned}$$

□

**Exercise 7.** Prove that surjectivity of $g$ is important in Theorem 1.6.

**Exercise 8.** Show that for every language, not necessarily regular, its syntactic homomorphism is the function

$$w \in \Sigma^* \qquad \mapsto \qquad \underbrace{(q \mapsto qw)}_{\substack{\text{state transformation} \\ \text{in the sytactic automaton}}},$$

where the syntactic automaton is the deterministic finite automaton from the Myhill-Nerode theorem.

# 2

# Green's relations and the structure of finite semigroups

In this chapter, we describe some of the structural theory of finite semigroups. This theory is based on Green's relations, which are pre-orders in a semigroup that are based on prefixes, suffixes and infixes.

We begin with idempotents, which are ubiquitous in the analysis of finite semigroups. A semigroup element $e$ is called *idempotent* if it satisfies

$$ee = e.$$

**Example 2.1.** In a group, there is a unique idempotent element, namely the group identity. There can be several idempotent elements, for example all elements are idempotent in the semigroup

$$(\{1, \ldots, n\}, \max).$$

One can think of idempotents as being a relaxed version of identity elements.

**Lemma 2.2** (Idempotent Power Lemma)**.** *Let $S$ be a finite semigroup. For every $a \in S$, there is exactly one idempotent in the set*

$$\{a^1, a^2, a^3, \ldots\} \subseteq S.$$

*Proof* Because the semigroup is finite, the sequence $a^1, a^2, \ldots$ must contain a repetition, i.e. there must exist $n, k \in \{1, 2, \ldots\}$ such that

$$a^n = a^{n+k} = a^{n+2k} = \cdots.$$

After multiplying both sides of the above equation by $a^{nk-n}$ we get

$$a^{nk} = a^{nk+k} = a^{nk+2k} = \cdots,$$

and therefore $a^{nk} = a^{nk+nk}$ is an idempotent. To prove uniqueness of the idempotent, suppose $n_1, n_2 \in \{1, 2, \ldots\}$ are powers such that that $a^{n_1}$ and $a^{n_2}$ are

idempotent. The we have

$$\underbrace{a^{n_1} = (a^{n_1})^{n_2}}_{\substack{\text{because } a^{n_1} \\ \text{is idempotent}}} = a^{n_1 n_2} = \underbrace{(a^{n_1})^{n_2} = a^{n_2}}_{\substack{\text{because } a^{n_2} \\ \text{is idempotent}}}$$

□

We use the name *idempotent power* for the unique idempotent in the above lemma. Note that it is the resulting semigroup element $a^n$ that is unique, and not the exponent $n$. Finiteness is crucial for the above lemma, for example the infinite semigroup

$$(\{1, 2, \ldots\}, +)$$

contains no idempotents. The analysis presented in the rest of this chapter will hold in any semigroup which satisfies the conclusion of the Idempotent Power Lemma.

## Green's relations

We now give the main definition of this chapter.

**Definition 2.3** (Green's relations)**.** Let $a, b$ be elements of a semigroup $S$. We say that $a$ is a *prefix* of $b$ if there exists a solution $x$ of

$$ax = b.$$

The solution $x$ can be an element of the semigroup, or empty (i.e. $a = b$). Likewise we define the suffix and infix relations, but with the equations

$$\underbrace{xa = b}_{\text{suffix}} \qquad \underbrace{xay = b}_{\text{infix}}.$$

In the case of the infix relation, one or both of $x$ and $y$ can be empty.

Figure 2.1 shows a monoid along with the accompanying Green's relations. The prefix, suffix and infix relations are pre-orders, i.e. they are transitive and reflexive[1]. They need not be anti-symmetric, for example in a group every element is an prefix (suffix, infix) of every other element. We say that two

[1] Another description of the prefix pre-order is that $a$ is a prefix of $b$ if

$$aS^1 \supseteq bS^1. \tag{2.1}$$

In the above, $S^1$ is the monoid obtained from $S$ by adding an identity element, unless it was already there. The sets $aS^1, bS^1$ are called *right ideals*. Because of the description in terms of

elements of a semigroup are in the same *prefix class* if they are prefixes of each other. Likewise we define *suffix classes* and *infix classes*.

Clearly every prefix class is contained in some infix class, because prefixes are special cases of infixes. Therefore, every infix class is partitioned into prefix classes. For the same reasons, every infix class is partitioned into suffix classes. The following lemma describes the structure of these partitions.

