Automata and Formal Languages

Stefan Hetzl
stefan.hetzl@tuwien.ac.at

Vienna University of Technology

Summer Term 2023

June 26, 2023

Contents

Preface v

1 Semirings and formal series 1

1.1 Complete partial orders 2

1.1.1 Suprema 2

1.1.2 Fixed points 2

1.2 Continuous semirings 3

1.2.1 Semirings 3

1.2.2 The natural order 4

1.2.3 Continuity 5

1.3 Algebraic systems 7

1.3.1 Polynomials 7

1.3.2 R'-algebraic systems 8

1.3.3 Solvability 9

1.3.4 Context-free grammars 10

1.4 Formal series 11

1.4.1 The continuous semiring $R\langle\langle A^*\rangle\rangle$ 11

1.4.2 $\mathcal{P}(A^*)$ and $\mathbb{B}\langle\langle A^*\rangle\rangle$ 15

1.4.3 Calculations in semirings of formal series 15

1.4.4 Context-free grammars and $\mathbb{N}^\infty$ 18

1.5 Matrices 19

1.5.1 The continuous semiring $R^{I\times I}$ 19

1.5.2 The star of a matrix 20

1.6 Regular languages 24

1.6.1 Finite automata 24

1.6.2 Kleene's theorem 27

1.6.3 Linear systems 29

2 Algebraic automata theory 31

2.1 The Myhill-Nerode theorem 31

2.1.1 The right-congruence of a DFA . . . 31
2.1.2 The right-congruence of a language . . . 33
2.1.3 The minimal DFA of a regular language . . . 35
2.2 Transition monoids . . . 37
2.2.1 Quotient monoids . . . 37
2.2.2 The transition monoid of a DFA . . . 38
2.2.3 Languages recognised by a monoid . . . 40
2.2.4 The syntactic monoid . . . 41
2.2.5 The monoids recognising a fixed language . . . 42
2.3 Star height . . . 43
2.3.1 Star-free languages . . . 43
2.3.2 Local divisors . . . 45
2.3.3 Schützenberger's theorem . . . 46
2.4 The variety theorem . . . 49

3 Further topics **53**
3.1 Automatic sequences . . . 53
3.1.1 DFAs with output . . . 53
3.1.2 k-automatic sequences . . . 54
3.1.3 Morphic sequences . . . 56

Preface

Automata theory is one of the most central subjects of theoretical computer science. Finite automata are the simplest possible machines and they appear explicitely as well as implicitely in a variety of different subjects and applications in both computer science and mathematics.

In mathematics, automata and formal languages are firmly tied to monoids and semirings. A large part of these course notes is devoted to developing the theory of automata and formal languages on this algebraic basis. In the other direction, automata theory finds many mathematical applications, e.g., in number theory and combinatorics (automatic sequences), in algebra (automatic groups), and logic (automatic structures). In the mentioned applications, automata are a tool for studying structures which are non-trivial but, at the same, sufficiently simple to have strong properties.

One of the historically first applications of formal language theory in computer science was for the improvement of interpreters and compilers. The theory allows to put the definition of a programming language on a firm theoretical basis and it has long become state of the art to automatically generate a parser for a programming language from a grammar of this language. Thus the reliability of crucial components of the infrastructure of computer science could be greatly increased. Now, there are many additional applications, too numerous to mention here, among them also quite recent ones, e.g., connected to the XML document format (XSLT and tree transducers, grammar-based XML compression, etc.).

Chapter 1

Semirings and formal series

In this chapter we will see a strong generalisation of the elementary theory of formal languages. It will turn out that many of the central results remain valid in the more general setting where the formal languages $\mathcal{P}(A^*)$ over some alphabet A with the operations $\cup$ and $\cdot$ is replaced by the formal series with coefficients in an arbitrary *continuous semiring*. Moreover, once the algebraic background is sufficiently developed, proofs known from the elementary theory of formal languages can – in most cases – be carried over to this more general setting in a straightforward way. We will thus obtain considerably more general results which do not only include the elementary theory of formal languages as special case, but also other interesting and useful types of languages and automata, e.g. weighted automata where costs are assigned to transitions. Moreover, this more general theory has been exploited for obtaining new results about the notions of the elementary theory.

As a starting point, let us consider the notion of context-free grammar. The language $L = \{a^n b^n \mid n \geqslant 0\}$ can be generated by the context-free grammar given by $S \to aSb \mid \varepsilon$. The usual perspective on a grammar is to consider the given production rules from a generative point of view: we start with S and, at each step, make a choice which production rule to apply until eventually S does no longer occur and we have thus generated a word in L. In our algebraic setting we will instead consider a nonterminal of a grammar as a variable for a language, i.e., a set of words, and the production rules as equations that have to be satisfied by the variables. This example is then rephrased as $S = aSb \cup \{\varepsilon\}$. We can observe that L is a set of words which satisfies the equation $L = aLb \cup \{\varepsilon\}$. While working with finite strings is crucial for the first, generative perspective, it is irrelevant for the second based on equations. We can thus lift the requirement that the object we insert for a variable must be a language, i.e., a set of words and consider the insertion of an element of an arbitrary continuous semiring.

In order to fully develop this point of view, we start with some basic results about complete partial orders and then proceed to discuss the most central notion of this first chapter: that of a continuous semiring. On this basis we then consider systems of algebraic equations which generalise grammars in the manner sketched above.

1.1 Complete partial orders

1.1.1 Suprema

Definition 1.1. $(S, \leqslant)$ is a *partial order* if $(S, \leqslant)$ is reflexive, transitive, and anti-symmetric[1].

Definition 1.2. Let $(S, \leqslant)$ be a partial order, $b \in S$ and $X \subseteq S$. Then b is called *upper bound* of X if $b \geqslant x$ for all $x \in X$. Furthermore, b is called *supremum* (least upper bound) of X if b is an upper bound of X and if, for every upper bound c of X: $b \leqslant c$.

Note that a supremum is unique if it exists (for suppose a set X had two suprema b_1 and b_2, then, since both are upper bounds, $b_1 \leqslant b_2$ and $b_2 \leqslant b_1$ and hence by the anti-symmetry of $\leqslant$ we have $b_1 = b_2$). We often write $\sup X$ for the supremum of a set X. Using the notation $\sup X$ we implicitly assert the *existence* of the supremum (its uniqueness then follows as above).

Definition 1.3. A partial order $(S, \leqslant)$ is called *complete* if every increasing sequence $s_0 \leqslant s_1 \leqslant \cdots$ has a supremum in S.

Note that an increasing sequence is a countable set with order type ω. Therefore the above notion is also called ω-complete in the literature. Since we will only be considering this type of complete partial orders in this course, we use the simpler terminology.

Example 1.1. $(\mathcal{P}(X), \subseteq)$ is complete with $\sup\{X_i \mid i \in \mathbb{N}\} = \bigcup_{i \in \mathbb{N}} X_i$. However, for infinite X, the partial order $(\mathcal{P}_{\text{fin}}(X), \subseteq)$ of finite subsets of X is not complete.

Definition 1.4. Let $(S, \leqslant)$ be a complete partial order and $f : S^n \to S$. Then f is called *continuous* if for all $k \in \{1, \ldots, n\}$, for all $a_1, \ldots, a_{k-1}, a_{k+1}, \ldots, a_n \in S$, and for all increasing sequences $b_0 \leqslant b_1 \leqslant \cdots$ we have:

$$f(a_1, \ldots, a_{k-1}, \sup\{b_i \mid i \in \mathbb{N}\}, a_{k+1}, \ldots, a_n) = \sup\{f(a_1, \ldots, a_{k-1}, b_i, a_{k+1}, \ldots, a_n) \mid i \in \mathbb{N}\}$$

Lemma 1.1. *Let $(S, \leqslant)$ be a complete partial order and $f : S^n \to S$ continuous. For all $i \in \{1, \ldots, n\}$ let $x_{i,0} \leqslant x_{i,1} \leqslant \cdots$ be an increasing sequence. Then*

$$f(\sup\{x_{1,j_1} \mid j_1 \in \mathbb{N}\}, \ldots, \sup\{x_{n,j_n} \mid j_n \in \mathbb{N}\}) = \sup\{f(x_{1,j}, \ldots, x_{n,j}) \mid j \in \mathbb{N}\}.$$

Proof. will be done as exercise. □

1.1.2 Fixed points

Definition 1.5. Let $(S, \leqslant)$ be a partial order and $f : S \to S$. Then $x \in S$ is called *fixed point of f* if $f(x) = x$ and x is called *least fixed point of f* if $x \leqslant y$ for every fixed point y of f.

If a least fixed point exists it is unique (since $\leqslant$ is a partial order). In analogy to the notational convention concerning the supremum we write $\text{lfp}(f)$ for the least fixed point of f and by using this notation implicitly assert the existence of a (and hence the) least fixed point of f.

Example 1.2. Consider the partial order $(\mathcal{P}(\mathbb{N}), \subseteq)$. Define

$$f : \mathcal{P}(\mathbb{N}) \to \mathcal{P}(\mathbb{N}), X \mapsto \{0\} \cup X \cup (X + 2)$$

where $X + k := \{x + k \mid x \in X\}$. Then $2\mathbb{N}$ is a fixed point of f because $\{0\} \cup 2\mathbb{N} \cup (2\mathbb{N} + 2) = 2\mathbb{N}$. Furthermore, also $\mathbb{N}$ is a fixed point of f and so is $2\mathbb{N} \cup (2\mathbb{N} + 2k + 1)$ for all $k \in \mathbb{N}$.

[1] $(S, \leqslant)$ is anti-symmetric if $x \leqslant y$ and $y \leqslant x$ implies $x = y$.

Definition 1.6. Let $(S, \leqslant)$ be a partial order. Then $f : S^n \to S$ is called *monotone* if for all $x_1, \dots, x_n \in S$, for all $k \in \{1, \dots, n\}$ and for all $x'_k \in S$ with $x'_k \geqslant x_k$ we have

$$f(x_1, \dots, x_n) \leqslant f(x_1, \dots, x_{k-1}, x'_k, x_{k+1}, \dots x_n).$$

Lemma 1.2. *Let $(S, \leqslant)$ be a complete partial order and $f : S^n \to S$ continuous. Then f is monotone.*

Proof. Let $x_1, \dots, x_n \in S$, $k \in \{1, \dots, n\}$ and $x_k \leqslant x'_k \in S$. Define $y_0 = x_k$ and $y_i = x'_k$ for $i \geqslant 1$. Then $y_0 \leqslant y_1 \leqslant \cdots$ is an increasing sequence and we have $\sup\{y_i \mid i \in \mathbb{N}\} = x'_k$. Then

$$\begin{aligned} f(x_1, \dots, x_{k-1}, x'_k, x_{k+1}, \dots, x_n) &= f(x_1, \dots, x_{k-1}, \sup\{y_i \mid i \in \mathbb{N}\}, x_{k+1}, \dots, x_n) \\ &= \sup\{f(x_1, \dots, x_{k-1}, y_i, x_{k+1}, \dots, x_n) \mid i \in \mathbb{N}\} \\ &\geqslant f(x_1, \dots, x_n). \end{aligned}$$

□

In a partial order $(S, \leqslant)$ we say that $x \in S$ is the least element of $(S, \leqslant)$ if $x \leqslant y$ for all $y \in S$. If a least element exists, it is unique (again, by anti-symmetry).

Theorem 1.1 (Kleene's fixed point theorem[2])**.** *Let $(S, \leqslant)$ be a complete partial order with least element $0 \in S$ and let $f : S \to S$ be continuous. Then $\mathrm{lfp}(f) = \sup\{f^i(0) \mid i \in \mathbb{N}\}$.*

Proof. We will consider the sequence $0 = f^0(0), f^1(0), f^2(0), \dots$. Since 0 is the least element we have $0 \leqslant f(0)$ and since f is monotone, $f^i(0) \leqslant f^{i+1}(0)$ implies $f^{i+1}(0) \leqslant f^{i+2}(0)$ and hence by induction $f^0(0) \leqslant f^1(0) \leqslant \cdots$ is an increasing sequence. Therefore $\sup\{f^i(0) \mid i \in \mathbb{N}\}$ exists. Furthermore,

$$f(\sup\{f^i(0) \mid i \in \mathbb{N}\}) = \sup\{f^{i+1}(0) \mid i \in \mathbb{N}\} = \sup\{f^i(0) \mid i \in \mathbb{N}\}$$

and hence $\sup\{f^i(0) \mid i \in \mathbb{N}\}$ is fixed point of f. It remains to show that it is the least fixed point of f. To that aim, let c be a fixed point of f. Since 0 is least element we have $0 \leqslant c$ and, if $f^i(0) \leqslant c$, then $f^{i+1}(0) = f(f^i(0)) \leqslant f(c) = c$ and hence by induction $f^i(0) \leqslant c$ for all $i \in \mathbb{N}$. So c is upper bound of $\{f^i(0) \mid i \in \mathbb{N}\}$ and hence $\sup\{f^i(0) \mid i \in \mathbb{N}\} \leqslant c$. □

Example 1.3. The complete partial order $(\mathcal{P}(\mathbb{N}), \subseteq)$ has the least element $\varnothing$. The function $f : \mathcal{P}(\mathbb{N}) \to \mathcal{P}(\mathbb{N})$ defined in Example 1.2 is continuous[3]. By the fixed point theorem, f has a least fixed point which we can approximate by the sequence

$$f^0(\varnothing) = \varnothing, f^1(\varnothing) = \{0\}, f^2(\varnothing) = \{0, 2\}, f^3(\varnothing) = \{0, 2, 4\}, \dots$$

A straightforward induction shows that $\mathrm{lfp}(f) = \bigcup_{i \in \mathbb{N}} f^i(\varnothing) = 2\mathbb{N}$.

1.2 Continuous semirings

1.2.1 Semirings

Definition 1.7. A *semiring* is a structure $\langle R, +, 0, \cdot, 1 \rangle$ s.t.

[2] named after Stephen Cole Kleene (1909-1994)

[3] Show this as exercise. More precisely, show $f(\bigcup_{i \in \mathbb{N}} X_i) = \bigcup_{i \in \mathbb{N}} f(X_i)$ by proving set-inclusion in both directions.

1. $\langle R, +, 0\rangle$ is a commutative monoid,
2. $\langle R, \cdot, 1\rangle$ is a monoid,
3. for all $x, y, z \in R$: $x \cdot (y + z) = x \cdot y + x \cdot z$ and $(x + y) \cdot z = x \cdot z + y \cdot z$, and
4. for all $x \in R$: $x \cdot 0 = 0 \cdot x = 0$.
5. $0 \neq 1$

Note that conditions 1-4 above together with $0 = 1$ imply $\forall x\, x = 0$ because $x = x \cdot 1 = x \cdot 0 = 0$. We therefore add condition 5 to avoid this trivial situation.

Example 1.4.

1. Every ring and hence also every field is a semiring, in particular $\langle \mathbb{Z}, +, 0, \cdot, 1\rangle$, $\langle \mathbb{Q}, +, 0, \cdot, 1\rangle$, $\langle \mathbb{R}, +, 0, \cdot, 1\rangle$, ...
2. $\langle \mathbb{N}, +, 0, \cdot, 1\rangle$ is a semiring.
3. The *Boolean semiring* is $\mathbb{B} = \langle \{0, 1\}, \vee, 0, \wedge, 1\rangle$ with logical (inclusive) disjunction as sum and logical conjunction as product.
4. Let $\mathbb{N}^\infty = \mathbb{N} \cup \{\infty\}$, then both $\langle \mathbb{N}^\infty, +, 0, \cdot, 1\rangle$ as well as $\langle \mathbb{N}^\infty, \min, \infty, +, 0\rangle$ are semirings (note that $0 \cdot \infty = \infty \cdot 0 = 0$). We use $\mathbb{N}^\infty$ to denote the first of these semirings. When we want to speak about the second we either mention the operations explicitly or call it the min-+-*semiring.*
5. Let $\bar{\mathbb{N}} = \mathbb{N} \cup \{\infty, -\infty\}$, then $\langle \bar{\mathbb{N}}, \max, -\infty, +, 0\rangle$ with $-\infty + \infty = \infty + -\infty = -\infty$ is a semiring. This semiring is also called the max-+-*semiring.*
6. Let $\mathbb{R}_+^\infty = \{x \in \mathbb{R} \mid x \geqslant 0\} \cup \{\infty\}$, then $\langle \mathbb{R}_+^\infty, +, 0, \cdot, 1\rangle$ is a semiring.
7. An alphabet is a finite set of symbols. The formal languages over an alphabet A form the semiring $\langle \mathcal{P}(A^*), \cup, \emptyset, \cdot, \{\varepsilon\}\rangle$. As a reminder, the concatenation of formal languages is defined as $L_1 \cdot L_2 = \{w_1 w_2 \mid w_1 \in L_1, w_2 \in L_2\}$.

1.2.2 The natural order

Definition 1.8. A semiring R is called *naturally ordered* if the relation $\sqsubseteq$ defined as

$$x \sqsubseteq y \iff \exists z\, x + z = y$$

is a partial order. In that case we call $\sqsubseteq$ the *natural order* of A.

The relation $\sqsubseteq$ is reflexive and transitive in every semiring. Hence it is a partial order iff it is anti-symmetric.

Example 1.5. The semiring $\langle \mathbb{N}, +, 0, \cdot, 1\rangle$ is naturally ordered since $\sqsubseteq$ is just the usual order $\leqslant$. The semiring $\langle \mathbb{Z}, +, 0, \cdot, 1\rangle$ is not naturally ordered since $x \sqsubseteq y$ is true for all $x, y \in \mathbb{Z}$ (but there is more than one integer). The semiring $\langle \mathcal{P}(A^*), \cup, \emptyset, \cdot, \{\varepsilon\}\rangle$ is naturally ordered with $\sqsubseteq = \subseteq$.

Proposition 1.1. *Let R be a naturally ordered semiring. Then 0 is the least element of $(R, \sqsubseteq)$ and R is zerosumfree, i.e., $x + y = 0$ implies $x = y = 0$.*

Proof. $0 + x = x$ for all $x \in R$ and hence 0 is the least element of R. Now, if $x + y = 0$, then $x \sqsubseteq 0$ and $y \sqsubseteq 0$ and since also $0 \sqsubseteq x$ and $0 \sqsubseteq y$ we have $x = y = 0$. □

This shows that, in particular, no ring and hence no field is a naturally ordered semiring (since rings are not zerosumfree).

1.2.3 Continuity

Definition 1.9. A naturally ordered semiring $\langle R, +, 0, \cdot, 1\rangle$ is called *continuous* if

1. $\sqsubseteq$ is a complete partial order, and
2. $+$ and $\cdot$ are continuous functions w.r.t. $(R, \sqsubseteq)$.

A remark analogous to that after the definition of complete partial order: also in the above definition of continuity we speak about countable sets with order type ω, hence the above notions are also often called ω-continuity and ω-continuous semirings respectively. Again, since we will work only with this type of continuous semirings in this course, we use the simpler terminology.

Example 1.6. The semiring $\langle \mathcal{P}(A^*), \cup, \varnothing, \cdot, \{\varepsilon\}\rangle$ is continuous. We already know that $\mathcal{P}(A^*)$ is a naturally ordered semiring, that the natural order is $\subseteq$, that this natural order is complete and that $\sup\{x_i \mid i \in \mathbb{N}\} = \bigcup_{i\in\mathbb{N}} x_i$. It remains to show that $\mathcal{P}(A^*)$ is continuous. To that aim note that

$$x \cup \sup\{y_i \mid i \in \mathbb{N}\} = x \cup \bigcup_{i\in\mathbb{N}} y_i = \bigcup_{i\in\mathbb{N}} (x \cup y_i) = \sup\{x \cup y_i \mid i \in \mathbb{N}\}.$$

Due to the commutativity of addition in a semiring, this suffices to show that addition is continuous. For multiplication, consider

$$x \cdot \sup\{y_i \mid i \in \mathbb{N}\} = x \cdot \bigcup_{i\in\mathbb{N}} y_i =^{(*)} \bigcup_{i\in\mathbb{N}} (x \cdot y_i) = \sup\{x \cdot y_i \mid i \in \mathbb{N}\},$$

where the equation $(*)$ can be proved carefully by proving set-inclusions in both directions. For multiplication from the right a calculation symmetric to the above can be carried out.

Continuous semirings are the central notion of this first chapter. Their importance stems from the fact that they allow the definition of (well-behaved) infinite sums and consequently a generalisation of the Kleene-star, one of the most important operations on formal languages.