**Lemma 2.4** (Eggbox lemma)**.** *The following hold in every finite semigroup.*

(1) *all distinct prefix classes in a given infix class are incomparable:*

$$a, b \text{ are infix equivalent, and } a \text{ is a prefix of } b \implies a, b \text{ are prefix equivalent}$$

(2) *if a prefix class and a suffix class are contained in the same infix class, then they have nonempty intersection;*
(3) *all prefix classes in the same infix class have the same size.*

Of course, by symmetry, the lemma remains true after swapping infixes with suffixes.

*Proof*

(1) This item says that distinct prefix classes in the same infix class are incomparable, with respect to the prefix relation. This item of the Eggbox Lemma is the one that will be used most often.

Suppose that $a, b$ are infix equivalent and $a$ is a prefix of $b$, as witnessed by solutions $x, y, z$ to the equations

$$b = ax \qquad a = ybz.$$

As usual, each of $x, y, z$ could be empty. This can be illustrated as

$$a \underset{c \mapsto ycz}{\overset{c \mapsto cx}{\rightleftarrows}} b$$

By the Idempotent Power Lemma, there is some $n \in \{1, 2, \ldots\}$ such that $y^n$ is idempotent. By following the loop around $a$ in the above diagram $n$ times,

inclusion of right ideals, the semigroup literature uses the notation

$$a \geq_{\mathcal{R}} b \overset{\text{def}}{=} aS^1 \supseteq bS^1$$

for the prefix relation. Likewise, $a \geq_{\mathcal{L}} b$ is used for the prefix relation, which is defined in terms of left ideals. Also, for some mysterious reason, $a \geq_{\mathcal{J}} b$ is used for the infix relation. We avoid this notation, because it makes longer words smaller.

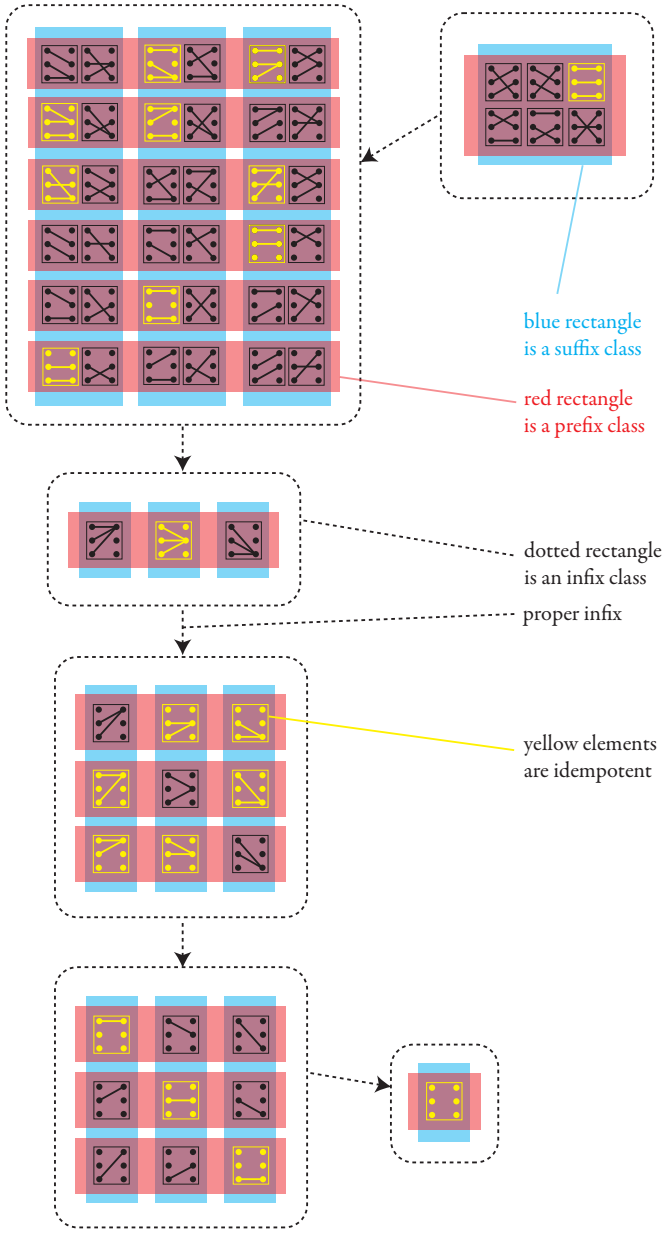



Figure 2.1 The semigroup of partial functions from a three element set to itself, partitioned into prefix, suffix and infix classes. In this particular example, the infix classes are totally ordered, which need not be the case in general.