If $\{x_i \mid i \in \mathbb{N}\}$ is an arbitrary subset (and hence not necessarily an increasing sequence) we still have the property that the partial sums $x_0 + \cdots + x_n = \sum_{i=0}^{n} x_i$ of $\sum_{i\in\mathbb{N}} x_i$ form an increasing sequence since $\sum_{i=0}^{n} x_i \sqsubseteq \sum_{i=0}^{n+1} x_i$ and therefore the supremum of the partial sums exists. Consequently:

Definition 1.10. Let $\langle R, +, 0, \cdot, 1\rangle$ be a continuous semiring and $\{x_i \mid i \in \mathbb{N}\} \subseteq R$. Then we define

$$\sum_{i\in\mathbb{N}} x_i := \sup\{\sum_{i=0}^{n} x_i \mid n \in \mathbb{N}\}$$

As a first observation about infinite sums in this context, one can show that the order of summation does not matter. More precisely: let $\varphi : \mathbb{N} \to \mathbb{N}$ be a bijection, then $\sum_{i\in\mathbb{N}} x_i = \sum_{i\in\mathbb{N}} x_{\varphi(i)}$.

We extend the definition of an infinite sum to an arbitrary countably infinite index set I by $\sum_{i\in I} x_i := \sum_{n\in\mathbb{N}} x_{\varphi(n)}$ for $\varphi : \mathbb{N} \to I$ being an arbitrary bijection. Infinite sums in continuous semirings also have the following properties:

Proposition 1.2. *Let* $\mathbb{N} = \biguplus_{j\in J} I_j$. *Then* $\sum_{i\in\mathbb{N}} x_i = \sum_{j\in J} \sum_{i\in I_j} x_i$.

Without Proof. □

Proposition 1.3. *Let R be a continuous semiring, let $x \in R$ and for all $i \in \mathbb{N}$ let $y_i \in R$. Then*

$$x \cdot \big(\sum_{i\in\mathbb{N}} y_i\big) = \sum_{i\in\mathbb{N}}(x \cdot y_i) \qquad \textit{and} \qquad \big(\sum_{i\in\mathbb{N}} y_i\big) \cdot x = \sum_{i\in\mathbb{N}}(y_i \cdot x).$$

Proof. We have

$$\begin{aligned} x \cdot \big(\sum_{i\in\mathbb{N}} y_i\big) &= x \cdot \sup\{\sum_{i=0}^{n} y_i \mid n \in \mathbb{N}\} \qquad \text{by definition,} \\ &= \sup\{x \cdot \sum_{i=0}^{n} y_i \mid n \in \mathbb{N}\} \qquad \text{by continuity of } \cdot, \\ &= \sup\{\sum_{i=0}^{n}(x \cdot y_i) \mid n \in \mathbb{N}\} \qquad \text{by distributivity, and} \\ &= \sum_{i\in\mathbb{N}}(x \cdot y_i) \qquad \text{by definition.} \end{aligned}$$

For multiplication from the right an analogous calculation can be done. □

Having defined infinite sums now allows to define the generalisation of the Kleene-star.

Definition 1.11. Let R be a continuous semiring and let $x \in R$. Then we define

$$x^* = \sum_{i\geqslant 0} x^i \qquad \text{and} \qquad x^+ = \sum_{i\geqslant 1} x^i$$

where $x^i = \underbrace{x \cdot \dots \cdot x}_{i \text{ times}}$ for $i \geqslant 1$ and $x^0 = 1$.

Example 1.7. Consider the continuous semiring $\langle \mathcal{P}(A^*), \cup, \varnothing, \cdot, \{\varepsilon\}\rangle$ and let $L \in \mathcal{P}(A^*)$. Then $L^* = \bigcup_{i\geqslant 0} L^i$ is the ordinary Kleene-star.

Proposition 1.4. *Let R be a continuous semiring, then we have*

$$x^+ = xx^* = x^*x \qquad \textit{and} \qquad x^* = 1 + x^+$$

for all $x \in R$.

Proof.

$$xx^* = x\sum_{i\geqslant 0} x^i = \sum_{i\geqslant 0} x^{i+1} = \sum_{i\geqslant 1} x^i = x^+$$

and analogously for $x^*x = x^+$. Moreover,

$$x^* = \sum_{i\geqslant 0} x^i = x^0 + \sum_{i\geqslant 1} x^i = 1 + x^+.$$

□

1.3 Algebraic systems

1.3.1 Polynomials

We know univariate polynomials, i.e., expressions of the form

$$P(x) = \sum_{i=0}^{n} a_i x^i$$

and multivariate polynomials, i.e., expressions of the form

$$P(x_1, \ldots, x_n) = \sum_{(i_1,\ldots,i_n)\in E} a_{i_1,\ldots,i_n} x_1^{i_1} \cdots x_n^{i_n} \quad \text{for finite } E \subseteq \mathbb{N}^n.$$

Several standard results about polynomials require multiplication to be commutative (for example the evaluation of a polynomial is not a homomorphism otherwise). Using the above notation already hints at the assumption of commutativity since we order the variables so that, in each product term, every x_i appears in front of all x_j with $i < j$. In this course we do not assume commutativity and hence we will work with more general expressions. Consequently we define:

Definition 1.12. Let R be a continuous semiring, $Y = \{y_1, \ldots, y_n\}$ be a set of variables and let $R' \subseteq R$. An *R'-product term with variables in Y* is an expression of the form

$$a_0 y_{i_1} a_1 y_{i_2} a_2 \cdots y_{i_k} a_k \quad \text{with } a_0, \ldots, a_k \in R' \text{ and } i_1, \ldots, i_k \in \{1, \ldots, n\}.$$

An *R'-polynomial with variables in Y* is an expression of the form $\sum_{i=1}^{m} t_i$ where $t_1, \ldots, t_m$ are R'-product terms with variables in Y. The set of R'-polynomials with variables in Y is denoted as $R'[Y]$.

Note that – depending on R' – the set $R'[Y]$ may not form a semiring, it may not even be closed under addition. In fact, we will often consider R' which are not closed under addition, see, e.g., Example 1.9.

Example 1.8. Let $R = \mathbb{N}^\infty$ and $Y = \{y_1, y_2\}$, then $3y_1y_2^2 + y_1 + 5$ is a $\{1, 3, 5\}$-polynomial with variables in Y. Note that – formally – this polynomial is written as $3y_11y_21y_21 + 1y_11 + 5$ but we will use the simplified notation where factors 1 are left out.

Example 1.9. Let $R = \mathcal{P}(A^*)$ and $Y = \{y_1, y_2\}$, then $\{a\}y_1\{\varepsilon\}y_1\{\varepsilon\}y_2\{\varepsilon\} \cup \{a\}y_2\{b\} \cup \{\varepsilon\}$ is a $\{\{w\} \mid w \in A^*\}$-polynomial with variables in Y. Note that A^* is not a subset of $\mathcal{P}(A^*)$. But since $\{\{w\} \mid w \in A^*\}$-polynomials appear frequently we often write them – in abuse of notation – in the form $ay_1^2y_2 \cup ay_2b \cup \varepsilon$.

Let R, R', and $Y = \{y_1, \ldots, y_n\}$ be as in the above definition and let $p \in R'[Y]$. Then, as usual, the *polynomial* p induces a *polynomial function* $\overline{p} : R^n \to R$ by replacing y_i by $r_i \in R$. As usual, we will not distinguish notationally between a polynomial and the polynomial function it induces.

Example 1.10. Let $A = \{a, b\}$, $R = \mathcal{P}(A^*)$, $R' = \{\{w\} \mid w \in A^*\}$, $Y = \{y_1, y_2\}$ and $p(y_1, y_2) = ay_1^2y_2 \cup ay_2b \cup \varepsilon$. Let $L_1 = \{a^i \mid i \in \mathbb{N}\}$ and $L_2 = \{a^ib^i \mid i \in \mathbb{N}\}$. Then

$$p(L_1, L_2) = aL_1^2L_2 \cup aL_2b \cup \varepsilon = \{a^ib^j \mid i > j \geqslant 0\} \cup \{a^ib^i \mid i \geqslant 1\} \cup \varepsilon = \{a^ib^j \mid i \geqslant j\}.$$

1.3.2 R'-algebraic systems

Definition 1.13. Let R be a continuous semiring and $R' \subseteq R$. An R'-algebraic system in the variables $Y = \{y_1, \ldots, y_n\}$ is a system of equations of the form

$$y_i = p_i(y_1, \ldots, y_n), \quad 1 \leqslant i \leqslant n$$

with $p_i(Y) \in R'[Y]$.

Definition 1.14. Let R be a continuous semiring and $R' \subseteq R$, let $Y = p(Y)$ be an R'-algebraic system with n equations and let $\sigma \in R^n$. Then σ is called *solution of* $Y = p(Y)$ if $\sigma = p(\sigma)$.

Example 1.11. Every $a \in R$ is unique solution of the trivial R-algebraic system $y_1 = a$.

This example illustrates that it is usually more interesting to consider R'-algebraic systems for a strict subset R' of R; if $R' = R$, then every element can be defined as solution of a trivial system as above.

Example 1.12. Consider the $\mathbb{R}_+^\infty$-algebraic system

$$y_1 = \frac{1}{4} y_1 + \frac{1}{2}.$$

This has – as can be checked by a quick calculation – the solution $\frac{2}{3} \in \mathbb{R}_+^\infty$. In addition, it also has the solution $\infty \in \mathbb{R}_+^\infty$. Moreover, it is clear that $\infty \in \mathbb{R}_+^\infty$ is a solution for many similar algebraic systems.

This example illustrates that the "non-infinite", i.e. smaller, solutions tend to be more interesting than the infinite solutions. We will make this more precise soon.

Example 1.13. Let $A = \{a, b\}$ and consider the following A^*-algebraic system[4] in the continuous semiring $\mathcal{P}(A^*)$:

$$y = ayb \cup \varepsilon$$

A solution of this system is $\{a^i b^i \mid i \in \mathbb{N}\}$ as the following calculation shows:

$$a\{a^i b^i \mid i \in \mathbb{N}\}b \cup \{\varepsilon\} = \{a^{i+1} b^{i+1} \mid i \in \mathbb{N}\} \cup \{a^0 b^0\} = \{a^i b^i \mid i \in \mathbb{N}\}.$$

Example 1.14. Again in $\mathcal{P}(\{a, b\}^*)$ consider the $\{a, b\}^*$-algebraic system

$$\begin{aligned} y_1 &= y_2 y_2 \\ y_2 &= y_1 y_1 \end{aligned}$$

A solution of this system is $\begin{pmatrix} \varnothing \\ \varnothing \end{pmatrix}$, other solutions are $\begin{pmatrix} \{\varepsilon\} \\ \{\varepsilon\} \end{pmatrix}$ or $\begin{pmatrix} a^* \\ a^* \end{pmatrix}$ or $\begin{pmatrix} (a^2)^* \\ (a^2)^* \end{pmatrix}$ or more generally $\begin{pmatrix} L^* \\ L^* \end{pmatrix}$ for any $L \in \mathcal{P}(A^*)$ but not $\begin{pmatrix} a(a^2)^* \\ a(a^2)* \end{pmatrix}$ because $a(a^2)^* a(a^2)^* = (a^2)^+ \neq a(a^2)^*$

[4] A^* is not a subset of $\mathcal{P}(A^*)$ and hence there is no such thing as a A^*-algebraic system in $\mathcal{P}(A^*)$. But – along the lines of the notation introduced in Example 1.9 – we abbreviate "$\{\{w\} \mid w \in A^*\}$-algebraic system" as "A^*-algebraic system".

1.3.3 Solvability

We will now show that every algebraic system is solvable and has a unique least solution. To that aim, we first have to observe that the product of complete partial orders is a complete partial order.

Lemma 1.3. *Let I be a set, for all $i \in I$ let $(S_i, \leqslant_i)$ be a complete partial order. Define a relation $\leqslant$ on $S := \prod_{i \in I} S_i$ by*

$$(x_i)_{i\in I} \leqslant (y_i)_{i\in I} \quad \textit{iff} \quad x_i \leqslant_i y_i \textit{ for all } i \in I.$$

Then $(S, \leqslant)$ is a complete partial order. Moreover, if $(x_{0,i})_{i\in I} \leqslant (x_{1,i})_{i\in I} \leqslant \cdots$ is an increasing sequence in S, then

$$\sup\{(x_{n,i})_{i\in I} \mid n \in \mathbb{N}\} = (\sup\{x_{n,i} \mid n \in \mathbb{N}\})_{i\in I}.$$

Proof. It is easy to check that $(S, \leqslant)$ is a partial order. For completeness let $(x_{0,i})_{i\in I} \leqslant (x_{1,i})_{i\in I} \leqslant \cdots$ be an increasing sequence in S. Then, for all $i \in I$, $x_{0,i} \leqslant_i x_{1,i} \leqslant_i \cdots$ is an increasing sequence in S_i and hence it has a supremum $x_i = \sup\{x_{n,i} \mid n \in \mathbb{N}\}$. We claim that $(x_i)_{i\in I} = \sup\{(x_{n,i})_{i\in I} \mid n \in \mathbb{N}\}$. First, $(x_i)_{i\in I}$ is an upper bound because $x_i \geqslant_i x_{n,i}$ for all $n \in \mathbb{N}$. Let now $(b_i)_{i\in I}$ be an upper bound as well, then $b_i \geqslant_i x_{n,i}$ for all $n \in \mathbb{N}, i \in I$. Since x_i was the least upper bound of $\{x_{n,i} \mid n \in \mathbb{N}\}$ we have $b_i \geqslant_i x_i$ and consequently $(b_i)_{i\in I} \geqslant (x_i)_{i\in I}$. □

Definition 1.15. Let R be a continuous semiring, $R' \subseteq R$ and $Y = \{y_1, \ldots, y_n\}$ a set of variables. A solution σ of an R'-algebraic system $Y = p(Y)$ is called *least solution* of $Y = p(Y)$ if for every solution τ of $Y = p(Y)$: $\sigma \sqsubseteq \tau$.

In the above definition, $\sqsubseteq$ is the order obtained from the n-fold product of the natural order on R. By Lemma 1.3, this is a complete partial order. Note that, if $Y = p(Y)$ has a least solution, then it is unique (due to anti-symmetry of $\sqsubseteq$).

There is a very specific reason for why we are interested in a *least* solution: keeping in mind the relation between context-free grammars and systems of algebraic equations discussed in the beginning of this chapter, consider the fact that the productions of a context-free grammar provide an *inductive definition* of the language generated by the grammar. An inductive definition has always two aspects: on the one hand the operation permitted for generating elements of the inductively defined set and, on the other hand, the understanding that the set thus defined contains *only* objects obtained from the permitted operation. This second aspect is usually self-evident and thus not mentioned. It is also implicit in the definition of the language of a context-free grammar. However, it gets lost when we move to considering *any* solution of a system of equations. In this context, we have to make it explicit by asking for a least solution.

Given polynomials $p_1, \ldots, p_k \in R'[y_1, \ldots, y_n]$ we can form the vector $p = \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix}$ which induces the polynomial function $p : R^n \to R^k$, $\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \mapsto \begin{pmatrix} p_1(a_1, \ldots, a_n) \\ \vdots \\ p_k(a_1, \ldots, a_n) \end{pmatrix}$. For the sake of brevity, such a vector of polynomials p will also be called a polynomial.

Lemma 1.4. *Let R be a continuous semiring, then every polynomial $p : R^n \to R^k$ is continuous.*

Proof. will be done as exercise. □

Theorem 1.2. *Let R be a continuous semiring, $R' \subseteq R$ and let $Y = p(Y)$ be an R'-algebraic system in n variables. Then $Y = p(Y)$ has a least solution in R^n. This least solution is* $\mathrm{lfp}(p) = \sup\{p^i(0) \mid i \in \mathbb{N}\}$

Proof. The solutions of $Y = p(Y)$ are the fixed points of $p : R^n \to R^n$. By Lemma 1.3 we know that R^n is a complete partial order. R^n has a least element 0. By Lemma 1.4 we know that $p : R^n \to R^n$ is continuous. Therefore the fixed point theorem applies and shows that p has a least fixed point (and hence that $Y = p(Y)$ has a least solution) and that this solution is $\mathrm{lfp}(p) = \sup\{p^i(0) \mid i \in \mathbb{N}\}$. □

Example 1.15. Continuing Example 1.13 let $A = \{a, b\}$ and consider the A^*-algebraic system

$$y = ayb \cup \varepsilon.$$

We can now use the fixed point theorem to compute a solution as follows. First, observe that

$$\begin{aligned}
p^0(\varnothing) &= \varnothing \\
p^1(\varnothing) &= a\varnothing b \cup \varepsilon = \{\varepsilon\} \\
p^2(\varnothing) &= a\{\varepsilon\}b \cup \varepsilon = \{ab, \varepsilon\} \\
p^3(\varnothing) &= a\{ab, \varepsilon\}b \cup \varepsilon = \{aabb, ab, \varepsilon\} \\
&\vdots
\end{aligned}$$

At this point one may form the conjecture that $p^n(\varnothing) = \{a^i b^i \mid 0 \leqslant i < n\}$. Let us show this for all $n \in \mathbb{N}$ by induction. For $n = 0$ this has been shown in the above calculation. For the induction step, observe that

$$\begin{aligned}
p^{n+1}(\varnothing) = p(p^n(\varnothing)) &= a\{a^i b^i \mid 0 \leqslant i < n\}b \cup \varepsilon \\
&= \{a^{i+1}b^{i+1} \mid 0 \leqslant i < n\} \cup \varepsilon = \{a^i b^i \mid 0 \leqslant i < n + 1\}.
\end{aligned}$$

Now, by Theorem 1.2, we know that the least solution of $y = p(y)$ is $\sup\{p^n(\varnothing) \mid n \in \mathbb{N}\}$. In $\mathcal{P}(A^*)$ the supremum is infinite union, so the least solution of $y = p(y)$ is

$$\sup\{p^n(\varnothing) \mid n \in \mathbb{N}\} = \bigcup_{n \in \mathbb{N}} p^n(\varnothing) = \bigcup_{n \in \mathbb{N}} \{a^i b^i \mid 0 \leqslant i < n\} = \{a^i b^i \mid i \in \mathbb{N}\}.$$

1.3.4 Context-free grammars

A particularly important special case of algebraic systems are context-free grammars. As a reminder (and already in a suitable notation for our purposes), a context-free grammar is a tuple $G = \langle Y, A, P, y_1 \rangle$ where $Y = \{y_1, \ldots, y_n\}$ are the nonterminals, A are the terminals and $P \subseteq Y \times (Y \cup A)^*$ are the production rules. A production rule (y, α) is usually written as $y \to \alpha$, a finite set of production rules with the same left-hand side is usually written as $y \to \alpha_1 \mid \cdots \mid \alpha_n$. The nonterminal y_1 is the starting symbol.

The one-step derivation relation of G is defined as $\alpha \Longrightarrow_G \alpha'$ if $\alpha = \alpha_1 y \alpha_2$, $\alpha' = \alpha_1 \beta \alpha_2$, and $y \to \beta$ is a production rule of G. Derivability in at most k steps is denoted as $\alpha \Longrightarrow_G^{\leqslant k} \alpha'$ and derivation in a finite number of steps is denoted as $\alpha \Longrightarrow_G^* \alpha'$. If the grammar is clear from the context we often omit the subscript G. The language of a grammar is defined as $L(G) = \{w \in A^* \mid y_1 \Longrightarrow^* w\}$.

Example 1.16. $G = \langle Y, A, P, y_1\rangle$ for $Y = \{y_1, y_2, y_3\}$, $A = \{a, b\}$ and $P =$

$$\begin{aligned}
y_1 &\to y_2by_3\\
y_2 &\to ay_2 \mid \varepsilon\\
y_3 &\to ay_3 \mid by_3 \mid \varepsilon
\end{aligned}$$

is a context-free grammar.

Definition 1.16. Let $G = \langle Y, A, P, y_1\rangle$ be a context-free grammar and $Y = \{y_1, \ldots, y_n\}$. Then we define the *corresponding* A^**-algebraic system* in the continuous semiring $\mathcal{P}(A^*)$ of formal languages to be $Y = p(Y)$ where

$$p_i(Y) = \bigcup_{(y_i,\alpha)\in P} \alpha \qquad \text{for } 1 \leqslant i \leqslant n.$$

Example 1.17. The grammar G from Example 1.16 corresponds to the following $\{a, b\}^*$-algebraic system in $\mathcal{P}(A^*)$:

$$\begin{aligned}
y_1 &= y_2by_3\\
y_2 &= ay_2 \cup \varepsilon\\
y_3 &= ay_3 \cup by_3 \cup \varepsilon
\end{aligned}$$

Theorem 1.3. *Let $G = \langle Y, A, P, y_1\rangle$ be a context-free grammar and let $Y = p(Y)$ the corresponding algebraic system with least solution $\sigma = (\sigma_1, \ldots, \sigma_n)$. Let $G_i = \langle Y, A, P, y_i\rangle$. Then $L(G_i) = \sigma_i$.*

Proof. For $t \in \mathbb{N}$, let $L_i^t = \{w \in A^* \mid y_i \Longrightarrow^{\leqslant t} w\}$. For the left-to-right inclusion we will show by induction on t that $L_i^t \subseteq p^t(\varnothing^n)_i$. For $t = 0$ we have $L_i^0 = \varnothing = p^0(\varnothing^n)_i$. So let $w \in L_i^{t+1}$, i.e., $y_i \Longrightarrow v_0y_{i_1}v_1 \cdots y_{i_k}v_k \Longrightarrow^{\leqslant t} v_0w_1v_1 \cdots w_kv_k = w$. Then $w_j \in L_{i_j}^t$ and by induction hypothesis $w_j \in p^t(\varnothing^n)_{i_j}$ and hence $w \in v_0p^t(\varnothing^n)_{i_1}v_1 \cdots p^t(\varnothing^n)_{i_k}v_k \subseteq p_i(p^t(\varnothing^n)) = p^{t+1}(\varnothing^n)_i$. Thus we obtain $L(G_i) = \bigcup_{t\in\mathbb{N}} L_i^t \subseteq \bigcup_{t\in\mathbb{N}} p^t(\varnothing^n)_i = \sigma_i$.