and then going to $b$, we get

$$\begin{aligned} b = & \quad \text{(follow } 2n \text{ times the loop around } a\text{, then go to } b\text{)} \\ y^{2n}a(xz)^{2n}x = & \quad \text{(}y^n \text{ is an idempotent)} \\ y^{n}a(xz)^{2n}x = & \quad \text{(follow } n \text{ times the loop around } a\text{)} \\ a(xz)^{n}z, & \end{aligned}$$

which establishes that $b$ is a prefix of $a$, and therefore $a, b$ are in the same prefix class.

(2) We now show that prefix and suffix classes in the same infix class must intersect. Suppose that $a, b$ are in the same infix class, as witnessed by

$$a = xby.$$

With respect to the infix relation, $by$ is between $b$ and $a$, and therefore it must be in the same infix class as both of them. We have

$$\overbrace{x\underbrace{b}y}^{by \text{ is a suffix of } xby = a} \qquad \begin{matrix} \\ , \end{matrix}$$
$$b \text{ is a prefix of } by$$

and therefore, thanks to the previous item, $by$ is prefix equivalent to $b$ and suffix equivalent to $a$. This witnesses that the prefix class of $b$ and the suffix class of $a$ have nonempty intersection.

(3) We now show that all prefix classes in the same infix class have the same size. Take some two prefix classes in the same infix class, given by representatives $a, b$. We can assume that $a, b$ are in the same suffix class, thanks to the previous item. Let

$$a = xb \qquad b = ya$$

be witnesses for the fact that $a, b$ are in the same suffix class. The following claim implies that the two prefix classes under consideration have the same size.

**Claim 2.5.** *The following maps are mutually inverse bijections*

$$\textit{prefix class of } a \underset{c \mapsto xc}{\overset{c \mapsto yc}{\rightleftarrows}} \textit{prefix class of } b$$

*Proof* Suppose that $c$ is in the prefix class of $a$, as witnessed by a decomposition $c = az$. If we apply sequentially both maps in the statement of the claim to $c$, then we get

$$xyc = xyaz \overset{ya=b}{=} xbz \overset{xb=a}{=} az \overset{az=c}{=} c.$$

This, and a symmetric argument for the case when $c$ is in the prefix class of $b$, establishes that the maps in the statement of the claim are mutually inverse. It remains to justify that the images of the maps are as in the statement of the claim, i.e. the image of the top map is the prefix class of $b$, and the image of the bottom map is the prefix class of $a$. Because the two maps are mutually inverse, and they prepend elements to their inputs, it follows that each of the maps has its image contained in the infix class of $a, b$. To show that the image of the top map is in the prefix class of $b$ (a symmetric argument works for the bottom map), we observe that every element of this image is of the form $yaz$, and therefore it has $b = ya$ as a prefix, but it is still in the same infix class as $a, b$ as we have observed before, and therefore it must be prefix equivalent to $b$ thanks to the item (1) of the lemma. □

□

The Eggbox Lemma establishes that each infix class has the structure of a rectangular grid (which apparently is reminiscent of a box of eggs), with the rows being prefix classes and the columns being suffix classes. Let us now look at the eggs in the box: define an $\mathcal{H}$-class to be a nonempty intersection of some prefix class and some suffix class. By item (2) of the Eggbox Lemma, every pair of prefix and suffix classes in the same infix class lead to some $\mathcal{H}$-class. The following lemma shows that all $\mathcal{H}$-classes in the same infix class have the same size.

**Lemma 2.6.** *If $a, b$ are in the same infix class, then there exist possibly empty $x, y$ such that the following is a bijection*

$$\mathcal{H}\text{-class of } a \xrightarrow{c \mapsto xcy} \mathcal{H}\text{-class of } b$$

*Proof* Consider first the special case of the lemma, when $a$ and $b$ are in the same suffix class. Take the map from Claim 2.5, which maps bijectively the prefix class of $a$ to the prefix class of $b$. Since this map preserves suffix classes, it maps bijectively the $\mathcal{H}$-class of $a$ to the $\mathcal{H}$-class of $b$. By a symmetric argument, the lemma is also true when $a$ and $b$ are in the same prefix class.

For the general case, we use item (2) of the Eggbox Lemma, which says that there must be some intermediate element that is in the same prefix class as $a$ and in the same suffix class as $b$, and we can apply the previously proved special cases to go from the $\mathcal{H}$-class of $a$ to the $\mathcal{H}$-class of the intermediate element, and then to the $\mathcal{H}$-class of $b$. □

The following lemma shows a dichotomy for an $\mathcal{H}$-class: either it is a group,

or the the product of every two elements in that $\mathcal{H}$-class falls outside the infix class.

**Lemma 2.7** ($\mathcal{H}$-class Lemma)**.** *Furthermore, the following conditions are equivalent for every $\mathcal{H}$-class $G$ in a finite semigroup:*

(1) *$G$ contains an idempotent;*
(2) *$ab$ is in the same infix class as $a$ and $b$, for some $a, b \in G$;*
(3) *$ab \in G$ for some $a, b \in G$;*
(4) *$ab \in G$ for all $a, b \in G$;*
(5) *$G$ is a group (with product inherited from the semigroup)*

*Proof* Implications (5) $\Rightarrow$ (1) $\Rightarrow$ (2) in the lemma are obvious, so we focus on the remaining implications.

(2)$\Rightarrow$(3) Suppose that $ab$ is in the same infix class as $a$ and $b$. Since $a$ is a prefix of $ab$, and the two elements are in the same infix class, item (1) of the Eggbox Lemma implies that $ab$ is in the prefix class of $a$, which is the same as the prefix class of $b$. For similar reasons, $ab$ is in the same suffix class as $a$ and $b$, and therefore $ab \in G$.

(3)$\Rightarrow$(4) Suppose that there exist $a, b \in G$ with $ab \in G$. We need to show that $G$ contains the product of every elements $c, d \in G$. Since $c$ is prefix equivalent to $a$ there is a decomposition $a = xc$, and for similar reasons there is a decomposition $b = dy$. Therefore, $cd$ is an infix of

$$\overbrace{xc}^{a}\overbrace{dy}^{b} \in G,$$

and therefore it is in the same infix class as $G$. Since $c$ is a prefix of $cd$, and both are in the same infix class, the Eggbox Lemma implies that $cd$ is in the prefix class of $c$. For similar reasons $cd$ is in the suffix class of $d$. Therefore, $cd \in G$.

(4)$\Rightarrow$(5) Suppose that $G$ is closed under products, i.e. it is a subsemigroup. We will show that it is a group. By the Idempotent Power Lemma, $G$ contains some idempotent, call it $e$. We claim that $e$ is an identity element in $G$, in particular it is unique. Indeed, let $a \in G$. Because $a$ and $e$ are in the same suffix class, it follows that $a$ can be written as $xe$, and therefore

$$ae = xee = xe = a.$$

For similar reasons, $ea = a$, and therefore $e$ is an identity element in $G$. The group inverse is defined as follows. For $a \in G$, choose some $k \in \{2, 3, \ldots\}$ such that $a^k$ is idempotent, such $k$ exists by Lemma 2.2. Since there is only

one idempotent in $G$, we have $a^k = e$. Therefore, $a^{k-1}$ is a group inverse of $a$.

□

**Exercise 9.** Show that for every finite monoid, the infix class of the monoid identity is a group.

**Exercise 10.** Consider a finite semigroup. Show that an infix class contains an idempotent if and only if it is *regular*, which means that there exist $a, b$ in the infix class such that $ab$ is also in the infix class.

**Exercise 11.** Show that if $G_1, G_2$ are two $\mathcal{H}$-classes in the same infix class of a finite semigroup, and they are both groups, then they are isomorphic as groups[2].

[2] Let us combine Exercises 10 and 11. By Exercises (10) and the $\mathcal{H}$-class lemma, an infix class is regular if and only if it contains an $\mathcal{H}$-class which is a group. By Exercise (11), the corresponding group is unique up to isomorphism. This group is called the *Schützenberger group* of the regular infix class.

# PART TWO

## SOLUTIONS TO THE EXERCISES

**Solution to Exercise 1.**
Let $M$ be any monoid, e.g. the trivial monoid of size one, and let $M'$ be its extension obtained by adding an extra identity element 1. The inclusion embedding

$$M \hookrightarrow M'$$

is a semigroup homomorphism, but not a monoid homomorphism.

**Solution to Exercise 2.**
Take any finite set $X$ and any function

$$f : X^2 \to \{0, \text{true}\}.$$

The number of choices for $f$ is exponential in $X^2$. We can extend $f$ to an associative product operation on the set

$$S = X + \{0, \text{true}\}$$

by defining all products outside $X^2$ to give value 0. Semigroups obtained this way correspond to languages where all words have length at most 2.