For the right-to-left inclusion, we will show $p^t(\varnothing^n)_i \subseteq L(G_i)$ for $1 \leqslant i \leqslant n$ by induction on t. For $t = 0$ we have $p^0(\varnothing^n)_i = \varnothing \subseteq L(G_i)$. So let $w \in p^{t+1}(\varnothing^n)_i = p_i(p^t(\varnothing^n))$. By induction hypothesis we have $w \in p_i(L(G_1), \ldots, L(G_n))$, i.e. there is a product term $v_0y_{i_1}v_1 \cdots y_{i_k}v_k$ in p_i s.t. $w \in v_0L(G_{i_1})v_1 \cdots L(G_{i_k})v_k$, i.e. there are words $w_j \in L(G_{i_j})$ s.t. $w = v_0w_1v_1 \cdots w_kv_k$. Since $y_{i_j} \Longrightarrow^* w_j$, we also have $y_i \Longrightarrow v_0y_{i_1}v_1 \cdots y_{i_k}v_k \Longrightarrow^* w$ and hence $w \in L(G_i)$. So $L(G_i)$ is an upper bound of $\{p^t(\varnothing^n)_i \mid t \in \mathbb{N}\}$ and since σ_i is the least upper bound of $\{p^t(\varnothing^n)_i \mid t \in \mathbb{N}\}$ we have $\sigma_i \subseteq L(G_i)$. □

1.4 Formal series

1.4.1 The continuous semiring $R\langle\langle A^*\rangle\rangle$

In previous courses you have already seen formal power series with, e.g., real-valued coefficients. These are expressions of the form $\sum_{i\in\mathbb{N}} a_iq^i$ with $a_i \in \mathbb{R}$ and q being a variable. Sometimes one would like to evaluate a power series (i.e. substitute a concrete value for q) but often one treats a formal power series as a formal object rather than as a function $q \mapsto \sum_{i\in\mathbb{N}} a_iq^i$, hence the term *formal* power series.

In this course we will consider formal series in another context: instead of working over a single variable q (and hence the multiplicative monoid freely generated by q) we work over an alphabet

A (and hence the multiplicative (non-commutative) monoid freely generated by A). Since the elements of A^* are no longer powers of a word we will simply speak about formal series.

Definition 1.17. Let A be an alphabet and R a continuous semiring. A *formal series over A^* with coefficients in R* is a function $r : A^* \to R$. We often write (r, w) for $r(w)$ and a formal series as a whole is written as $\sum_{w \in A^*}(r, w)w$. We write $R\langle\!\langle A^*\rangle\!\rangle$ for the set of all formal series over A^* with coefficients in R.

We define the structure $\langle R\langle\!\langle A^*\rangle\!\rangle, +, 0, \cdot, 1\rangle$ as follows: addition and (Cauchy-)multiplication on $R\langle\!\langle A^*\rangle\!\rangle$ are defined as usual:

$$\begin{aligned}
(\sum_{w\in A^*}(r,w)w) + (\sum_{w\in A^*}(s,w)w) &= \sum_{w\in A^*}((r,w)+(s,w))w \\
(\sum_{w\in A^*}(r,w)w) \cdot (\sum_{w\in A^*}(s,w)w) &= \sum_{w\in A^*}(\sum_{\substack{u,v\in A^*\\ uv=w}}(r,u)(s,v))w
\end{aligned}$$

0 is the formal series where all coefficients are 0 and 1 is the formal series 1ε.

The relationship between formal series over A^* and formal languages is that a formal language $L \in \mathcal{P}(A^*)$ can be identified with a function $\chi_L : A^* \to \{0, 1\}$, i.e., a formal series with coefficients in $\{0, 1\}$. This relationship will be analysed in more depth in Proposition 1.7.

It is also possible to consider formal series over an arbitrary monoid M (instead of the monoid freely generated by A). But then many results of formal language theory do not longer hold (in particular: Kleene's theorem which states that the rational languages are exactly the recognisable languages fails). Therefore, in this course, we will only consider formal series over A^*.

Proposition 1.5. *Let A be an alphabet and let R be a continuous semiring. Then $R\langle\!\langle A^*\rangle\!\rangle$ is a continuous semiring.*

Proof. It is straightforward to show that $R\langle\!\langle A^*\rangle\!\rangle$ is a semiring[5]. We have $r_1 \sqsubseteq r_2$ iff $(r_1, w) \sqsubseteq (r_2, w)$ for all $w \in A^*$, i.e. the natural order on $R\langle\!\langle A^*\rangle\!\rangle$ is the product of countably many copies of the natural order in A. By Lemma 1.3 this is a complete partial order and its supremum is given component-wise, i.e. for any increasing sequence $r_0 \sqsubseteq r_1 \sqsubseteq \cdots$ we have

$$\sup\{\sum_{w\in A^*}(r_i,w)w \mid i \in \mathbb{N}\} = \sum_{w\in A^*}\sup\{(r_i,w) \mid i\in\mathbb{N}\}w. \tag{*}$$

It remains to show that $+$ and $\cdot$ are continuous. We have

$$\begin{aligned}
r + \sup\{s_i \mid i\in\mathbb{N}\} &= \sum_{w\in A^*}(r,w)w + \sup\{\sum_{w\in A^*}(s_i,w)w \mid i\in\mathbb{N}\} \\
&=^{(*)} \sum_{w\in A^*}(r,w)w + \sum_{w\in A^*}\sup\{(s_i,w)\mid i\in\mathbb{N}\}w \\
&= \sum_{w\in A^*}\big((r,w) + \sup\{(s_i,w)\mid i\in\mathbb{N}\}\big)w \\
&= \sum_{w\in A^*}\sup\{(r,w)+(s_i,w)\mid i\in\mathbb{N}\}w \\
&=^{(*)} \sup\{\sum_{w\in A^*}\big((r,w)+(s_i,w)\big)w \mid i\in\mathbb{N}\} \\
&= \sup\{\sum_{w\in A^*}(r,w)w + \sum_{w\in A^*}(s_i,w)w \mid i\in\mathbb{N}\} \\
&= \sup\{r + s_i \mid i\in\mathbb{N}\}.
\end{aligned}$$

[5] Prove some properties as exercise!

This suffices for $+$ since $+$ is commutative. For $\cdot$ we have:

$$\begin{aligned}
r \cdot \sup\{s_i \mid i \in \mathbb{N}\} &= \sum_{w \in A^*} (r,w)w \cdot \sup\{ \sum_{w \in A^*} (s_i,w)w \mid i \in \mathbb{N}\} \\
&=^{(*)} \sum_{w \in A^*} (r,w)w \cdot \sum_{w \in A^*} \sup\{(s_i,w) \mid i \in \mathbb{N}\}w \\
&= \sum_{w \in A^*} \big(\sum_{\substack{u,v \in A^* \\ uv=w}} (r,u) \sup\{(s_i,v) \mid i \in \mathbb{N}\}\big)w \\
&= \sum_{w \in A^*} \big(\sum_{\substack{u,v \in A^* \\ uv=w}} \sup\{(r,u)(s_i,v) \mid i \in \mathbb{N}\}\big)w
\end{aligned}$$

Now there is only finitely many pairs u,v s.t. $uv = w$ and a finite sum in R is a continuous function. Therefore we can apply Lemma 1.1 in order to obtain

$$\begin{aligned}
&= \sum_{w \in A^*} \sup\{ \sum_{\substack{u,v \in A^* \\ uv=w}} (r,u)(s_i,v) \mid i \in \mathbb{N}\}w \\
&=^{(*)} \sup\{ \sum_{w \in A^*} \big(\sum_{\substack{u,v \in A^* \\ uv=w}} (r,u)(s_i,v)\big)w \mid i \in \mathbb{N}\} \\
&= \sup\{r \cdot s_i \mid i \in \mathbb{N}\}.
\end{aligned}$$

For multiplication from the right an analogous calculation shows continuity. □

Having shown that $R\langle\langle A^*\rangle\rangle$ is a continuous semiring we obtain a $*$-operation on formal series. Before we analyse this operation in more detail, we make the preparatory observation that infinite sums of formal series are given component-wise in the following Lemma.

Lemma 1.5. *Let A be an alphabet and R be a continuous semiring. Let $r_0, r_1, \ldots \in R\langle\langle A^*\rangle\rangle$. Then*

$$\sum_{n \in \mathbb{N}} r_n = \sum_{n \in \mathbb{N}} \sum_{w \in A^*} (r_n,w)w = \sum_{w \in A^*} \sum_{n \in \mathbb{N}} (r_n,w)w.$$

Proof. We have

$$\begin{aligned}
\sum_{n \in \mathbb{N}} \sum_{w \in A^*} (r_n,w)w &= \sup\{\sum_{i=0}^{n} \sum_{w \in A^*} (r_i,w)w \mid n \in \mathbb{N}\} \\
&= \sup\{ \sum_{w \in A^*} \sum_{i=0}^{n} (r_i,w)w \mid n \in \mathbb{N}\} \\
&=^{(*)} \sum_{w \in A^*} \sup\{\sum_{i=0}^{n} (r_i,w) \mid n \in \mathbb{N}\}w \\
&= \sum_{w \in A^*} \big(\sum_{n \in \mathbb{N}} (r_n,w)\big)w
\end{aligned}$$

where $(*)$ refers to the equation $(*)$ in the proof of Proposition 1.5. □

The star-operation in a continuous semiring is defined as $r^* = \sum_{n \in \mathbb{N}} r^n$. The following proposition further explicates this definition for the case of a semiring of formal series: the coefficient (r^*, w) is obtained as sum over the values of all decompositions of w where the value of a decomposition is the product of the values of its components. In order to state and prove this proposition we first make the notion of decomposition precise. If S is a set and $k \in \mathbb{N}$ we write S^k for the set of k-tuples of elements of S. In particular, $S^0 = \{()\}$, the set consisting of the empty tuple. A decomposition of $w \in A^*$ is a tuple $(u_1, \ldots, u_n) \in (A^*)^n$ for some $n \geqslant 0$ s.t. $u_1 \cdots u_n = w$. Note that $u_i = \varepsilon$ is possible so that every word has infinitely many decompositions. In particular, $w = \varepsilon$ has the decompositions $(), (\varepsilon), (\varepsilon, \varepsilon), \ldots$.

Proposition 1.6. *Let A be an alphabet, R a continuous semiring, and $r \in R\langle\langle A^*\rangle\rangle$. Then*

$$r^* = \sum_{w \in A^*} \sum_{n \in \mathbb{N}} \sum_{\substack{(u_1,\ldots,u_n) \in (A^*)^n \\ u_1 \cdots u_n = w}} \prod_{i=1}^{n} (r, u_i) w$$

Proof. We will first prove that, for all $w \in A^*$ and all $n \in \mathbb{N}$

$$(r^n, w) = \sum_{\substack{(u_1,\ldots,u_n) \in (A^*)^n \\ u_1 \cdots u_n = w}} \prod_{i=1}^{n} (r, u_i) \tag{*}$$

by induction on n. For the case $n = 0$ observe that

$$\sum_{\substack{(u_1,\ldots,u_0) \in (A^*)^0 \\ u_1 \cdots u_0 = w}} \prod_{i=1}^{0} (r, u_i) = \sum_{\substack{() \in (A^*)^0 \\ \varepsilon = w}} 1 = \begin{cases} 1 & \text{if } w = \varepsilon \\ 0 & \text{if } w \neq \varepsilon \end{cases}$$

which is (r^0, w). For the induction step, we have

$$\begin{aligned}(r^{n+1}, w) = (r \cdot r^n, w) = \sum_{\substack{u_1, v \in A^* \\ u_1 v = w}} (r, u_1)(r^n, v) &=^{\text{IH}} \sum_{\substack{u_1, v \in A^* \\ u_1 v = w}} (r, u_1) \sum_{\substack{v_1,\ldots,v_n \in A^* \\ v_1 \cdots v_n = v}} \prod_{i=1}^{n} (r, v_i) \\ &= \sum_{\substack{(u_1,\ldots,u_{n+1}) \in (A^*)^{n+1} \\ u_1 \cdots u_{n+1} = w}} \prod_{i=1}^{n+1} (r, u_i).\end{aligned}$$

We then have

$$r^* = \sum_{n \in \mathbb{N}} r^n =^{\text{Lem. 1.5}} \sum_{w \in A^*} \Big(\sum_{n \in \mathbb{N}} (r^n, w)\Big) w = \sum_{w \in A^*} \Big(\sum_{n \in \mathbb{N}} \sum_{\substack{(u_1,\ldots,u_n) \in (A^*)^n \\ u_1 \cdots u_n = w}} \prod_{i=1}^{n} (r, u_i)\Big) w.$$

□

Corollary 1.1. *Let A be an alphabet, R a continuous semiring, and $r \in R\langle\langle A^*\rangle\rangle$ with $(r, \varepsilon) = 0$. Then*

$$r^* = \sum_{w \in A^*} \sum_{n=0}^{|w|} \sum_{\substack{(u_1,\ldots,u_n) \in (A^*)^n \\ u_1 \cdots u_n = w}} \prod_{i=1}^{n} (r, u_i) w$$

Proof. If $u_1, \ldots, u_n \in A^*$ with $u_1 \cdots u_n = w$ and $u_i = \varepsilon$ for some $i \in \{1, \ldots, n\}$, then $\prod_{i=1}^{n} (r, u_i) = 0$, so it suffices to consider decompositions into non-emtpy words. Every such decomposition of w consists of at most $|w|$ words. □

In particular, $(r, \varepsilon) = 0$ implies $(r^*, \varepsilon) = 1$.

1.4.2 $\mathcal{P}(A^*)$ and $\mathbb{B}\langle\langle A^*\rangle\rangle$

Definition 1.18. Let R and S be semirings. A function $\varphi : R \to S$ is called *semiring homomorphism* if $\varphi(0) = 0$, $\varphi(1) = 1$, $\varphi(x + y) = \varphi(x) + \varphi(y)$ and $\varphi(x \cdot y) = \varphi(x) \cdot \varphi(y)$.

Often, it will be clear from the context that we speak about *semiring* homomorphisms. Then we will just say homomorphism. As always, an isomorphism is a bijective homomorphism.

Lemma 1.6. *Let R and S be continuous semirings and $\varphi : R \to S$ an isomorphism. Then $\varphi(x^*) = \varphi(x)^*$ for all $x \in R$.*

Proof. First, note that $x \sqsubseteq y$ iff $\varphi(x) \sqsubseteq \varphi(y)$: for if $\exists z\, x + z = y$, then $\varphi(x) + \varphi(z) = \varphi(y)$. In the other direction, if $\exists z\, \varphi(x) + z = \varphi(y)$, then $x + \varphi^{-1}(z) = y$. So φ is an isomorphism for the partial orders $(R, \sqsubseteq)$ and $(S, \sqsubseteq)$ and therefore

$$\varphi(\sup\{x_i \mid i \in \mathbb{N}\}) = \sup\{\varphi(x_i) \mid i \in \mathbb{N}\}.$$

Then, for any $x \in R$, we have

$$\varphi(x^*) = \varphi(\sup\{\sum_{i=0}^{n} x^i \mid n \geqslant 0\}) = \sup\{\sum_{i=0}^{n} \varphi(x)^i \mid n \geqslant 0\} = \varphi(x)^*,$$

□

Proposition 1.7. *The continuous semirings $\langle \mathcal{P}(A^*), \cup, \varnothing, \cdot, \{\varepsilon\}\rangle$ and $\mathbb{B}\langle\langle A^*\rangle\rangle$ are isomorphic.*

Proof. The mapping $\varphi : \mathcal{P}(A^*) \to \mathbb{B}\langle\langle A^*\rangle\rangle, L \mapsto \sum_{w \in A^*} \chi_L(w)w$ is clearly a bijection. It is also a homomorphism since $\varphi(\varnothing) = 0, \varphi(\{\varepsilon\}) = \varepsilon$, $\varphi(L_1 \cup L_2) = \varphi(L_1) + \varphi(L_2)$ and

$$\begin{aligned}\varphi(L_1) \cdot \varphi(L_2) &= (\sum_{w_1 \in A^*} \chi_{L_1}(w_1)w_1)(\sum_{w_2 \in A^*} \chi_{L_2}(w_2)w_2) \\ &= \sum_{w \in A^*} (\sum_{\substack{w = w_1 w_2 \\ w_1 \in A^*, w_2 \in A^*}} \chi_{L_1}(w_1)\chi_{L_2}(w_2))w = \sum_{w \in L_1 \cdot L_2} w = \varphi(L_1 \cdot L_2).\end{aligned}$$

□

So we can identify a formal language $L \subseteq A^*$ with a formal series over A^* with coefficients in $\mathbb{B}$. This observation clarifies the nature of the generalisation considered: the generalisation of a language is a formal series in the continuous semiring $R\langle\langle A^*\rangle\rangle$. This generalises the notion of a word being element of a language from a Boolean function to a function $r : A^* \to R$. Most of the theory of formal languages thus generalises from $\mathcal{P}(A^*)$ to $R\langle\langle A^*\rangle\rangle$ for an arbitrary continuous semiring R.

1.4.3 Calculations in semirings of formal series

Example 1.18. In $\mathbb{B}\langle\langle \{a,b\}^*\rangle\rangle$ we have:

$$\begin{aligned}(a+b)^n &= (a+b) \cdot \dots \cdot (a+b) = \sum_{w \in \{a,b\}^n} w \\ (a+b)^* &= \sum_{n \geqslant 0} (a+b)^n = \sum_{w \in \{a,b\}^*} w \\ (a+b)^* a (a+b)^* &= \sum_{\substack{w \in \{a,b\}^* \\ n_a(w) \geqslant 1}} w\end{aligned}$$

where $n_x(w) \in \mathbb{N}$ is the number of occurrences of the letter x in the word w. So, computations in $\mathbb{B}\langle\langle A^*\rangle\rangle$ can be carried out just as we know them from working with regular expressions.