**Solution to Exercise 3.**

**Solution to Exercise 4.**
For $n \in \{1, 2, \ldots\}$, consider the language of words in $\{a, b\}^*$ where the $n$-th letter is $a$. This language is recognised by a deterministic finite automaton with $O(n)$ states. To recognise the language with a monoid homomorphism, we need to remember the first $n$ letters of a word.

**Solution to Exercise 5.**
An example is words of length $2n + 1$, over alphabet $\{a, b\}$, where the middle letter is $a$. For the same reasons as in Exercise 4, this language needs an exponential size monoid to be recognised. On the other hand, the language is clearly recognised by a deterministic automaton (running in either direction), with $O(n)$ states.

**Solution to Exercise 6.**
No surprises here – these are the commutative languages, i.e. languages $L \subseteq \Sigma^*$ which are commutative in the sense that

$$wabv \in L \quad \text{iff} \quad wbav \in L \qquad \text{for all } w, a, b, v \in \Sigma^*.$$

There is, however, an extended description of commutative languages, which is given below.

Consider a language $L \subseteq \Sigma^*$ recognised by a monoid homomorphism

$$h : \Sigma^* \to M,$$

where $M$ is a commutative monoid. By commutativity of $M$, we have

$$h(w) = \prod_{a \in \Sigma} h(a)^{\#_a(w)} \qquad \text{for every } w \in \Sigma^*,$$

where $\#_a(w) \in \{0, 1, \ldots\}$ is the number of appearances of letter $a \in \Sigma$ in $w$. For every $a \in M$ the sequence

$$a^0, a^1, a^2, \ldots \in M$$

is easily seen to be ultimately periodic, which means that after cutting of a finite prefix of the sequence we get a sequence that is periodic. This in turn implies that for every $a, b \in M$, the set

$$\{i \in \{0, 1, \ldots\} : a^i = b\}$$

is defined by a formula $\varphi(i)$ which is a finite Boolean combination of formulas which have one of the following forms:

(1) $i = k$ for some $k \in \{0, 1, \ldots\}$; or
(2) $i \equiv k \mod m$ for some $k \in \{0, 1, \ldots\}$.

Putting these observations together, we see that for every $b \in M$, the set

$$\{w \in \Sigma^* : b = \prod_{a \in \Sigma} h(a)^{\#_a(w)}\},$$

which is equal to the inverse image $h^{-1}(b)$, is defined by a finite Boolean combination of formulas of one of the following forms:

(i) $\#_a(w) = k$ for some $a \in \Sigma$ and $k \in \{0, 1, \ldots\}$; or
(ii) $\#_a(w) \equiv k \mod m$ for some $a \in \Sigma$ and $k \in \{0, 1, \ldots\}$.

It follows that the following conditions are equivalent for every language:

- recognised by a finite commutative monoid;
- regular and commutative as a language;
- defined by a finite Boolean combination of conditions as in (i) and (ii).

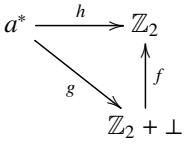
**Solution to Exercise 7.**
Consider the language $(aa)^*$ of even length words, which is recognised by the homomorphism

$$h : a^* \to \mathbb{Z}_2 = (\{0, 1\}, +_{\text{mod}2}).$$

Define $\mathbb{Z}_2 + \bot$ to be the extension of $\mathbb{Z}_2$ with an absorbing element $\bot$. Define

$$g : a^* \to \mathbb{Z}_2 + \bot$$

to be the same function as $h$, except that the co-domain is bigger. In particular, $g$ is not surjective. We claim that there is no monoid homomorphism $f$ which makes the following diagram commute




Since $\bot$ is absorbing in $\mathbb{Z}_2 + \bot$, then the image of $f(\bot)$ must be an absorbing in $\mathbb{Z}_2$, and there are no absorbing elements in $\mathbb{Z}_2$. It follows that there is no $f$ which makes the diagram commute.

**Solution to Exercise 8.**
The syntactic congruence identifies two words $w_1$ and $w_2$ if

$$uw_1v \in L \Leftrightarrow uw_2v \in L \qquad \text{for all } u, v \in \Sigma^*.$$

To prove the exercise, we will show that two words are equivalent in the above sense if and only if they induce the same state transformations in the syntactic automaton. Clearly if the words have the same state transformations, then they are equivalent. We are left with proving the opposite, i.e. different state transformations imply non-equivalence under the syntactic congruence.