Example 1.19. In $\mathbb{N}^\infty\langle\langle\{a,b\}^*\rangle\rangle$ we have:

$$\begin{aligned}(a+b)^n &= (a+b)\cdot\dots\cdot(a+b) = \sum_{w\in\{a,b\}^n} w \\ (a+b)^* &= \sum_{n\geqslant 0}(a+b)^n = \sum_{w\in\{a,b\}^*} w \\ (a+b)^*a(a+b)^* &= \sum_{w\in\{a,b\}^*} n_a(w)w\end{aligned}$$

The third equation is proved in detail as follows: let $r = (a+b)^*a(a+b)^* \in \mathbb{N}^\infty\langle\langle\{a,b\}^*\rangle\rangle$. We will compute the coefficient (r,w) for $w \in \{a,b\}^*$. To that aim, let $r_1 = (a+b)^*, r_2 = a \in \mathbb{N}^\infty\langle\langle\{a,b\}^*\rangle\rangle$. Then the definition of the Cauchy-product implies that

$$(r,w) = \sum_{\substack{u_1,u_2,u_3\in\{a,b\}^*\\ u_1u_2u_3=w}} (r_1,u_1)(r_2,u_2)(r_1,u_3)$$

Since all coefficients of r_1 and r_2 are either 0 or 1, also the product $(r_1,u_1)(r_2,u_2)(r_1,u_3)$ can only be 0 or 1. Moreover, it is 1 iff all factors are 1. But the only word u_2 s.t. $(r_2,u_2) = 1$ is a, so

$$= \sum_{\substack{u_1,u_3\in\{a,b\}^*\\ u_1au_3=w}} 1$$

and this is the number of ways one can write w as u_1au_2 and hence

$$= n_a(w).$$

Example 1.20. Let R be the min-+-semiring $\langle\mathbb{N}^\infty, \min, \infty, +, 0\rangle$ and let $A = \{a,b\}$. The elements of $R\langle\langle A^*\rangle\rangle$ are formal Min-series, i.e., mappings $r : A^* \to \mathbb{N}^\infty$ which we write as $\mathrm{Min}_{w\in A^*}(r,w)w$. The usual notational convention for power series is that all words which are not mentioned explicitly have the additive unit as coefficient. In R the additive unit is ∞, so, for example,

$$0\varepsilon = \begin{cases} 0 & \text{if } w = \varepsilon \\ \infty & \text{otherwise}\end{cases}$$

in $R\langle\langle A^*\rangle\rangle$. The continuous semiring $R\langle\langle A^*\rangle\rangle$ has a min operation which is defined pointwise and whose unit is $\infty : A^* \to R, w \mapsto \infty$. Moreover, it has a plus operation $\oplus$ which is defined via the Cauchy product formula as

$$(r_1 \oplus r_2, w) = \min\{(r_1,u) + (r_2,v) \mid u,v \in A^*, uv = w\}$$

and has 0ε as unit. The formula for the star of a formal series shown in Proposition 1.6 becomes

$$r^* = \underset{w\in A^*}{\mathrm{Min}}\ \underset{n\in\mathbb{N}}{\mathrm{Min}} \min\{\sum_{i=1}^{n}(r,u_i) \mid u_1,\dots,u_n \in A^*, u_1\cdots u_n = w\}w$$

or, put differently:

$$(r^*,w) = \underset{n\in\mathbb{N}}{\mathrm{Min}} \min\{\sum_{i=1}^{n}(r,u_i) \mid u_1,\dots,u_n \in A^*, u_1\cdots u_n = w\}$$

for all $w \in A^*$. In particular, for the calculation of (r^*, w) for $w = babaab$ we have to consider all decompositions of w into $\{a, b, ab\}$ and minimise over their respective values (where the value of a decomposition is the sum of the coefficients of its components, in the case of r this is just the number of components). We have

$$\begin{array}{r|l} b \cdot a \cdot b \cdot a \cdot a \cdot b & 6 \\ b \cdot ab \cdot a \cdot a \cdot b & 5 \\ b \cdot a \cdot b \cdot a \cdot ab & 5 \\ b \cdot ab \cdot a \cdot ab & 4 \end{array}$$

Since all other decompositions lead to a sum of ∞, we have $(r^*, w) = 4$. For the general case, we obtain

$$(r^*, w) = \begin{cases} \infty & \text{if } w = \varepsilon \\ |w| - n_{ab}(w) & \text{otherwise} \end{cases}$$

We have already seen that every context-free grammar can be considered an algebraic system in $\mathcal{P}(A^*)$ and – by the above isomorphism-result Proposition 1.7 – as an algebraic system in $\mathbb{B}\langle\langle A^*\rangle\rangle$. The generalisation from a language $L \in \mathcal{P}(A^*)$ to a formal series $r \in R\langle\langle A^*\rangle\rangle$ hence immediately gives rises to a class of algebraic systems that corresponds to $R\langle\langle A^*\rangle\rangle$ as context-free grammars correspond to $\mathcal{P}(A^*)$: the $\{\{aw\} \mid a \in R, w \in A^*\}$-algebraic systems in $R\langle\langle A^*\rangle\rangle$. The set of components of solutions of these systems hence generalises the notion of a context-free language from $\mathcal{P}(A^*)$ to $R\langle\langle A^*\rangle\rangle$.

Example 1.21. Let us consider the following algebraic system in $\mathbb{R}_+^\infty\langle\langle\{a, b\}^*\rangle\rangle$:

$$y = \frac{1}{2}ay + \frac{1}{4}by + \frac{1}{4}\varepsilon = p(y).$$

We can think of this as a stochastic process which – at each point in time – outputs a with probability $\frac{1}{2}$, b with probability $\frac{1}{4}$ and stops with probability $\frac{1}{4}$. The least solution of this algebraic system is a formal series $\sigma \in \mathbb{R}_+^\infty\langle\langle\{a, b\}^*\rangle\rangle$ s.t. (σ, w) is the probability that this process outputs the word $w \in \{a, b\}^*$. The least element of $\mathbb{R}_+^\infty\langle\langle\{a, b\}^*\rangle\rangle$ is 0, the series all of whose coefficients are $0 \in \mathbb{R}_+^\infty$. We compute σ by proceeding as in Example 1.15:

$$\begin{aligned} p^0(0) &= 0 \\ p^1(0) &= \frac{1}{4}\varepsilon \\ p^2(0) &= \frac{1}{2}a\frac{1}{4}\varepsilon + \frac{1}{4}b\frac{1}{4}\varepsilon + \frac{1}{4}\varepsilon = \frac{1}{8}a + \frac{1}{16}b + \frac{1}{4}\varepsilon \\ p^3(0) &= \ldots \\ &\vdots \end{aligned}$$

We form the conjecture

$$p^n(0) = \sum_{w \in \{a,b\}^{<n}} \frac{1}{4}\left(\frac{1}{2}\right)^{n_a(w)}\left(\frac{1}{4}\right)^{n_b(w)} w$$

and prove it by induction. For $n = 0, 1, 2$ it is shown above. For the induction step we have

$$\begin{aligned} p^{n+1}(0) &=^{\text{IH}} \frac{1}{2}a \sum_{w \in \{a,b\}^{<n}} \frac{1}{4}\left(\frac{1}{2}\right)^{n_a(w)}\left(\frac{1}{4}\right)^{n_b(w)} w + \frac{1}{4}b \sum_{w \in \{a,b\}^{<n}} \frac{1}{4}\left(\frac{1}{2}\right)^{n_a(w)}\left(\frac{1}{4}\right)^{n_b(w)} w + \frac{1}{4}\varepsilon \\ &= \sum_{w \in a\{a,b\}^{<n}} \frac{1}{4}\left(\frac{1}{2}\right)^{n_a(w)}\left(\frac{1}{4}\right)^{n_b(w)} w + \sum_{w \in b\{a,b\}^{<n}} \frac{1}{4}\left(\frac{1}{2}\right)^{n_a(w)}\left(\frac{1}{4}\right)^{n_b(w)} w + \frac{1}{4}\varepsilon \end{aligned}$$

and since $\{a,b\}^{<n+1} = a\{a,b\}^{<n} \cup b\{a,b\}^{<n} \cup \{\varepsilon\}$ we have

$$= \sum_{w \in \{a,b\}^{<n+1}} \frac{1}{4}\left(\frac{1}{2}\right)^{n_a(w)}\left(\frac{1}{4}\right)^{n_b(w)} w.$$

By the fixed point theorem we know that $\sigma = \sup\{p^n(0) \mid n \in \mathbb{N}\}$ so we obtain

$$\sigma = \sup\Big\{ \sum_{w \in \{a,b\}^{<n}} \frac{1}{4}\left(\frac{1}{2}\right)^{n_a(w)}\left(\frac{1}{4}\right)^{n_b(w)} w \mid n \in \mathbb{N}\Big\}$$

and since the supremum of a sequence of formal series is computed component-wise and the components only change once each (from 0 to the final value) we have

$$= \sum_{w \in \{a,b\}^{*}} \frac{1}{4}\left(\frac{1}{2}\right)^{n_a(w)}\left(\frac{1}{4}\right)^{n_b(w)} w.$$

1.4.4 Context-free grammars and $\mathbb{N}^{\infty}$

Moving to a more general semiring can also provide a means for analysing aspects of elementary notions from formal language theory. We will see an example for this now. Let $G = \langle Y, A, P, y_1\rangle$ be a context-free grammar and $Y = \{y_1, \ldots, y_n\}$. We define the one-step leftmost derivation relation as $\alpha \overset{lm}{\Longrightarrow}_G \alpha'$ if $\alpha = \alpha_1 y \alpha_2$, $\alpha' = \alpha_1\beta\alpha_2$, $y \to \beta$ is a production rule of G, $\alpha_1 \in A^*$ and $\alpha_2 \in (A \cup Y)^*$. It is easy to show that a word is derivable in a grammar iff it has a leftmost derivation.

Example 1.22. Let $A = \{a, b, c, \circ\}$ and define a grammar by the production rules

$$S \to S \circ S \mid a \mid b \mid c$$

Then the word $b \circ a \circ b$ has the following two distinct leftmost derivations:

$$S \Longrightarrow S \circ S \Longrightarrow b \circ S \Longrightarrow b \circ S \circ S \Longrightarrow^{*} b \circ a \circ b$$
$$S \Longrightarrow S \circ S \Longrightarrow S \circ S \circ S \Longrightarrow^{*} b \circ a \circ b$$

In contexts where $\circ$ is interpreted as a non-associative operation these two different leftmost derivations, and hence different parse trees, will yield different interpretations. In many applications this is undesirable, therefore (non-)ambiguity is an important aspect of grammars.

For $w \in L(G_i)$ write $d_i(w)$ for the number of leftmost derivations of the word w in $G_i = \langle Y, A, P, y_i\rangle$. Ambiguity can be characterised nicely by transforming a context-free grammar into an algebraic system in $\mathbb{N}^{\infty}\langle\langle A^*\rangle\rangle$ instead of $\mathcal{P}(A^*)$. This is done just as in Definition 1.16. Then one can obtain the following:

Theorem 1.4. *Let $G = \langle Y, A, P, y_1\rangle$ be a context-free grammar, let $Y = \{y_1, \ldots, y_n\}$, and let $Y = p(Y)$ the corresponding algebraic system in $\mathbb{N}^{\infty}\langle\langle A^*\rangle\rangle$ with least solution $\sigma = (\sigma_1, \ldots, \sigma_n)$. Then $\sigma_i = \sum_{w \in A^*} d_i(w)w$ for $1 \leqslant i \leqslant n$.*

1.5 Matrices

Automata are a central notion of formal language theory. In order to develop our theory of automata we first study matrices over a continuous semiring. This section is devoted to them, in particular to proving that the square matrices over a continuous semiring form a continuous semiring. This will, as in the case of formal series, give us a star of matrices. We will then see how to calculute the star of a matrix in terms of $+$, $\cdot$ and star of the underlying semiring.

1.5.1 The continuous semiring $R^{I\times I}$

Definition 1.19. Let R be a semiring and I, J finite sets. A mapping $M : I \times J \to R$ is called *matrix*. The values of M are denoted by $M_{i,j}$ for $i \in I, j \in J$. The set of all such matrices is denoted by $R^{I\times J}$.

For $M, N \in R^{I\times J}$ we define:

$$(M + N)_{i,j} = M_{i,j} + N_{i,j}.$$

We also define $0 \in R^{I\times J}$ by $0_{i,j} = 0$

For $M \in R^{I\times J}$ and $N \in R^{J\times K}$ we define:

$$(M \cdot N)_{i,k} = \sum_{j\in J} M_{i,j}N_{j,k}$$

We also define $1 \in R^{I\times I}$ by $1_{i,j} = 1$ if $i = j$ and 0 otherwise.

We thus obtain a structure $\langle R^{I\times I}, +, 0, \cdot, 1\rangle$. It is straightforward to show that, if R is a semiring, then also $\langle R^{I\times I}, +, 0, \cdot, 1\rangle$ is[6].

Proposition 1.8. *Let R be a continuous semiring, then $R^{I\times I}$ is a continuous semiring.*

Proof. $R^{I\times I}$ is a semiring. Note that $M \sqsubseteq N$ iff $M_{i,j} \sqsubseteq N_{i,j}$ for all $i, j \in I$. Therefore we can apply Lemma 1.3 to conclude that that $(R^{I\times I}, \sqsubseteq)$ is a complete partial order and that, for any increasing sequence $M_0 \sqsubseteq M_1 \sqsubseteq \cdots$, and for all $i, j \in I$ we have

$$(\sup\{M_n \mid n \in \mathbb{N}\})_{i,j} = \sup\{(M_n)_{i,j} \mid n \in \mathbb{N}\} \tag{*}$$

Let us first show that $+$ is continuous. To that aim let $M_0 \sqsubseteq M_1 \sqsubseteq \cdots$ be an increasing sequence and fix $i, j \in I$. Then we have

$$\begin{aligned}(\sup\{M_n \mid n \in \mathbb{N}\} + N)_{i,j} &= (\sup\{M_n \mid n \in \mathbb{N}\})_{i,j} + N_{i,j}\\ &=^{(*)} \sup\{(M_n)_{i,j} \mid n \in \mathbb{N}\} + N_{i,j}\\ &= \sup\{(M_n)_{i,j} + N_{i,j} \mid n \in \mathbb{N}\}\\ &= \sup\{(M_n + N)_{i,j} \mid n \in \mathbb{N}\}\\ &=^{(*)} (\sup\{M_n + N \mid n \in \mathbb{N}\})_{i,j}\end{aligned}$$

And therefore we have $\sup\{M_n + N \mid n \in \mathbb{N}\} = \sup\{M_n \mid n \in \mathbb{N}\} + N$.

[6] Prove some properties as exercise!

For showing that $\cdot$ is continuous, let again $M_0 \sqsubseteq M_1 \sqsubseteq \cdots$ be an increasing sequence and fix $i, k \in I$. Then we have

$$\begin{aligned}(\sup\{M_n \mid n \in \mathbb{N}\} \cdot N)_{i,k} &= \sum_{j \in I} (\sup\{M_n \mid n \in \mathbb{N}\})_{i,j} N_{j,k} \\ &=^{(*)} \sum_{j \in I} \sup\{(M_n)_{i,j} \mid n \in \mathbb{N}\} N_{j,k}\end{aligned}$$

then by continuity of polynomials in R together with Lemma 1.1

$$\begin{aligned}&= \sup\{\sum_{j \in I} (M_n)_{i,j} N_{j,k} \mid n \in \mathbb{N}\} \\ &= \sup\{(M_n \cdot N)_{i,k} \mid n \in \mathbb{N}\} \\ &=^{(*)} (\sup\{M_n \cdot N \mid n \in \mathbb{N}\})_{i,k}\end{aligned}$$

For multiplication from the left proceed analogously. □

Since $R^{I \times I}$ is a continuous semiring, there is also a star-operation on matrices: $M^* = \sum_{i \in \mathbb{N}} M^i$. The star of a matrix will play an important role for finite automata. We will therefore study it more closely here (before moving on to automata).

1.5.2 The star of a matrix

Example 1.23. Let $I = \{1, \ldots, n\}$ and $M \in R^{I \times I}$ be a diagonal matrix. Then

$$M^* = \sum_{i \in \mathbb{N}} M^i = \sum_{i \in \mathbb{N}} \begin{pmatrix} m_1 & & \\ & \ddots & \\ & & m_n \end{pmatrix}^i = \sum_{i \in \mathbb{N}} \begin{pmatrix} m_1^i & & \\ & \ddots & \\ & & m_n^i \end{pmatrix} = \begin{pmatrix} m_1^* & & \\ & \ddots & \\ & & m_n^* \end{pmatrix}$$

In general, the computation of the star of a matrix is considerably more complicated. Before we study an algorithm to compute the star of a matrix, we relate it to more familiar notions: we will now show that the star of a matrix is closely related to the paths in a graph. As usual, a graph is a pair (V, E) where V is a finite set of vertices and $E \subseteq V \times V$ is the set of edges. A path is a list $e_1, \ldots, e_n$ of edges s.t. for all $i = 1 \ldots, n-1$ there are $x, y, z \in V$ s.t. $e_i = (x, y)$ and $e_{i+1} = (y, z)$. If the graph is clear from the context, and $i, j \in V$, we write $P_n(i, j)$ for the set of paths from i to j of length n and $P(i, j)$ for the set of all paths from i to j. If R is a continuous semiring, an R-weighted graph is a tuple (V, E, w) s.t. (V, E) is a graph and $w : E \to R$. We extend w to $w : V \times V \to R$ by setting $w(i, j) = 0$ if $(i, j) \notin E$. Then we can identify an R-weighted graph with a matrix $M \in R^{V \times V}$ by letting $M_{i,j} = w((i, j))$. The weight of a path $p = e_1, \ldots, e_n$ is then defined as $w(p) = \prod_{i=1}^n w(e_i)$. Note that this implicitely defines the weight of the empty path to be $1 \in R$.

Proposition 1.9. *Let R be a continuous semiring and $M = (V, E, w)$ an R-weighted graph. Then $(M^*)_{i,j} = \sum_{p \in P(i,j)} w(p)$.*

Proof. We will first show by induction on n that $(M^n)_{i,j} = \sum_{p \in P_n(i,j)} w(p)$. For $n = 0$ this holds trivially since M^0 is the identity matrix. For the induction step, observe that

$$(M^{n+1})_{i,j} = \sum_{k \in V} M_{i,k} (M^n)_{k,j} =^{\text{IH}} \sum_{k \in V} w((i,k)) \sum_{p \in P_n(k,j)} w(p) = \sum_{\substack{k \in V \\ p \in P_n(k,j)}} w((i,k)) w(p) = \sum_{p \in P_{n+1}(i,j)} w(p).$$

But now

$$(M^*)_{i,j} = (\sum_{n\in\mathbb{N}} M^n)_{i,j} = \sum_{n\in\mathbb{N}}(M^n)_{i,j} = \sum_{n\in\mathbb{N}}\sum_{p\in P_n(i,j)} w(p) = \sum_{p\in P(i,j)} w(p).$$

□

The aim of this section is to prove a theorem that reduces the computation of the star of a matrix to the computation of the stars of smaller matrices. Applying this result recursively gives a procedure to compute the star of any matrix in $R^{I\times I}$ provided we know how to compute the star in R. In order to prove that theorem we need some preliminary results first.

Lemma 1.7. *Let R be a continuous semiring and $x, y \in R$. Then the sum-star equation*

$$(x + y)^* = (x^*y)^*x^* = x^*(yx^*)^*$$

holds.

Proof. First note that, for all $a, b \in R$, we have:

$$(ab)^*a = \big(\sum_{i\in\mathbb{N}}(ab)^i\big)a = \sum_{i\in\mathbb{N}}(ab)^i a = \sum_{i\in\mathbb{N}} a(ba)^i = a\sum_{i\in\mathbb{N}}(ba)^i = a(ba)^*.$$

Therefore it suffices to prove $(x + y)^* = x^*(yx^*)^*$. For $q, r \in \mathbb{N}$ define[7]

$$S_{q,r} = \sum_{\substack{p_0,\ldots,p_r\in\mathbb{N}\\ p_0+\cdots+p_r=q}} x^{p_0}yx^{p_1}\cdots yx^{p_r}.$$

The definition of $S_{q,r}$ directly entails that $(x + y)^i = \sum_{\substack{q,r\in\mathbb{N}\\ q+r=i}} S_{q,r}$, so we have

$$(x + y)^* = \sum_{i\in\mathbb{N}}(x + y)^i = \sum_{\substack{q,r,i\in\mathbb{N}\\ q+r=i}} S_{q,r} = \sum_{q,r\in\mathbb{N}} S_{q,r}.$$

Furthermore, for all $r \in \mathbb{N}$ we have

$$\begin{aligned} x^*(yx^*)^r &= \sum_{p_0\in\mathbb{N}} x^{p_0}y\sum_{p_1\in\mathbb{N}} x^{p_1}\cdots y\sum_{p_r\in\mathbb{N}} x^{p_r} \\ &= \sum_{p_0,\ldots,p_r\in\mathbb{N}} x^{p_0}yx^{p_1}\cdots yx^{p_r} = \sum_{q\in\mathbb{N}}\sum_{\substack{p_0,\ldots,p_r\in\mathbb{N}\\ p_0+\cdots+p_r=q}} x^{p_0}yx^{p_1}\cdots yx^{p_r} = \sum_{q\in\mathbb{N}} S_{q,r} \end{aligned}$$

and therefore also

$$x^*(yx^*)^* = x^*\sum_{r\in\mathbb{N}}(yx^*)^r = \sum_{r\in\mathbb{N}} x^*(yx^*)^r = \sum_{r\in\mathbb{N}}\sum_{q\in\mathbb{N}} S_{q,r} = \sum_{q,r\in\mathbb{N}} S_{q,r}.$$

□

Lemma 1.8. *In a continuous semiring R we have $(x + y)^* = (x + yx^*y)^*(1 + yx^*)$.*

[7] For example, if $R = \mathcal{P}(A^*)$ and $x, y \in A$, then $S_{q,r}$ is the set of all words which consist of q occurrences of x and r occurrences of y.

Proof.

$$\begin{aligned}(x+y)^* &=^{\text{Lem. } 1.7} (x^*y)^*x^* = \sum_{j\geqslant 0}(x^*y)^j x^* = \sum_{j\geqslant 0}(x^*y)^{2j}x^* + \sum_{j\geqslant 0}(x^*y)^{2j+1}x^* \\ &= (x^*yx^*y)^*x^* + (x^*yx^*y)^*x^*yx^* = (x^*yx^*y)^*x^*(1+yx^*) \\ &=^{\text{Lem. } 1.7} (x+yx^*y)^*(1+yx^*).\end{aligned}$$

□

We have seen that $R^{I\times I}$ is a continuous semiring. If J is another finite set, then applying Proposition 1.8 again shows that $(R^{I\times I})^{J\times J}$ is a continuous semiring as well. An element M in $(R^{I\times I})^{J\times J}$ is a matrix of matrices but can be considered a matrix of elements of R as the following proposition shows:

Proposition 1.10. *Let R be a continuous semiring, I, J finite sets. Then the continuous semirings $(R^{I\times I})^{J\times J}$ and $R^{(I\times J)\times(I\times J)}$ are isomorphic.*

Proof. Define $\varphi : (R^{I\times I})^{J\times J} \to R^{(I\times J)\times(I\times J)}$ by $(\varphi(M))_{(i_1,j_1),(i_2,j_2)} = (M_{i_1,i_2})_{j_1,j_2}$. It is straightforward to verify that φ is an isomorphism. □

We will often decompose a matrix. Let $M \in R^{I\times I}$, $I_1 \uplus I_2 = I$. For $k, l \in \{1,2\}$ we write $M(I_k, I_l)$ for the matrix obtained from keeping the rows with indices in I_k and the columns with indices in I_l and deleting all others. In the context of a fixed partition $I = I_1 \uplus I_2$ we will often write $M_{k,l}$ as an abbreviation for $M(I_k, I_l)$. Proposition 1.10 above shows that we can identify $M \in R^{I\times I}$ with the matrix $\begin{pmatrix} M_{1,1} & M_{1,2} \\ M_{2,1} & M_{2,2}\end{pmatrix}$.

Theorem 1.5. *Let R be a continuous semiring, let $M \in R^{I\times I}$ and let $I = I_1 \uplus I_2$. Then*

$$\begin{aligned}(M^*)_{1,1} &= (M_{1,1} + M_{1,2}M_{2,2}^*M_{2,1})^* \\ (M^*)_{1,2} &= (M^*)_{1,1}M_{1,2}M_{2,2}^* \\ (M^*)_{2,2} &= (M_{2,2} + M_{2,1}M_{1,1}^*M_{1,2})^* \\ (M^*)_{2,1} &= (M^*)_{2,2}M_{2,1}M_{1,1}^*\end{aligned}$$

Proof. Let $M_1 = \begin{pmatrix} M_{1,1} & 0 \\ 0 & M_{2,2}\end{pmatrix}$ and $M_2 = \begin{pmatrix} 0 & M_{1,2} \\ M_{2,1} & 0\end{pmatrix}$, then

$$M^* = (M_1 + M_2)^* =^{\text{Lem. } 1.8} (M_1 + M_2M_1^*M_2)^*(1 + M_2M_1^*).$$

We have $M_1^* = \begin{pmatrix} M_{1,1}^* & 0 \\ 0 & M_{2,2}^*\end{pmatrix}$ and hence

$$M_1 + M_2M_1^*M_2 = \begin{pmatrix} M_{1,1} + M_{1,2}M_{2,2}^*M_{2,1} & 0 \\ 0 & M_{2,2} + M_{2,1}M_{1,1}^*M_{1,2}\end{pmatrix}.$$

Furthermore

$$1 + M_2M_1^* = \begin{pmatrix} 1 & M_{1,2}M_{2,2}^* \\ M_{2,1}M_{1,1}^* & 1\end{pmatrix}.$$

Therefore

$$
\begin{aligned}
M^* &= (M_1 + M_2 M_1^* M_2)^* (1 + M_2 M_1^*) \\
&= \begin{pmatrix} (M_{1,1} + M_{1,2} M_{2,2}^* M_{2,1})^* & 0 \\ 0 & (M_{2,2} + M_{2,1} M_{1,1}^* M_{1,2})^* \end{pmatrix} \begin{pmatrix} 1 & M_{1,2} M_{2,2}^* \\ M_{2,1} M_{1,1}^* & 1 \end{pmatrix} \\
&= \begin{pmatrix} (M_{1,1} + M_{1,2} M_{2,2}^* M_{2,1})^* & (M_{1,1} + M_{1,2} M_{1,1}^* M_{2,1})^* M_{1,2} M_{2,2}^* \\ (M_{2,2} + M_{2,1} M_{1,1}^* M_{1,2})^* M_{2,1} M_{1,1}^* & (M_{2,2} + M_{2,1} M_{1,1}^* M_{1,2})^* \end{pmatrix}
\end{aligned}
$$

□

Example 1.24. Let $I = \{1, 2, 3\}$ and $R = \mathbb{B}\langle\langle \{a, b\}^* \rangle\rangle$ and $M \in R^{I \times I}$ be

$$
M = \begin{pmatrix} a & b & 0 \\ 0 & 0 & b \\ b & 0 & 0 \end{pmatrix}
$$

Preparing for the application of Theorem 1.5, let $I_1 = \{1\}$ and $I_2 = \{2, 3\}$. We then have

$$
\begin{aligned}
(M_{1,1})^* &= \begin{pmatrix} a \end{pmatrix}^* = \begin{pmatrix} a^* \end{pmatrix} \\
(M_{2,2})^* &= \sum_{i \in \mathbb{N}} \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix}^i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}
\end{aligned}
$$

Now, by Theorem 1.5 we have

$$
\begin{aligned}
(M^*)_{1,1} &= \left(\begin{pmatrix} a \end{pmatrix} + \begin{pmatrix} b & 0 \end{pmatrix} \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ b \end{pmatrix} \right)^* = \begin{pmatrix} (a + b^3)^* \end{pmatrix} \\
(M^*)_{1,2} &= \begin{pmatrix} (a + b^3)^* \end{pmatrix} \begin{pmatrix} b & 0 \end{pmatrix} \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} (a + b^3)^* b & (a + b^3)^* b^2 \end{pmatrix} \\
(M^*)_{2,2} &= \left(\begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 \\ b \end{pmatrix} \begin{pmatrix} a^* \end{pmatrix} \begin{pmatrix} b & 0 \end{pmatrix} \right)^* = \begin{pmatrix} 0 & b \\ b a^* b & 0 \end{pmatrix}^*
\end{aligned}
$$

Now another application of Theorem 1.5 shows that the star of an anti-diagonal matrix is $\begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix}^* = \begin{pmatrix} (\alpha\beta)^* & (\alpha\beta)^* \alpha \\ (\beta\alpha)^* \beta & (\beta\alpha)^* \end{pmatrix}$ and hence

$$
= \begin{pmatrix} (b^2 a^* b)^* & (b^2 a^* b)^* b \\ (b a^* b^2)^* b a^* b & (b a^* b^2)^* \end{pmatrix}.
$$

Finally,

$$
(M^*)_{2,1} = \begin{pmatrix} (b^2 a^* b)^* & (b^2 a^* b)^* b \\ (b a^* b^2)^* b a^* b & (b a^* b^2)^* \end{pmatrix} \begin{pmatrix} 0 \\ b \end{pmatrix} \begin{pmatrix} a^* \end{pmatrix} = \begin{pmatrix} (b^2 a^* b)^* b^2 a^* \\ (b a^* b^2)^* b a^* \end{pmatrix}
$$

Summing up we have obtained:

$$
M^* = \begin{pmatrix} (a + b^3)^* & (a + b^3)^* b & (a + b^3)^* b^2 \\ (b^2 a^* b)^* b^2 a^* & (b^2 a^* b)^* & (b^2 a^* b)^* b \\ (b a^* b^2)^* b a^* & (b a^* b^2)^* b a^* b & (b a^* b^2)^* \end{pmatrix}
$$

1.6 Regular languages

1.6.1 Finite automata

Definition 1.20. Let R be a continuous semiring and $R' \subseteq R$. An R'-automaton in R is a tuple $\mathcal{A} = (I, M, S, P)$ where

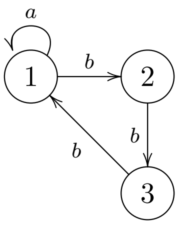
- I is a finite set of *states*,
- $M \in R'^{I \times I}$ is the *transition matrix*,
- $S \in R'^{1 \times I}$ is the *initial state vector*, and
- $P \in R'^{I \times 1}$ is the *final state vector*.

The behaviour $\|\mathcal{A}\| \in R$ of the R'-automaton $\mathcal{A}$ is defined as $\|\mathcal{A}\| = SM^*P$.

The intention of the above definition is that the entry $M_{i,j}$ of the matrix M determines the element of R' associated to the edge from state i to state j. Note that this definition permits more than one initial state and (as the traditional definition) also more than one final state. Moreover, to each starting and final state (each entering and exiting edge) we can associate an element of R' as well.

It is possible to extend the above definition of finite automata to an infinite set of states I. The results of Section 1.5 also hold in this more general setting. This allows to treat pushdown automata in an arbitrary continuous semiring A. However, we do not follow this direction here – in this course all automata will have a finite number of states.

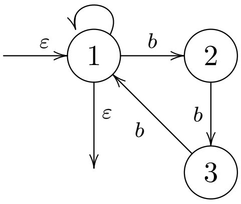
Example 1.25. The matrix M discussed in Example 1.24 corresponds to the following diagram



The entry $(M^*)_{i,j}$ is the language recognised by the paths from i to j. Indicating 1 as the only initial and the only final state can be done by setting

$$S = \begin{pmatrix} \varepsilon & 0 & 0 \end{pmatrix} \qquad \text{and} \qquad P = \begin{pmatrix} \varepsilon \\ 0 \\ 0 \end{pmatrix}.$$

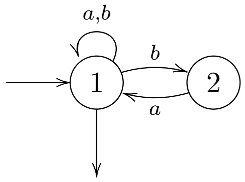
This is represented in the diagram as



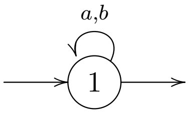
(but often we will omit the label ε on entering and exiting edges of a diagram). Then the behaviour of the automaton $\mathcal{A} = (\{1,2,3\}, M, S, P)$ in $\mathcal{P}(A^*)$ is

$$\|\mathcal{A}\| = SM^*P = (a + b^3)^*$$

Example 1.26. Let $A = \{a, b\}$ and consider the automaton

$\mathcal{A} \quad =$ 

in $\mathbb{B}\langle\!\langle A^*\rangle\!\rangle$. Then $M = \begin{pmatrix} a+b & b \\ a & 0 \end{pmatrix}$ and note that $\|\mathcal{A}\| = \begin{pmatrix}\varepsilon & 0\end{pmatrix} M^* \begin{pmatrix}\varepsilon \\ 0\end{pmatrix} = (M^*)_{1,1}$. By Theorem 1.5 we have $(M^*)_{1,1} = (M_{1,1} + M_{1,2}M_{2,2}^*M_{2,1})^* = (a + b + ba)^*$. But in $\mathbb{B}\langle\!\langle A^*\rangle\!\rangle$ we have $(a + b + ba)^* = (a + b)^*$ and hence $\|\mathcal{A}\| = \sum_{w\in A^*} w$. In particular the simpler automaton

$\mathcal{A}' \quad =$ 

has the same behaviour.

If we move to $\mathbb{N}^\infty\langle\!\langle A^*\rangle\!\rangle$ we have a more complicated behaviour. The coefficient of a word w turns out to be the *number of accepting paths of* w. Considering $\mathcal{A}$ in $\mathbb{N}^\infty\langle\!\langle A^*\rangle\!\rangle$ we claim that $\|\mathcal{A}\| = \sum_{w\in A^*} 2^{n_{ba}(w)}w$ where $n_{ba}(w)$ is the number of occurrences of ba in w. This can be shown as follows: first, as above we have $\|\mathcal{A}\| = (a + b + ba)^*$. Now, by Corollary 1.1, we have

$$\begin{aligned}(a+b+ba)^* &= \sum_{w\in A^*} \sum_{n=0}^{|w|} \sum_{\substack{(u_1,\dots,u_n)\in(A^*)^n \\ u_1\cdots u_n = w}} \prod_{i=1}^{n} (a+b+ba, u_i)\ w \\ &= \sum_{w\in A^*} \sum_{n=0}^{|w|} \sum_{\substack{u_1,\dots,u_n\in\{a,b,ba\}^* \\ u_1\cdots u_n = w}} 1\ w \\ &= \sum_{w\in A^*} 2^{n_{ba}(w)}\ w\end{aligned}$$

because for each occurrence of ba in w we can make the choice of whether to consider it as $u_i = ba$ or as $u_i = b$ and $u_{i+1} = a$ and these choices are independent.

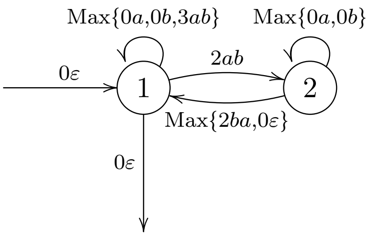
The above examples are classical automata; more precisely: for $A = \{a_1, \dots, a_n\}$ and $R' = \{\lambda_0\varepsilon + \lambda_1 a_1 + \cdots + \lambda_n a_n \mid \lambda_0, \dots, \lambda_n \in \{0, 1\}\}$ they are R'-automata in $R\langle\!\langle A^*\rangle\!\rangle$. But our more general context allows to put weights on the transitions (hence the terminology "weighted automata") by taking $\lambda_i \notin \{0, 1\}$. We thus define:

Definition 1.21. Let R be a continuous semiring, $A = \{a_1, \dots, a_n\}$ be an alphabet and $R' = \{\lambda_0\varepsilon + \lambda_1 a_1 + \cdots + \lambda_n a_n \mid \lambda_0, \dots, \lambda_n \in R\}$. An R'-automaton in $R\langle\!\langle A^*\rangle\!\rangle$ is called *weighted automaton* in $R\langle\!\langle A^*\rangle\!\rangle$.

When working in the min-plus semiring one can think of the weight as cost (since we want to minimise it) and, dually, in the max-plus semiring as gains (which we want to maximise). Under certain conditions, real-valued weights can be considered probabilities.

Example 1.27. Let $A = \{a, b\}$, $R = \langle\bar{\mathbb{N}}, \max, -\infty, +, 0\rangle$ where $\bar{\mathbb{N}} = \mathbb{N} \cup \{-\infty, +\infty\}$ with $+\infty +$

$(-\infty) = -\infty$. Consider the following automaton $\mathcal{A}$ in $R\langle\langle A^*\rangle\rangle$:



The above automaton has 1 as only initial and only final state: it reads a word of a's and b's looking out for the subword ab. The "greedy" strategy would be to take the edge $3ab$ when an occurrence of ab is encountered. However, this does not maximise the gain. If the current ab is matched by a ba later without any ab's in between, then it pays out to move to state 2, thus collecting gains of 4 instead of only 3 (since a ba in state 1 is useless).

We have

$$M = \begin{pmatrix} \mathrm{Max}\{0a, 0b, 3ab\} & 2ab \\ \mathrm{Max}\{2ba, 0\varepsilon\} & \mathrm{Max}\{0a, 0b\} \end{pmatrix}$$

and $\|\mathcal{A}\| = (M^*)_{1,1} = \mathrm{Max}\{M_{1,1}, M_{1,2} \oplus M^*_{2,2} \oplus M_{2,1}\}$ where we write $\oplus$ for the addition in $R\langle\langle A^*\rangle\rangle$ which is defined via the Cauchy formula. Concerning $M^*_{2,2}$ first note that, by Corollary 1.1,

$$(\mathrm{Max}\{0a, 0b\}^*, w) = \underset{n=0,\ldots,|w|}{\mathrm{Max}} \ \underset{(u_1,\ldots,u_n)\in(A^*)^n}{\mathrm{Max}} \sum_{i=1}^{n} (\mathrm{Max}\{0a, 0b\}, u_i) = 0$$

In $R\langle\langle A^*\rangle\rangle$, the sum of three formal series is defined via the Cauchy formula as

$$(r_1 \oplus r_2 \oplus r_3, w) = \underset{\substack{u_1,u_2,u_3\in A^* \\ u_1u_2u_3=w}}{\mathrm{Max}} \{(r_1, u_1) + (r_2, u_2) + (r_3, u_3)\}.$$

Thus we obtain

$$\begin{aligned}
(r, w) &:= (M_{1,2} \oplus M^*_{2,2} \oplus M_{2,1}, w) \\
&= \underset{\substack{u_1,u_2,u_3\in A^* \\ u_1u_2u_3=w}}{\mathrm{Max}} \{(2ab, u_1) + 0 + (\max\{2ba, 0\varepsilon\}, u_3)\} \\
&= \underset{\substack{u_1,u_2,u_3\in A^* \\ u_1u_2u_3=w}}{\mathrm{Max}} \{(2ab, u_1) + (2ba, u_3), (2ab, u_1) + (0\varepsilon, u_3)\} \\
&= \begin{cases} 4 & \text{if } w \in abA^*ba \\ 2 & \text{if } w \in abA^* \backslash (abA^*ba) \\ -\infty & \text{otherwise} \end{cases}
\end{aligned}$$

and

$$\begin{aligned}
(s, w) &:= (\mathrm{Max}\{M_{1,1}, M_{1,2} \oplus M^*_{2,2} \oplus M_{2,1}\}, w) \\
&= (\mathrm{Max}\{0a, 0b, 3ab, r\}, w) \\
&= \begin{cases} 4 & \text{if } w \in abA^*ba \\ 3 & \text{if } w = ab \\ 2 & \text{if } w \in abA^* \backslash ((abA^*ba) \cup ab) = abA^+ \backslash abA^*ba \\ 0 & \text{if } w = a \text{ or } w = b \\ -\infty & \text{otherwise} \end{cases}
\end{aligned}$$

By applying Proposition 1.6 in $R\langle\langle A^*\rangle\rangle$ we obtain

$$(s^*, w) = \operatorname*{Max}_{\substack{u_1,\dots,u_n \in A^* \\ u_1\cdots u_n = w}} \{(s, u_1) + \cdots + (s, u_n)\}$$

Now, whenever $u_i \in abA^+ \backslash abA^*ba$ then the decomposition $u_i = abu_{i+1}\cdots u_{i+k}$ leads to a higher coefficient, so the line with coefficient 2 is never used in the computation of a maximum. Similarly, if $u_i = abvba$ with $v \in A^*$ s.t. v contains an a, then $n_{ab}(u_i) \geqslant 2$ (where $n_{ab}(u)$ is the number of occurrences of ab in u). So the line with coefficient 4 is only used on $u_i \in abb^*ba$ in the computation of a maximum. Moreover, note that if $u_i \in bA^*$ in a maximal decomposition, then $u_i = b$ for otherwise $-\infty$ must appear as summand. Furthermore, the line with coefficient 4 is never used on a $u_i \in abb^*ba$ if $u_{i+1} = b$ in a maximal decomposition, for then $n_{ab}(u_iu_{i+1}) \geqslant 2$ and using the line with coefficient 3 would give a better decomposition.

Therefore, the decomposition of w which maximises (s^*, w) is $w = u_0v_1u_1\cdots v_nu_n$ where $v_1, \dots, v_n \in abb^*ba$, $u_1, \dots, u_n \in aA^* \cup \{\varepsilon\}$, $u_0 \in A^*$ and n is maximal. Then we obtain

$$(s^*, w) = 4n + 3n_{ab}(u_0\cdots u_n)$$

where n is obtained from that decomposition of w.

Note that the notion of finite automaton as defined above does not make any assumption on the continuous semiring R, in particular it is not required that $R = S\langle\langle A^*\rangle\rangle$ for some continuous semiring S and some alphabet A. The generality of this notion of automaton (just as that of the notion of algebraic system) thus goes beyond what is required for formal language theory, as the following example shows.

Example 1.28. Let $G(V, E)$ be a graph. Let $R = \langle \mathbb{N}^\infty, \min, \infty, +, 0\rangle$ and let $w : E \to R$. Then $M = (V, E, w)$ is an R-weighted graph and the weight of a path is the sum of the weights of its edges. Let $i, j \in V$, let $S \in \{\infty, 0\}^{1\times V}$ be ∞ everywhere except at position i and $P \in \{\infty, 0\}^{V\times 1}$ be ∞ everywhere except at position j. Then $\mathcal{A} = (V, M, S, P)$ is an $\{0, 1, \infty\}$-automaton in R (but not in $R\langle\langle A^*\rangle\rangle$ (!)) and we have $\|\mathcal{A}\| = S(M^*)P = (M^*)_{i,j} = \min\{\sum_{i=1}^n w(e_i) \mid (e_1, \dots, e_n) \in P(i, j)\}$, i.e., the automaton computes the length of the shortest path from i to j.

1.6.2 Kleene's theorem

In the context of the elementary theory of formal languages, Kleene's theorem states that regular expressions define the same class of languages as finite automata: the regular languages. In this section we prove an analogous result in our more general context.