Suppose that $w_1$ and $w_2$ have different state transformations. This means that there is some state $q$ of the syntactic automaton such that

$$qw_1 \neq qw_2.$$

Like any state of the syntactic automaton, $q$ is reached from the initial state by reading some word $u$. Since the states $qw_1$ and $qw_2$ are different, there must be some word $u$ such that exactly one of the states

$$(qw_1)u \qquad (qw_2)u$$

is accepting. Summing up, we have found two words $u, v$ such that exactly one of the words

$$uw_1v \qquad uw_2v$$

is in the language, thus proving that $w_1$ and $w_2$ are not equivalent under the syntactic congruence.

**Solution to Exercise 9.**
Let $J$ be the infix class of the monoid identity 1. Since 1 is prefix of every monoid element, it follows from the Eggbox Lemma that $J$ is equal to the prefix class of 1. For the same reasons, $J$ is equal to the suffix class of 1. Therefore $J$ is a single $\mathcal{H}$-class. Since $J$ contains an idempotent, namely 1, it must be a group by the $\mathcal{H}$-class Lemma.

**Solution to Exercise 10.**
If an infix class contains an idempotent, then it clearly contains elements $a, b$ such that $ab$ is in the infix class. For the converse implication, suppose that $a, b$ and $ab$ are in the same infix class $J$. It follows that each of $a, b$ can be decomposed as products of two elements from $J$. By iterating this procedure, we see that $a$ can be decomposed as product of $n$ elements from $J$, for every $n \in \{1, 2, \ldots\}$. Thanks Exercise 3, a product of $n$ elements from $J$ must contain an idempotent infix.

**Solution to Exercise 11.**
Suppose that $G_1$ and $G_2$ are groups in the same infix class. Let $e_1, e_2$ be the identities in the groups $G_1, G_2$. From Claim 2.5 it follows that there exist $x_1, x_2, y_1, y_2$ such that

$$G_1 \underset{g \mapsto x_2 g y_2}{\overset{g \mapsto x_1 g y_1}{\rightleftarrows}} G_2 \tag{2.2}$$

are mutually inverse bijections. By replacing

$$\underbrace{x_1 e_1}_{\text{new } x_1} \quad \underbrace{e_1 x_2}_{\text{new } x_2} \quad \underbrace{e_1 y_1}_{\text{new } y_1} \quad \underbrace{y_2 e_1}_{\text{new } y_2},$$

we still get mutually inverse bijections between $G_1$ and $G_2$. Summing up, we can assume without loss of generality that $x_1, y_2$ end with $e_1$, while $x_2, y_1$ begin with $e_1$.

The element $y_1 x_1$ begins and ends with $e_1$, and it is also an infix of $e_2$ (and therefore also of $e_1$) thanks to

$$e_2 = e_2 e_2 = \underbrace{x_1 x_2 e_2 y_2 y_1}_{e_2} \underbrace{x_1 x_2 e_2 y_2 y_1}_{e_2}.$$

It follows from the Eggbox Lemma that $y_1 x_1$ is both in the prefix class and suffix class of $e_1$, which means that $y_1 x_1 \in G_1$. Since $G_1$ is a group, there must be some $a, b \in G_1$ such that

$$a y_1 x_1 a = e_1$$

Let $\overline{y_1x_1}$ be the group inverse of $y_1x_1$, in the group $G_1$. Define $\alpha : G_1 \to G_2$ to be the composition of the following two functions

$$G_1 \xrightarrow{g \mapsto g\overline{y_1x_1}} G_1 \xrightarrow{g \mapsto x_1gy_1} G_2.$$

The first function is a permutation of $G_1$, while the second function is a bijection of $G_1$ and $G_2$. It follows that $\alpha$ is a bijection. We now claim that $\alpha$ is a homomorphism:

$$\alpha(gh) = x_1gh\overline{y_1x_1}y_1 = \overbrace{x_1g\underbrace{\overline{y_1x_1}y_1}_{e_1}}^{\alpha(g)}\overbrace{x_1h\overline{y_1x_1}y_1}^{\alpha(h)} = \alpha(g)\alpha(h).$$

Summing up, $\alpha$ is a bijective semigroup homomorphism between the groups $G_1$ and $G_2$. It follows that these groups are isomorphic as groups, because a bijective semigroup homomorphism also preserves the group structure.

# Bibliography

# Author index

# Subject index