Definition 1.22. Let R be a continuous semiring and $R' \subseteq R$. Then the *automatic closure of* R' is defined as $\mathfrak{Aut}(R') = \{x \in R \mid \text{there is } R'\text{-automaton } \mathcal{A} \text{ s.t. } x = \|\mathcal{A}\|\}$.

Definition 1.23. Let R be a continuous semiring and $R' \subseteq R$. The *rational closure* $\mathfrak{Rat}(R')$ of R' is the smallest set which contains 0, 1, all $x \in R'$ and is closed under $+$, $\cdot$ and $*$.

Expressions built from $R', 0, 1$ as well as $+, \cdot, *$ are also called R'-rational expressions. For $R = \mathcal{P}(A^*)$ and $R' = \{\{w\} \mid w \in A^*\}$, the R'-rational-expressions are just the regular expressions well-known from theoretical computer science. Kleene's theorem will then be formulated as follows: for all $\{0, 1\} \subseteq R' \subseteq R$ we have $\mathfrak{Rat}(R') = \mathfrak{Aut}(R')$. The proof of this result will occupy the remainder of this section. Before we start with the actual proof, it will be helpful to define a notion of normal form for automata.

Definition 1.24. An R'-automaton $\mathcal{A} = (I, M, S, P)$ is called *normalised* if

1. There is an $i_0 \in I$ s.t. $S_{i_0} = 1$ and $S_i = 0$ for $i_0 \neq i$.
2. There is an $i_\text{f} \in I$, $i_\text{f} \neq i_0$ s.t. $P_{i_\text{f}} = 1$ and $P_i = 0$ for $i \neq i_\text{f}$.
3. $M_{i,i_0} = M_{i_\text{f},i} = 0$ for all $i \in I$.

Lemma 1.9. *For every R'-automaton $\mathcal{A}$ there is a normalised $R' \cup \{0,1\}$-automaton $\mathcal{A}'$ s.t. $\|\mathcal{A}'\| = \|\mathcal{A}\|$.*

Proof. Let $\mathcal{A} = (I, M, S, P)$ be an R'-automaton, let $i_0, i_\text{f} \notin I$ be new states and define $\mathcal{A}' = (\{i_0, i_\text{f}\} \cup I, M', S', P')$ where

$$S' = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, \quad M' = \begin{pmatrix} 0 & 0 & S \\ 0 & 0 & 0 \\ 0 & P & M \end{pmatrix}, \quad P' = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

We have

$$\|\mathcal{A}'\| = S'M'^*P' = (M'^*)_{i_0,i_\text{f}}.$$

Using the partition $I' = I_1 \uplus I_2 = \{i_0, i_\text{f}\} \cup I$ for Theorem 1.5 we obtain

$$\begin{aligned}(M'^*)_{1,1} &= (M_{1,1} + M_{1,2}M_{2,2}^*M_{2,1})^* = \left(\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} S \\ 0 \end{pmatrix} M^* \begin{pmatrix} 0 & P \end{pmatrix} \right)^* \\ &= \begin{pmatrix} 0 & SM^*P \\ 0 & 0 \end{pmatrix}^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & SM^*P \\ 0 & 0 \end{pmatrix} + \sum_{i \geqslant 2} \begin{pmatrix} 0 & SM^*P \\ 0 & 0 \end{pmatrix}^i = \begin{pmatrix} 1 & SM^*P \\ 0 & 1 \end{pmatrix}.\end{aligned}$$

Therefore $(M'^*)_{i_0,i_f} = SM^*P$ and we obtain $\|\mathcal{A}'\| = SM^*P = \|\mathcal{A}\|$. □

Lemma 1.10. *Let R be a continuous semiring, $\{0,1\} \subseteq R' \subseteq R$. Then $\mathfrak{Rat}(R') \subseteq \mathfrak{Aut}(R')$.*

Proof. For any $x \in R'$ we have $\|(\{1\}, \begin{pmatrix} 0 \end{pmatrix}, \begin{pmatrix} 1 \end{pmatrix}, \begin{pmatrix} x \end{pmatrix})\| = \begin{pmatrix} 1 \end{pmatrix} \begin{pmatrix} 0 \end{pmatrix}^* \begin{pmatrix} x \end{pmatrix} = x$.

For closure under addition, let $\mathcal{A}_1 = (I_1, M_1, S_1, P_1)$ and $\mathcal{A}_2 = (I_2, M_2, S_2, P_2)$ be R'-automata with $I_1 \cup I_2 = \varnothing$ and define $\mathcal{A} = (I_1 \cup I_2, M, S, P)$ by

$$M = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix}, \quad S = \begin{pmatrix} S_1 & S_2 \end{pmatrix}, \quad P = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}.$$

Then, clearly $\mathcal{A}$ is an R'-automaton and we have $\|\mathcal{A}\| = SM^*P = S_1M_1^*P_1 + S_2M_2^*P_2 = \|\mathcal{A}_1\| + \|\mathcal{A}_2\|$.

For closure under product, let $\mathcal{A}_1$ and $\mathcal{A}_2$ be as above. By Lemma 1.9 we can assume w.l.o.g. that $\mathcal{A}_1$ and $\mathcal{A}_2$ are normalised. Define $\mathcal{A} = (I_1 \cup I_2, M, S, P)$ by

$$M = \begin{pmatrix} M_1 & P_1S_2 \\ 0 & M_2 \end{pmatrix}, \quad S = \begin{pmatrix} S_1 & 0 \end{pmatrix}, \quad P = \begin{pmatrix} 0 \\ P_2 \end{pmatrix}.$$

Since $\mathcal{A}_1$ and $\mathcal{A}_2$ are normalised, $P_1S_2 \in \{0,1\}^{I_1 \times I_2} \subseteq R'^{I_1 \times I_2}$ and hence $\mathcal{A}$ is an R'-automaton. Using the partition $I_1 \uplus I_2$ for Theorem 1.5 we obtain

$$(M^*)_{1,2} = (M_{1,1} + M_{1,2}M_{2,2}^*M_{2,1})^*M_{1,2}M_{2,2}^* = (M_1 + P_1S_1M_2^*0)^*P_1S_1M_2^* = M_1^*P_1S_2M_2^*$$

and hence

$$\|\mathcal{A}\| = SM^*P = S_1(M^*)_{1,2}P_2 = S_1M_1^*P_1S_2M_2^*P_2 = \|\mathcal{A}_1\|\|\mathcal{A}_2\|.$$

For closure under star, let $\mathcal{A} = (I, M, S, P)$ be an R' automaton. Let $q_0 \notin I$ be a new state and define $\mathcal{A}' = (\{q_0\} \cup I, M', S', P')$ by

$$M' = \begin{pmatrix} 0 & S \\ P & M \end{pmatrix}, \quad S' = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad P' = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Using the partition $\{q_0\} \uplus I$ for Theorem 1.5 we obtain

$$\|\mathcal{A}'\| = S'M'^*P' = (M'^*)_{1,1} = (0 + SM^*P)^* = \|\mathcal{A}\|^*$$

□

Lemma 1.11. *Let R be a continuous semiring, $R' \subseteq R$. Then $\mathfrak{Aut}(R') \subseteq \mathfrak{Rat}(R')$.*

Proof. We first show that $M \in R'^{I \times I}$ implies $M^* \in \mathfrak{Rat}(R')^{I \times I}$ by induction on $|I|$. If $|I| = 1$, then clearly $\begin{pmatrix} m \end{pmatrix}^* = \begin{pmatrix} m^* \end{pmatrix} \in \mathfrak{Rat}(R')^{I \times I}$. For the induction step, consider any partition $I = I_1 \uplus I_2$ into two non-empty parts. Then, by Theorem 1.5, $(M^*)_{1,1} = (M_{1,1} + M_{1,2}M_{2,2}^*M_{2,1})^*$. By induction hypothesis $M_{2,2}^* \in \mathfrak{Rat}(R')^{I_2 \times I_2}$, hence $M_{1,1} + M_{1,2}M_{2,2}^*M_{2,1} \in \mathfrak{Rat}(R')^{I_1 \times I_1}$ and therefore, again by induction hypothesis, $(M^*)_{1,1} \in \mathfrak{Rat}(\mathfrak{Rat}(R'))^{I_1 \times I_1} = \mathfrak{Rat}(R')^{I_1 \times I_1}$. For the $(M^*)_{i,j}$ with $(i,j) \neq (1,1)$ proceed analogously. We obtain $M^* \in \mathfrak{Rat}(R')^{I \times I}$.

Let $\mathcal{A} = (I, M, S, P)$ be an R'-automaton, then $\|\mathcal{A}\| = SM^*P = \sum_{i,j \in I} S_i(M^*)_{i,j}P_j$ and since $(M^*)_{i,j} \in \mathfrak{Rat}(R')$, also $\|\mathcal{A}\| \in \mathfrak{Rat}(R')$. □

We have thus proved Kleene's theorem:

Theorem 1.6. *Let R be a continuous semiring and $\{0, 1\} \subseteq R' \subseteq R$. Then $\mathfrak{Aut}(R') = \mathfrak{Rat}(R')$.*

This result shows that two structurally quite different specification formalisms, rational expressions on the one hand and automata on the other hand, define the same closure operator. A situation like this is evidence that we are dealing with an important closure operator. It is a strong generalisation of the usual Kleene theorem from the elementary theory of formal languages:

Corollary 1.2. *Let $R = \mathcal{P}(A^*)$ and $R' = \{\varnothing, \{\varepsilon\}\} \cup \{\{x\} \mid x \in A\}$. Then $\mathfrak{Aut}(R') = \mathfrak{Rat}(R')$. An $L \in \mathfrak{Aut}(R')$ is called regular language.*

1.6.3 Linear systems

So far we have seen that regular expressions and automata are equivalent in the sense that they define the same closure operator. As in the elementary theory of formal languages, a certain kind of grammars (or in our setting: algebraic systems) is equivalent to these as well.

Definition 1.25. Let R be a continuous semiring, $R' \subseteq R$ and $Y = \{y_1, \ldots, y_n\}$. An R'-algebraic system $Y = p(Y)$ is called *linear* if there are $m_{i,j}, q_i \in R'$ s.t.

$$p_i(y_1, \ldots, y_n) = m_{i,1}y_1 + \ldots + m_{i,n}y_n + q_i, \qquad \text{for } 1 \leqslant i \leqslant n.$$

Thus linear systems generalise right-linear grammars (of which we know that they generate exactly the regular languages). An R'-linear system is often written as $Y = MY + Q$ where $M = (m_{i,j})_{1 \leqslant i,j \leqslant n}$ and $Q = \begin{pmatrix} q_1 \\ \vdots \\ q_n \end{pmatrix}$.

Definition 1.26. Let R be a continuous semiring and $R' \subseteq R$. Then the linear closure $\mathfrak{Lin}(R')$ of R' is defined as $\mathfrak{Lin}(R') = \{x \in R \mid x$ is component of the least solution of an R'-linear system$\}$.

Proposition 1.11. *Let R be a continuous semiring, $R' \subseteq R$ and let $Y = MY + Q$ be an R'-linear system. Then M^*Q is its least solution.*

Proof. Write $p(Y) = MY + Q$. We show $p^n(0) = \sum_{0 \leqslant i < n} M^i Q$ by induction on n. For $n = 0$, $p^0(0) = 0$. For the induction step we have $p^{n+1}(0) = p(\sum_{0 \leqslant i < n} M^i Q) = M \sum_{0 \leqslant i < n} M^i Q + Q = \sum_{0 \leqslant i < n+1} M^i Q$.
Now, the least solution of $Y = p(Y)$ is

$$\sup\{p^n(0) \mid n \in \mathbb{N}\} = \sup\{\sum_{0 \leqslant i < n} M^i Q \mid n \in \mathbb{N}\} = \sup\{\sum_{0 \leqslant i < n} M^i \mid n \in \mathbb{N}\} Q = M^* Q.$$

□

Theorem 1.7. *Let R be a continuous semiring and $\{0, 1\} \subseteq R' \subseteq R$. Then $\mathfrak{Lin}(R') = \mathfrak{Aut}(R')$.*

Proof. For the left-to-right direction let $x \in \mathfrak{Lin}(R')$. Then x is i-th component of the least solution of an R'-linear system $Y = MY + Q$ where $Y = \{y_1, \ldots, y_n\}$. From Proposition 1.11 we know that this solution is M^*Q. Now let $S \in \{0, 1\}^{1 \times n}$ be 0 everywhere except for the i-th component. Then $x = SM^*Q$. This is the behaviour of the R'-automaton $(\{1, \ldots, n\}, S, M, Q)$.

For the right-to-left direction let $x \in \mathfrak{Aut}(R')$. By Lemma 1.9 we can assume that $x = \|\mathcal{A}\|$ for a normalised R'-automaton $\mathcal{A} = (I, S, M, P)$, i.e. there is some i s.t. all components of S are 0 except the i-th which is 1. Therefore $x = (M^*P)_i$ which, by Proposition 1.11, is the i-th component of the least solution of the R'-linear system $Y = MY + P$, hence $x \in \mathfrak{Lin}(R')$. □

Chapter 2

Algebraic automata theory

In this chapter we will focus on regular languages and deterministic finite automata. It will turn out that there is an intimate relationship between finite automata and finite monoids. This relationship goes so far that one can establish a one-to-one correspondence between certain classes, so-called varieties, of finite monoids and classes of regular languages. We will see one of the most important such correspondences, Schützenberger's characterisation of the star-free languages as the languages recognisable by aperiodic monoids.

In order to arrive at this characterisation and to motivate the notions underlying this correspondence, it pays out to first study the Myhill-Nerode theorem. This result is an algebraic characterisation of the class of regular languages and is based on the construction of the minimal deterministic finite automaton of a regular language.

2.1 The Myhill-Nerode theorem

2.1.1 The right-congruence of a DFA

Let M be a monoid and Q be a set. A right monoid action of M on Q is a function $\cdot : Q \times M \to Q$ s.t.

$$\begin{aligned} q \cdot e &= q, \text{ and} \\ q \cdot (m_1 m_2) &= (q \cdot m_1) \cdot m_2. \end{aligned}$$

A right monoid action is best thought of as each $m \in M$ inducing a function from Q to Q with e inducing the identity function and composition in the monoid M being composition of functions. Let A be an alphabet and $M = A^*$ the freely generated monoid. Then a monoid action $\cdot$ of M is uniquely determined by the action of the generators $x \in A$. In this chapter, we will notate a deterministic finite automaton (DFA) as a tuple $D = \langle Q, A, \cdot, q_0, F \rangle$ where Q, A, q_0 and F have the usual meaning and $\cdot$ is a right monoid action of the freely generated monoid A^* on Q. The action of the generators A on Q is represented as a diagram as usual for DFAs. While – at this point – this is only a change in notation, we will later consider monoids different from A^* and their action on Q.

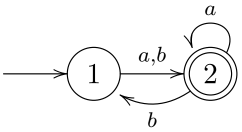
An important property of finite state automata which recognise infinite languages is that there are differents words which are indistinguishable to the automaton since they lead to the same state. This is, for example, used in the proof of the pumping lemma where we observe that a word w with $|w| > |Q|$ induces a path which must contain one state twice, say $q_0 \cdot u_1$ and

$q_0 \cdot u_1 u_2$. This loop can then be "pumped" because the words u_1 and $u_1 u_2$, since they are leading to the same state, are indistinguishable to the automaton. Moreover, in a deterministic finite automaton, every word induces a unique path of states. We can hence make this notion of indistinguishability precise as follows:

Definition 2.1. Let $D = \langle Q, A, \cdot, q_0, F\rangle$ be a DFA. Define the relation $\sim_D$ on A^* by $w_1 \sim_D w_2$ iff $q_0 \cdot w_1 = q_0 \cdot w_2$.

The relation $\sim_D$ is clearly an equivalence relation. Moreover, $\sim_D$ is right-congruent, i.e., whenever $w_1 \sim_D w_2$, then for all $v \in A^*$: $w_1 v \sim_D w_2 v$ because $q_0 \cdot w_1 = q_0 \cdot w_2$ implies $q_0 \cdot w_1 v = q_0 \cdot w_2 v$ for all $v \in A^*$. Therefore $\sim_D$ is also called the *right-congruence of D*. Note that $\sim_D$ is not a congruence relation, see the following example:

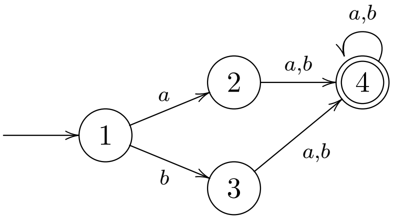
Example 2.1. Let D be the following DFA:



Then $a \sim_D b$ and therefore $aw \sim_D bw$ for all $w \in A^*$. However, $\sim_D$ is not left-congruent, in particular we have $aa \not\sim_D ab$.

Moreover, since $\sim_D$ is right congruent, $w_1 \sim_D w_2$ implies $w_1 v \in L(D)$ iff $w_2 v \in L(D)$ for all $v \in A^*$. However, the converse is not true. Consider the following example.

Example 2.2. Let D be the following DFA:



Here $av \in L(D)$ iff $v \in \{a,b\}^+$ iff $bv \in L(D)$ but $a \not\sim_D b$ because $1 \cdot a = 2$ and $1 \cdot b = 3$.

Definition 2.2. Let $(X, \sim)$ be an equivalence relation. The index of $\sim$ is the number of equivalence classes of $\sim$.

Note that it is an immediate consequence of the definition of $\sim_D$ that the index of $\sim_D$ is at most $|Q|$. We also write $[q]_{\sim_D}$ or just $[q]$ for the class induced by a state, i.e., $[q] = \{w \in A^* \mid q_0 \cdot w = q\}$. Assume that D is a DFA where every state is accessible, i.e., for all $q \in Q$ there is a $w_q \in A^*$ s.t. $q_0 \cdot w_q = q$. Then, for $q_1, q_2 \in Q$ with $q_1 \neq q_2$ we have $w_{q_1} \not\sim_D w_{q_2}$. Therefore, in a DFA where all states are accessible, the index of $\sim_D$ is exactly $|Q|$.

We now proceed to study what the relation $\sim_D$ tells us about the automaton D. To that aim, the notion of isomorphism of automata will turn out to be crucial.

Definition 2.3. Let $D = \langle Q, A, \cdot, q_0, F\rangle$ and $D' = \langle Q', A, \cdot, q_0', F'\rangle$ be DFAs. Then D and D' are called *isomorphic*, written as $D \simeq D'$, if there is a bijection $\varphi : Q \to Q'$ s.t.

1. $\varphi(q_0) = q_0'$,
2. $\varphi(F) = F'$, and

3. for all $q \in Q$ and $x \in A$: $\varphi(q) \cdot x = \varphi(q \cdot x)$.

It is not difficult to show that two isomorphic automata accept the same language[1]. The implication in the other direction is of course not true: there are non-isomorphic automata which accept the same language. However:

Lemma 2.1. *Let $D = \langle Q, A, \cdot, q_0, F\rangle$ and $D' = \langle Q', A, \cdot, q_0', F'\rangle$ be DFAs where every state is accessible. If $L(D) = L(D')$ and $\sim_D = \sim_{D'}$, then $D \simeq D'$.*

Proof. Let us first observe that, for all $v, w \in A^*$:

$$q_0 \cdot v = q_0 \cdot w \text{ iff } v \sim_D w \text{ iff } v \sim_{D'} w \text{ iff } q_0' \cdot v = q_0' \cdot w. \tag{*}$$

Let $q \in Q$. Since q is accessible, every $q \in Q$ can be written as $q_0 \cdot w = q$ for some $w \in A^*$. We define $\varphi : Q \to Q', q_0 \cdot w \mapsto q_0' \cdot w$. Reading (*) from left to right shows that φ is well-defined. Reading (*) from right to left shows that φ is injective. The function φ is also surjective: let $q' \in Q'$. Then, since q' is accessible, there is w s.t. $q_0' \cdot w = q'$. Then $\varphi(q_0 \cdot w) = q_0' \cdot w = q'$.
For 1. we have $\varphi(q_0) = \varphi(q_0 \cdot \varepsilon) = q_0' \cdot \varepsilon = q_0'$. For 3., let $w \in A^*$ s.t. $q_0 \cdot w = q$. Then we have

$$\begin{aligned}\varphi(q) \cdot x &= \varphi(q_0 \cdot w) \cdot x = (q_0' \cdot w) \cdot x = q_0' \cdot wx, \text{ and}\\ \varphi(q \cdot x) &= \varphi((q_0 \cdot w) \cdot x) = \varphi(q_0 \cdot wx) = q_0' \cdot wx.\end{aligned}$$

For 2. note that, for every $q \in Q$, we have

$$\begin{aligned}[\varphi(q)]_{\sim_{D'}} &= \{w \in A^* \mid q_0' \cdot w = \varphi(q)\}\\ &= \{w \in A^* \mid q_0' \cdot w = \varphi(q_0 \cdot v)\}\end{aligned}$$

where $v \in A^*$ s.t. $q_0 \cdot v = q$ and thus

$$\begin{aligned}[\varphi(q)]_{\sim_{D'}} &= \{w \in A^* \mid q_0' \cdot w = q_0' \cdot v\}\\ &= \{w \in A^* \mid w \sim_{D'} v\}\\ &= \{w \in A^* \mid w \sim_D v\}\\ &= \{w \in A^* \mid q_0 \cdot w = q_0 \cdot v = q\}\\ &= [q]_{\sim_D}.\end{aligned}$$

Therefore

$$L(D) = \bigcup_{q \in F} [q]_{\sim_D} = \bigcup_{q \in F} [\varphi(q)]_{\sim_{D'}} = \bigcup_{q' \in \varphi(F)} [q']_{\sim_{D'}}$$

and since $L(D') = \bigcup_{q' \in F'} [q']_{\sim_{D'}}$ we have $F' = \varphi(F)$. □

2.1.2 The right-congruence of a language

We have seen that, for a DFA D which accepts the language L, the relation $\sim_D$ satisfies: whenever $w_1 \sim_D w_2$, then for all $v \in A^*$: $w_1 v \in L$ iff $w_2 v \in L$. Instead of starting with the definition of $\sim_D$ based on D and deriving this property, we can start with the language L and define the coarsest equivalence relation which satisfies this property.

Definition 2.4. Let $L \subseteq A^*$. Define the relation $\sim_L$ on A^* by $w_1 \sim_L w_2$ iff for all $v \in A^*$: $w_1 v \in L$ iff $w_2 v \in L$.

[1] Do it as exercise!

The relation $\sim_L$ contains important information about L. In particular it allows a characterisation of the regular languages, the Myhill-Nerode theorem: $L \subseteq A^*$ is regular iff index($\sim_L$) is finite. This theorem is the main result of Section 2.1. A first important property of $\sim_L$ is that it is also a right-congruence, i.e., if $w_1 \sim_L w_2$ and $u \in A^*$, then $w_1 u \sim_L w_2 u$, we hence also speak of the *right-congruence of* L. Another important property is:

Lemma 2.2. *Let D be a DFA, then $w_1 \sim_D w_2$ implies $w_1 \sim_{L(D)} w_2$.*

Proof. Let $D = \langle Q, A, \cdot, q_0, F \rangle$. If $w_1 \sim_D w_2$, then $q_0 \cdot w_1 = q_0 \cdot w_2$ and therefore, for all $v \in A^*$: $q_0 \cdot w_1 v \in F$ iff $q_0 \cdot w_2 v \in F$, i.e. for all $v \in A^*$: $w_1 v \in L(D)$ iff $w_2 v \in L(D)$, i.e., $w_1 \sim_{L(D)} w_2$. □

Definition 2.5. Let $(X, \sim_1)$ and $(X, \sim_2)$ be two equivalence relations. We say that $\sim_2$ is a *refinement* of $\sim_1$ if $\forall x, y \in X : x \sim_2 y \Rightarrow x \sim_1 y$.

If $\sim_2$ is a refinement of $\sim_1$ we also say that $\sim_2$ is *finer* than $\sim_1$ and that $\sim_1$ is *coarser* than $\sim_2$. If $\sim_2$ is a refinement of $\sim_1$, then each $\sim_1$-equivalence class is a disjoint union of $\sim_2$-equivalence classes. Consequently, the index of $\sim_1$ is at most that of $\sim_2$. So the above Lemma shows that $\sim_D$ is finer than $\sim_{L(D)}$ for all DFAs D.

Example 2.3. Continuing Example 2.2, note that, since all states are accessible, the equivalence classes of $\sim_D$ are [1], [2], [3] and [4]. On the other hand, the language accepted by D is $L = \{w \in A^* \mid |w| \geqslant 2\}$. Therefore the equivalence classes of $\sim_L$ are

$$\{w \in A^* \mid |w| = 0\} = \{\varepsilon\}$$
$$\{w \in A^* \mid |w| = 1\} = \{a, b\}$$
$$\{w \in A^* \mid |w| \geqslant 2\} = L$$

And we have $[1] = \{\varepsilon\}$, $[2] \cup [3] = \{a, b\}$ and $[4] = L$.

A priori, it is not clear how to compute the equivalence classes of $\sim_L$ in general. However, $\sim_L$ is closely related to the left-quotients of L which can be computed systematically in a straightforward way.

Definition 2.6. Let $L \subseteq A^*$ and $v \in A^*$. We define the *left-quotient* $v^{-1}L = \{w \in A^* \mid vw \in L\}$.

The left-quotient $v^{-1}L$ can be thought of as the set of all $w \in A^*$ s.t., if we have already read v, reading w will lead us into L. Note that $w \in v^{-1}L$ iff $vw \in L$, so, in particular, $v \in L$ iff $\varepsilon \in v^{-1}L$.

Lemma 2.3. *Let $L \subseteq A^*$ and $v_1, v_2 \in A^*$. Then $v_1^{-1}L = v_2^{-1}L$ iff $v_1 \sim_L v_2$.*

Proof. $v_1^{-1}L = v_2^{-1}L$ iff $\{w \in A^* \mid v_1 w \in L\} = \{w \in A^* \mid v_2 w \in L\}$ iff for all $w \in A^*$: $v_1 w \in L \Leftrightarrow v_2 w \in L$ iff $v_1 \sim_L v_2$. □

In order to compute with left-quotients, the following observations are helpful.

Lemma 2.4. *Let $L, L_1, L_2 \subseteq A^*$ and $u, v \in A^*$, then*

1. $v^{-1}(L_1 \cup L_2) = v^{-1}L_1 \cup v^{-1}L_2$, *and*

2. $(uv)^{-1}L = (v^{-1}(u^{-1}L))$.

Proof. For 1. we have

$$\begin{aligned} v^{-1}(L_1 \cup L_2) &= \{w \in A^* \mid vw \in L_1 \cup L_2\} \\ &= \{w \in A^* \mid vw \in L_1\} \cup \{w \in A^* \mid vw \in L_2\} \\ &= v^{-1}L_1 \cup v^{-1}L_2. \end{aligned}$$

For 2., note first that $u^{-1}L = \{w_0 \in A^* \mid uw_0 \in L\}$. Then we have

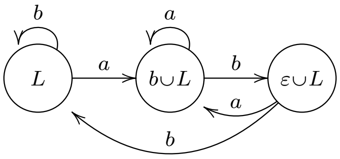
$$\begin{aligned} v^{-1}(u^{-1}L) &= \{w \in A^* \mid vw \in u^{-1}L\} \\ &= \{w \in A^* \mid vw = w_0, uw_0 \in L\} \\ &= \{w \in A^* \mid uvw \in L\} \\ &= (uv)^{-1}L. \end{aligned}$$

□

Example 2.4. Let $A = \{a, b\}$ and $L = A^*ab$. We compute the left-quotients of L:

$$\begin{aligned} \varepsilon^{-1}L &= L \\ a^{-1}L &= b \cup L \\ b^{-1}L &= L \\ a^{-1}(b \cup L) &= a^{-1}b \cup a^{-1}L = a^{-1}L = b \cup L \\ b^{-1}(b \cup L) &= b^{-1}b \cup b^{-1}L = \varepsilon \cup L \\ a^{-1}(\varepsilon \cup L) &= a^{-1}\varepsilon \cup a^{-1}L = a^{-1}L = b \cup L \\ b^{-1}(\varepsilon \cup L) &= b^{-1}\varepsilon \cup b^{-1}L = b^{-1}L = L \end{aligned}$$

Note how every line except the first of the above calculation contributes an edge to the below diagram. A line which contains a left-quotient that is new w.r.t. the lines so far creates a new vertex.



The above list is saturated in the sense that it contains $\varepsilon^{-1}L = L$ and for every left-quotient $v^{-1}L$ it contains, it also contains $a^{-1}(v^{-1}L)$ and $b^{-1}(v^{-1}L)$. A set which is saturated in this sense contains all left-quotients, since every $w \in A^*$ can be written as $w = x_1 \cdots x_n$ with $x_i \in A$ and thus $w^{-1}L = x_n^{-1}(\cdots\cdots x_2^{-1}(x_1^{-1}L))) \cdots)$ occurs in the set.

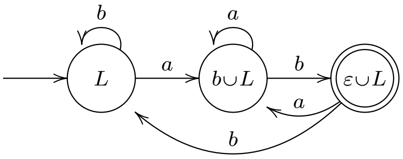
Note that L from the above example has a finite number of left-quotients, or, equivalently: a finite number of $\sim_L$-equivalence classes.

2.1.3 The minimal DFA of a regular language

Definition 2.7. Let $L \subseteq A^*$ s.t. the index of $\sim_L$ is finite. Define the *canonical automaton of* L as $D(L) = \langle Q, A, \cdot, q_0, F\rangle$ where $Q = \{w^{-1}L \mid w \in A^*\}$, $q_0 = \varepsilon^{-1}L = L$, $F = \{w^{-1}L \mid w \in L\}$ and $w^{-1}L \cdot x = (wx)^{-1}L$.

Since $\sim_L$ is of finite index there are, by Lemma 2.3, only finitely many left-quotients and so Q is finite. Clearly $q_0 \in Q$ and $F \subseteq Q$. Also note that $F = \{w^{-1}L \mid w \in A^*, \varepsilon \in w^{-1}L\}$. If $w_1^{-1}L = w_2^{-1}L$, then $(w_1x)^{-1}L = x^{-1}(w_1^{-1}L) = x^{-1}(w_2^{-1}L) = (w_2x)^{-1}L$, so $\cdot$ is well-defined. Hence $D(L)$ is indeed a DFA.

Example 2.5. Let $A = \{a, b\}$ and $L = A^*ab$. Continuing Example 2.4 and following Definition 2.7, $D(L)$ is:



The initial state is $\varepsilon^{-1}L = L$, the set of final states is fixed to $\{\varepsilon \cup L\}$ because $\varepsilon \cup L$ is the only left-quotient that contains ε.

Lemma 2.5. *Let $L \subseteq A^*$ s.t. $\sim_L$ is of finite index and let $D(L) = \langle Q, A, \cdot, q_0, F\rangle$ be the canonical automaton. Then*

1. $v^{-1}L \cdot w = (vw)^{-1}L$,
2. $L(D(L)) = L$, *and*
3. $\sim_{D(L)} = \sim_L$.

Proof. Let us first prove $v^{-1}L \cdot w = (vw)^{-1}L$ by induction on $|w|$. If $w = \varepsilon$, we have $v^{-1}L \cdot \varepsilon = v^{-1}L$. For the induction step, let $w = xw'$, then we have

$$v^{-1}L \cdot xw' = (v^{-1}L \cdot x) \cdot w' = (vx)^{-1}L \cdot w' =^{\text{IH}} (vxw')^{-1}L.$$

For 2. we have

$$L(D(L)) = \{w \in A^* \mid L \cdot w \in F\} = \{w \in A^* \mid \varepsilon \in L \cdot w\} =^{1.} \{w \in A^* \mid \varepsilon \in w^{-1}L\}$$

and since $\varepsilon \in w^{-1}L$ iff $w \in L$ we obtain $L(D(L)) = L$.

For 3. note that $w_1 \sim_{D(L)} w_2$ iff $L \cdot w_1 = L \cdot w_2$ iff (by 1.) $w_1^{-1}L = w_2^{-1}L$ iff (by Lemma 2.3) $w_1 \sim_L w_2$. □

We are now ready to prove the Myhill-Nerode theorem.

Theorem 2.1. *Let $L \subseteq A^*$. Then L is regular iff the index of $\sim_L$ is finite.*

Proof. For the left-to-right direction, let D be a DFA that accepts L. Then $\sim_D$ is a refinement of $\sim_L$ and therefore the index of $\sim_L$ is at most the index of $\sim_D$. But the index of $\sim_D$ is at most the number of states in D which is finite.

The other direction follows directly from Lemma 2.5/2. □

The Myhill-Nerode theorem is an algebraic characterisation of the regular languages. It does not refer to any notion of automaton, grammar or a similar formalism.

Example 2.6. Let $L = \{a^n b^n \mid n \in \mathbb{N}\}$. We can show that L is not regular as follows: let $p, q \in \mathbb{N}$ s.t. $p \neq q$. Then $a^p \not\sim_L a^q$ because $a^p b^p \in L$ but $a^q b^p \notin L$. So there are infinitely many $\sim_L$-classes, hence by the Myhill-Nerode theorem, L is not regular.

The size of an automaton is the number of its states. Consequently an automaton is called minimal if it has a minimal number of states

Theorem 2.2. *Let L be a regular language. Then $D(L)$ is the unique minimal DFA for L up to isomorphism.*

Proof. We first show that $D(L)$ is a minimal automaton. To that aim, let D be any DFA that accepts L. Then $\sim_D$ is a refinement of $\sim_L$, hence the index of $\sim_D$ is at least the index of $\sim_L$, and therefore the number of states of D is at least the index of $\sim_L = \sim_{D(L)}$ which, since every state in $D(L)$ is reachable, is the number of states in $D(L)$.

Now let D be any minimal DFA and assume w.l.o.g. that all states in D are accessible. Then the number of states in D is equal to the number of states in $D(L)$. Therefore the index of $\sim_D$ is equal to the index of $\sim_{D(L)}$. But $\sim_D$ is a refinement of $\sim_L = \sim_{D(L)}$ hence $\sim_D = \sim_{D(L)}$. Therefore by Lemma 2.1 we see that $D \simeq D(L)$. □

2.2 Transition monoids

2.2.1 Quotient monoids

Definition 2.8. Let M be a monoid. An equivalence relation $\approx$ on M is called *congruence* if $x \approx y$ implies that, for all $z_1, z_2 \in M$: $z_1 x z_2 \approx z_1 y z_2$.

Lemma 2.6. *Let M be a monoid and $\approx$ a congruence on M. Then ${}^{M}\!/\!_{\approx}$ with the natural operations forms a monoid.*

Proof. The unit element of ${}^{M}\!/\!_{\approx}$ is $[e]$ and the operation is defined as $[x][y] = [xy]$. To see that the operation is well-defined let $x_1 \approx x_2$ and $y_1 \approx y_2$. Then, for all $i, j, k, l \in \{1, 2\}$, we have $x_i y_j \approx x_k y_l$ because $x_i \approx x_k$ implies $x_i y_j \approx x_k y_j$ and $y_j \approx y_l$ implies $x_k y_j \approx x_k y_l$ and therefore $x_i y_j \approx x_k y_l$. Then associativity and $[e]$ being a unit element follow directly from the respective properties of M. □

Definition 2.9. Let $\varphi : M \to N$ be a monoid homomorphism. Define the relation $\approx_\varphi$ on M by $m_1 \approx_\varphi m_2 \Leftrightarrow \varphi(m_1) = \varphi(m_2)$.

Clearly, $\approx_\varphi$ is an equivalence relation. Moreover, it is also a congruence: let $m_1, m_2, m_3, m_4 \in M$ and $m_1 \approx_\varphi m_2$. Then

$$\varphi(m_3 m_1 m_4) = \varphi(m_3)\varphi(m_1)\varphi(m_4) = \varphi(m_3)\varphi(m_2)\varphi(m_4) = \varphi(m_3 m_2 m_4)$$

and therefore $m_3 m_1 m_4 \approx_\varphi m_3 m_2 m_4$.

Lemma 2.7. *Let M, N be monoids, $\varphi : M \to N$ a homomorphism. Then ${}^{M}\!/\!_{\approx_\varphi} \simeq \varphi(M)$.*

Proof. Define $\bar{\varphi} : {}^{M}\!/\!_{\approx_\varphi} \to \varphi(M), [m] \mapsto \varphi(m)$. First observe that $m_1 \approx_\varphi m_2$ iff $\varphi(m_1) = \varphi(m_2)$. Reading this from left to right shows that $\bar{\varphi}$ is well-defined. Reading it from right to left shows that $\bar{\varphi}$ is injective. Moreover, $\bar{\varphi}$ is surjective since $\forall n \in \varphi(M)\, \exists m \in M$ s.t. $n = \varphi(m)$. It remains to show that $\bar{\varphi}$ is a homomorphism. To that aim, observe that

$$\begin{aligned} \bar{\varphi}([e]) &= \varphi(e) = e, \text{ and,} \\ \bar{\varphi}([m_1][m_2]) &= \bar{\varphi}([m_1 m_2]) = \varphi(m_1 m_2) = \varphi(m_1)\varphi(m_2) = \bar{\varphi}([m_1])\bar{\varphi}([m_2]). \end{aligned}$$

□

It will occasionally be convenient to present a monoid in terms of generators and relations. Remember that, for an alphabet A, we write A^* for the monoid freely generated by A, i.e. for the set of all words of finite length consisting of letters of A.

Definition 2.10. Let $u_1, v_1, \ldots, u_n, v_n \in A^*$. Then we say that the coarsest[2] congruence $\approx$ which satisfies $u_1 \approx v_1, \ldots, u_n \approx v_n$ is the *congruence induced by the equations* $u_1 = v_1, \ldots, u_n = v_n$. In this situation, the monoid ${}^{A^*}\!/\!_{\approx}$ is called the *monoid given by the generators A and the relations* $u_1 = v_1, \ldots, u_n = v_n$. ${}^{A^*}\!/\!_{\approx}$ is written as $\langle A \mid u_1 = v_1, \ldots, u_n = v_n \rangle$.

Example 2.7. $\langle a, b \mid a^2 = a, b^2 = b, ab = ba \rangle$ consists of four elements: $[\varepsilon]$, $[a]$, $[b]$, $[ab]$ which represent, respectively, the empty word, the words consisting of a only, the words consisting of b only, and the words containing both a and b.

It will often be useful to consider quotients up to isomorphism. To that aim, we define the following abstract notion of quotient.

Proposition 2.1. *Let M, N be monoids. Then the following are equivalent:*

1. *there is a surjective homomorphism $\varphi : M \to N$*

2. *there is a congruence $\approx$ on M s.t. $N \simeq {}^{M}\!/\!_{\approx}$.*

In this case we say that "N is a quotient of M".

Proof. For 1. $\Rightarrow$ 2. let $\varphi : M \to N$ be surjective. Then $\approx_\varphi$ is a congruence and ${}^{M}\!/\!_{\approx_\varphi}$ is a monoid with ${}^{M}\!/\!_{\approx_\varphi} \simeq \varphi(M)$. Since φ is surjective, $\varphi(M) = N$.

For 2. $\Rightarrow$ 1. let $\approx$ be a congruence on M and $\varphi : {}^{M}\!/\!_{\approx} \to N$ an isomorphism. Define $\psi : M \to N, m \mapsto \varphi([m])$. Since ψ is the composition of the two surjective homomorphisms $m \mapsto [m]$ and φ it is a surjective homomorphism too. □

2.2.2 The transition monoid of a DFA

For a finite set Q we will write Q^Q for the set of all functions from Q to Q. Given a DFA $D = \langle Q, A, \cdot, q_0, F \rangle$, a word $w \in A^*$ induces the transition

$$\tau_{D,w} : Q \to Q, q \mapsto q \cdot w.$$

Note that $\tau_{D,\varepsilon} = \mathrm{id}$ for every DFA D. If D is clear from the context we will often just write τ_w.

Definition 2.11. Let $D = \langle Q, A, \cdot, q_0, F \rangle$ be a DFA, then the *transition monoid $M(D)$ of D* is defined as

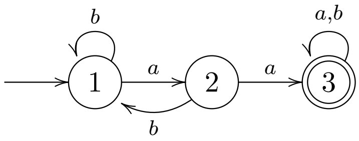
$$M(D) = \{\tau_{D,w} \in Q^Q \mid w \in A^*\}$$

with unit $\tau_{D,\varepsilon} \in M(D)$ and the monoid operation defined by $\tau_{D,w_1} \circ \tau_{D,w_2} = \tau_{D,w_1 w_2}$.

Note that the operation in this monoid is *not* the usual composition of functions $(f \circ g)(x) = f(g(x))$ but instead the reversed composition $(f \circ g)(x) = g(f(x))$. This notation is more convenient as it corresponds directly to the concatenation of words. Note that $M(D)$ is closed under composition (of functions) since A^* is closed under composition (of words) and that $\tau_{D,\varepsilon} = \mathrm{id}$. Therefore $M(D)$ is a submonoid of Q^Q. Since Q^Q is finite, so is $M(D)$. Also note that $\tau_D : A^* \to Q^Q, w \mapsto \tau_{D,w}$ is a monoid homomorphism. It allows to define the monoid $M(D)$ as the submonoid $\tau_D(A^*)$ of Q^Q.

[2] i.e. the one which only makes the necessary identifications

Example 2.8. Let D be the following DFA:



In order to compute the transition monoid $M(D)$ we create the following table:

	1	2	3	
τ_ε	1	2	3	
τ_a	2	3	3	
τ_b	1	1	3	
τ_{aa}	3	3	3	
τ_{ab}	1	3	3	
τ_{ba}	2	2	3	
τ_{bb}	1	1	3	$\tau_{bb} = \tau_b$
τ_{aaa}	3	3	3	$\tau_{aaa} = \tau_{aa}$
τ_{aab}	3	3	3	$\tau_{aab} = \tau_{aa}$
τ_{aba}	2	3	3	$\tau_{aba} = \tau_a$
τ_{baa}	3	3	3	$\tau_{baa} = \tau_{aa}$
τ_{bab}	1	1	3	$\tau_{bab} = \tau_b$

Now the table is saturated, because every $w \in A^*$ not in the table contains a subword with an equation on the right-hand side (and hence it induces the same function as a shorter word). This concludes the computation and we have:

$$M(D) = \{\mathrm{id}, \tau_a, \tau_b, \tau_{aa}, \tau_{ab}, \tau_{ba}\}$$

Definition 2.12. Let $D = \langle Q, A, \cdot, q_0, F\rangle$ be a DFA and $\tau_D : A^* \to Q^Q, w \mapsto \tau_{D,w}$. Then the congruence $\approx_{\tau_D}$ is called *congruence of* D and written more succinctly as $\approx_D$.

By definition we have $w_1 \approx_D w_2$ iff $w_1 \approx_{\tau_D} w_2$ iff $\tau_{D,w_1} = \tau_{D,w_2}$ iff $\forall q \in Q : q \cdot w_1 = q \cdot w_2$.

Lemma 2.8. *Let D be a DFA. Then $M(D) \simeq {}^{A^*}\!/\!_{\approx_D}$.*

Proof. Observe that $\tau_D : A^* \to M(D), w \mapsto \tau_{D,w}$ is a surjective homomorphism, so, by Lemma 2.7, ${}^{A^*}\!/\!_{\approx_{\tau_D}} \simeq \tau_D(A^*) = M(D)$. □

So we see that there are two ways to think about the transition monoid of an automaton: either, literally, as the monoid of transitions with composition of functions as operation, or as monoid of $\approx_D$-equivalence classes with composition of words.

Example 2.9. The monoid $M(D)$ from Example 2.8 is isomorphic to the generator and relations representation

$$\langle a, b \mid a = aba, b = b^2 = bab, a^2 = a^2b = a^3 = ba^2\rangle$$

which consists of the $\approx_D$-equivalence classes

$$[\varepsilon], [a], [b], [a^2], [ab], [ba].$$

Note that the congruence $\approx_D$ of D is a refinement of the right-congruence $\sim_D$ of D, i.e., $w_1 \approx_D w_2$ implies $w_1 \sim_D w_2$.

Example 2.10. Letting D be the DFA from Example 2.8, we have $w_1 \sim_D w_2$ iff $1 \cdot w_1 = 1 \cdot w_2$ iff w_1 and w_2 have the same entry in the first column. So the $\sim_D$-equivalence classes are

$$\begin{aligned} [\varepsilon]_{\sim_D} &= [\varepsilon]_{\approx_D} \cup [b]_{\approx_D} \cup [ab]_{\approx_D} \\ [a]_{\sim_D} &= [a]_{\approx_D} \cup [ba]_{\approx_D} \\ [a^2]_{\sim_D} &= [a^2]_{\approx_D} \end{aligned}$$

For example, $a \sim_D ba$, since, for membership in $L(D) = A^*a^2A^*$, the leading b is irrelevant.

2.2.3 Languages recognised by a monoid

Definition 2.13. Let $L \subseteq A^*$ and $\varphi : A^* \to M$ be a monoid homomorphism. We say that L *is recognised by* φ if there is a $P \subseteq M$ s.t. $L = \varphi^{-1}(P)$. In this case we also say that L *is recognised by* M.

Note that if M_1 is a monoid that recognises L and M_2 is isomorphic to M_1, then also M_2 recognises L. Moreover, if $\varphi : A^* \to M$ recognises $L \subseteq A^*$, then L can also be recognised by the surjective homomorphism $\varphi : A^* \to \varphi(A^*)$ in the submonoid $\varphi(A^*)$ of M.

Example 2.11. Let $A = \{a, b\}$. Consider the monoid $(\mathbb{Z}/2\mathbb{Z}, +, 0)$ and the homomorphism $\varphi : A^* \to \mathbb{Z}/2\mathbb{Z}$ defined by $\varphi(a) = 1$ and $\varphi(b) = 0$. Then $\varphi^{-1}(\{0\})$ is the set of words that contain an even number of a's.

Lemma 2.9. *Let $L \subseteq A^*$ and $\varphi : A^* \to M$ be a homomorphism. The following are equivalent:*

1. *L is recognised by φ.*
2. *$\forall w \in A^*$: $w \in L \Leftrightarrow \varphi(w) \in \varphi(L)$.*
3. *$\varphi^{-1}(\varphi(L)) = L$.*

Proof. (1) $\Rightarrow$ (2): Let $P \subseteq M$ s.t. $L = \varphi^{-1}(P)$ and let $w \in A^*$. If $w \in L$ then $\varphi(w) \in \varphi(L)$. For the other direction let $\varphi(w) \in \varphi(L)$. Then there is a $v \in L$ s.t. $\varphi(w) = \varphi(v) \in P$. Therefore $w \in L$.

(2) $\Rightarrow$ (3): $\varphi^{-1}(\varphi(L)) = \{v \in A^* \mid \varphi(v) \in \varphi(L)\} =^{(2)} \{v \in A^* \mid v \in L\} = L$.

(3) $\Rightarrow$ (1): let $P = \varphi(L) \subseteq M$. Then $L = \varphi^{-1}(P)$, so L is recognised by φ. □

So, quite naturally, $\varphi(L)$ is a $P \subseteq M$ s.t. $L = \varphi^{-1}(P)$. In case φ is surjective $\varphi(L)$ is[3] the only such P.

Lemma 2.10. *Let D be a DFA recognising a language $L \subseteq A^*$. Then $M(D)$ recognises L.*

Proof. Let $D = \langle Q, A, \cdot, q_0, F\rangle$, $\tau_D : A^* \to M(D), w \mapsto \tau_{D,w}$, and $P = \{\tau \in M(D) \mid q_0 \cdot \tau \in F\}$. Then

$$w \in L \Leftrightarrow q_0 \cdot \tau_{D,w} \in F \Leftrightarrow \tau_{D,w} \in P \Leftrightarrow \tau_D(w) \in P \Leftrightarrow w \in \tau_D^{-1}(P)$$

and hence $L = \tau_D^{-1}(P)$. □

Example 2.12. The monoid $M(D)$ of Example 2.8 recognises the language $A^*a^2A^*$, more precisely: the homomorphism $\tau_D : A^* \to M(D)$ satisfies $\tau_D^{-1}(\tau_D(L)) = L$. We have $\tau_D(L) = \{\tau_{aa}\}$.

[3] Show this as exercise.

Theorem 2.3. *A language $L \subseteq A^*$ is regular iff L is recognised by a finite monoid.*

Proof. The implication from left to right follows directly from Lemma 2.10. For the other direction, let $L \subseteq A^*$, M be a finite monoid, $\varphi : A^* \to M$ a homomorphism and $P \subseteq M$ s.t. $L = \varphi^{-1}(P)$. We define the DFA $D = \langle M, A, \cdot, e, P\rangle$ by $m \cdot x = m\varphi(x)$ for $m \in M, x \in A$. First we claim that $m \cdot w = m\varphi(w)$ for all $w \in A^*$. To show that proceed by induction on $|w|$. For $w = \varepsilon$ we have $m \cdot \varepsilon = m = m\varphi(\varepsilon)$. For the words on length 1 this follows directly from the definition. For a word w with $|w| \geqslant 2$, let $w = w_1w_2$ s.t. both w_1 and w_2 have length at least 1. We then have

$$m \cdot w = m \cdot w_1w_2 = (m \cdot w_1) \cdot w_2 = m\varphi(w_1) \cdot w_2 = m\varphi(w_1)\varphi(w_2) = m\varphi(w_1w_2) = m\varphi(w).$$

Summing up, we have

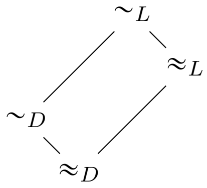
$$w \in L(D) \Leftrightarrow e \cdot w \in P \Leftrightarrow \varphi(w) \in P \Leftrightarrow w \in \varphi^{-1}(P) = L$$

and therefore $L(D) = L$. □

2.2.4 The syntactic monoid

Definition 2.14. Let $L \subseteq A^*$. The *syntactic congruence* $\approx_L$ is the relation on A^* defined as: $w_1 \approx_L w_2$ iff for all $u, v \in A^*$: $uw_1v \in L \Leftrightarrow uw_2v \in L$.

So we have now seen four equivalence relation: the two right-congruences $\sim_D$ and $\sim_L$ and the two congruences $\approx_D$ and $\approx_L$. Given a DFA D which recognises a language L we have the following diagram



where an upward line indicates "is a refinement of". We have already seen that $\approx_D$ is a refinement of $\sim_D$ and that $\sim_D$ is a refinement of $\sim_L$. Let now $w_1 \approx_D w_2$ and $u, v \in A^*$, then $uw_1v \in L$ iff $q_0 \cdot uw_1v \in F$ iff $q_0 \cdot uw_2v \in F$ iff $uw_2v \in L$. So $\approx_D$ is a refinement of $\approx_L$. For the remaining arrow, let $w_1 \approx_L w_2$ and $v \in A^*$, then $w_1v \in L$ iff $w_2v \in L$ and hence $w_1 \sim_L w_2$.

Lemma 2.11. *Let $L \subseteq A^*$ be regular. Then $\approx_L \; = \; \approx_{D(L)}$.*

Proof. Let $D(L) = \langle Q, A, \cdot, q_0, F\rangle$. We have

$$\begin{aligned} w_1 \approx_{D(L)} w_2 \;&\text{iff } \forall q \in Q\colon q \cdot w_1 = q \cdot w_2\\ &\text{iff } \forall u \in A^*\colon u^{-1}L \cdot w_1 = u^{-1}L \cdot w_2\\ &\text{iff (by Lemma 2.5/1) } \forall u \in A^*\colon (uw_1)^{-1}L = (uw_2)^{-1}L\\ &\text{iff (by Lemma 2.3) } \forall u \in A^*\colon uw_1 \sim_L uw_2\\ &\text{iff } \forall u, v \in A^*\colon uw_1v \in L \Leftrightarrow uw_2v \in L\\ &\text{iff } w_1 \approx_L w_2. \end{aligned}$$

□

In Lemma 2.5/3 we have already shown that $\sim_L \,=\sim_{D(L)}$. So, in terms of the above diagram we see that, if D is the minimal automaton, then the D-line and the L-line coincide.

Definition 2.15. Let $L \subseteq A^*$. The *syntactic monoid of* L is defined as $M(L) = A^*/{\approx_L}$.

Theorem 2.4. *Let $L \subseteq A^*$ be a regular language. Then $M(L) \simeq M(D(L))$.*

Proof. We have

$$M(L) = A^*/{\approx_L} =^{\text{Lem. 2.11}} A^*/{\approx_{D(L)}} \simeq^{\text{Lem. 2.8}} M(D(L))$$

□

Corollary 2.1. *Let $L \subseteq A^*$ be a regular language. Then $M(L)$ recognises L.*

This corollary follows from Theorem 2.4 and Lemma 2.10. A closer look at the proofs of these two results reveals that $\tau_{D(L)} : A^* \to M(D(L)), w \mapsto \tau_{D(L),w}$ recognises L. Concatenating $\tau_{D(L)}$ with the isomorphism $\iota : M(D(L)) \to M(L), \tau_{D(L),w} \mapsto [w]_{\approx_L}$ we obtain the *syntactic homomorphism* $\eta : A^* \to M(L), w \mapsto [w]_{\approx_L}$ which recognises L, i.e., $L = \eta^{-1}(\eta(L))$.

2.2.5 The monoids recognising a fixed language

We will now see a generalisation of the minimality and uniqueness property of $D(L)$: the monoids which recognise L are exactly the monoids which are divided by the syntactic monoid $M(L)$ of L. Let us first make precise what being a divisor means.

Definition 2.16. Let M, N be monoids. Then N is a *divisor* of M, written as $N \preccurlyeq M$, if N is a quotient of a submonoid of M.

Theorem 2.5. *Let $L \subseteq A^*$ and M be a monoid. Then M recognises L iff $M(L) \preccurlyeq M$.*

Proof. For the left to right direction, let $\varphi : A^* \to M$ be a homomorphism that recognises L, then so does the surjective homomorphism $\varphi : A^* \to \varphi(A^*)$. Note that $\varphi(A^*)$ is a submonoid of M and that $\varphi(A^*) \simeq A^*/{\approx_\varphi}$. Let $\eta : A^* \to M(L), w \mapsto [w]_{\approx_L}$ be the syntactic homomorphism. Now we claim that $\approx_\varphi$ is a refinement of $\approx_L$: let $w_1 \approx_\varphi w_2$ and let $u, v \in A^*$, then $uw_1v \in L$ iff $\varphi(uw_1v) \in \varphi(L)$ iff $\varphi(u)\varphi(w_1)\varphi(v) \in \varphi(L)$ iff $\varphi(u)\varphi(w_2)\varphi(v) \in \varphi(L)$ iff $\varphi(uw_2v) \in \varphi(L)$ iff $uw_2v \in L$. We define $\pi : A^*/{\approx_\varphi} \to A^*/{\approx_L}, [w]_{\approx_\varphi} \mapsto [w]_{\approx_L}$ and observe that π is a surjective homomorphism. Therefore $M(L)$ is a quotient of $\varphi(A^*)$ which is a submonoid of M, hence $M(L) \preccurlyeq M$.

$$\begin{tikzcd}
A^* \arrow[rd, "\varphi"] \arrow[dd, "\eta"'] & \\
& A^*/{\approx_\varphi} \simeq \varphi(A^*) \subseteq M \arrow[ld, dashed, "\pi"] \\
M(L) = A^*/{\approx_L} &
\end{tikzcd}$$

For the right to left direction let $\eta : A^* \to M(L), w \mapsto [w]_{\approx_L}$. If $M(L) \preccurlyeq M$, then there is a submonoid N of M and a surjective homomorphism $\beta : N \to M(L)$. Define $\varphi_0 : A \to N$ by picking for each $x \in A$ an element $\varphi_0(x) \in \beta^{-1}(\eta(x))$. Then there is a unique homomorphism $\varphi : A^* \to N$ which extends φ_0. Thus $\beta \circ \varphi$ is a homomorphism. By definition of φ we have $(\beta \circ \varphi)(x) = \eta(x)$ for all $x \in A$ and therefore $\beta \circ \varphi = \eta$. ¸So we have

$$L =^{\text{Lem. 2.9}} \eta^{-1}(\eta(L)) = \varphi^{-1}(\beta^{-1}(\eta(L)))$$

and letting $P = \beta^{-1}(\eta(L)) \subseteq N$ we see that N, and therefore also M, recognises L.

$$\begin{tikzcd}
A^* \arrow[dr, dashed, "\varphi"] \arrow[dd, "\eta"'] & \\
& N \subseteq M \arrow[dl, "\beta"] \\
M(L) &
\end{tikzcd}$$

□

2.3 Star height

In this section we study the star height of regular languages, i.e., the number of nested stars necessary to specify a language by a regular expression. Of particular interest are the star-free languages, i.e., those that can be specified by a star-free regular expression. We will prove Schützenberger's theorem which characterises the star-free languages as those whose syntactic monoid is aperiodic.

2.3.1 Star-free languages

Definition 2.17. Let A be an alphabet. An *A-regular expression* is an expression formed from $x \in A$, ε and $\varnothing$ with $\cdot, \cup, *$ and ${}^{\mathrm{c}}$.

A regular expression E defines a regular language, written as $L(E)$, in the straightforward way. Note that the alphabet must be fixed for $L(E^{\mathrm{c}}) = A^*\backslash L(E)$ to be well-defined. We also write $E_1 \cap E_2$ for $(E_1^{\mathrm{c}} \cup E_2^{\mathrm{c}})^{\mathrm{c}}$ and $E_1\backslash E_2$ for $E_1 \cap E_2^{\mathrm{c}}$.

Example 2.13. Letting $A = \{a, b\}$ and $E = b(a(aa)^*)^{\mathrm{c}}b$, observe that $L(E)$ is the set of words in A^* which start and end with a b between which is anything but an odd number of a's.

Definition 2.18. Let E be an A-regular expression. The *star height* $h(E)$ is defined as follows:

1. $h(x) = h(\varepsilon) = h(\varnothing) = 0$ for all $x \in A$.
2. $h(E_1 \cup E_2) = h(E_1 \cdot E_2) = \max\{h(E_1), h(E_2)\}$.
3. $h(E_0^*) = h(E_0) + 1$.
4. $h(E_0^{\mathrm{c}}) = h(E_0)$.

Definition 2.19. Let $L \subseteq A^*$. Then the *star height of* L is $h(L) = \min\{h(E) \mid L(E) = L\}$.

A language L with $h(L) = 0$ is also called *star-free*.

Example 2.14. Let $A = \{a, b\}$. The A-regular expression $(ab)^*$ has star height 1. However, $L((ab)^*)$ has star height 0 as the following equivalence shows:

$$(ab)^* = ((a\varnothing^{\mathrm{c}} \cap \varnothing^{\mathrm{c}}b)\backslash(\varnothing^{\mathrm{c}}a^2\varnothing^{\mathrm{c}} \cup \varnothing^{\mathrm{c}}b^2\varnothing^{\mathrm{c}})) \cup \varepsilon$$

What is the star height of the language $(aa)^*$? The above trick does not work directly. At this point, the answer is not clear. In fact, at the end of this section we will be able to show $h(L((aa)^*)) = 1$.

Usually, the alphabet is clear from the context. However, in this section we will work with different alphabets and so we want to make it explicit in the notation for the set of star-free languages.

Definition 2.20. Let A be an alphabet. Then $\mathrm{SF}(A)$ is the set of star-free languages in A.

Clearly, $\mathrm{SF}(A)$ is closed under concatenation and the Boolean operations.

Lemma 2.12. *Let $B \subseteq A$. Then 1. $B^* \in \mathrm{SF}(A)$ and 2. $\mathrm{SF}(B) \subseteq \mathrm{SF}(A)$.*

Proof. For 1. consider the A-regular expression $E_B = \varnothing^{\mathrm{c}} \backslash (\bigcup_{x \in A \backslash B} \varnothing^{\mathrm{c}} x \varnothing^{\mathrm{c}})$ and observe that $L(E_B) = B^*$. For 2. let $L \in \mathrm{SF}(B)$ and let E be a star-free B-regular expression s.t. $L(E) = L$. Then E is also an A-regular expression and $L(E \cap E_B) = L(E) \cap B^* = L$. □

Schützenberger's theorem provides an algebraic characterisation of the star-free languages as those recognised by aperiodic monoids.

Definition 2.21. A monoid M is called *aperiodic* if for every $x \in M$ there is an $n \in \mathbb{N}$ s.t. $x^n = x^{n+1}$.

Example 2.15. Let G be an aperiodic group, then multiplying $x^n = x^{n+1}$ with x^{-n} yields $x = 1$. Thus $G = \{1\}$. So there are no non-trivial aperiodic groups.

Note that, if M is aperiodic and N is a submonoid of M then trivially also N is aperiodic. Moreover, if M is aperiodic and N is a quotient of M, i.e. there exists a surjective homomorphism $\varphi : M \to N$, then N is aperiodic. This can be shown as follows: let $y \in N$, then there is an $x \in M$ and an $n \in \mathbb{N}$ with $\varphi(x) = y$ and $x^n = x^{n+1}$. Therefore we have

$$y^n = \varphi(x)^n = \varphi(x^n) = \varphi(x^{n+1}) = \varphi(x)^{n+1} = y^{n+1}.$$

As a consequence, if M is aperiodic and $N \preccurlyeq M$ then also N is aperiodic.

Theorem (Schützenberger)**.** *Let $L \subseteq A^*$. Then the following are equivalent:*

1. *L is star-free.*
2. *$M(L)$ is a finite aperiodic monoid.*
3. *L is recognised by a finite aperiodic monoid.*

Example 2.16. Based on this theorem we can now show that the star height of $L((aa)^*)$ is 1. That it is at most 1 is clear from the given regular expression. But it must also be at least 1 since the syntactic monoid of $L((aa)^*)$ is not aperiodic. To see that, one can carry out the routine computation of $M(D(L))$ and check, for each of its finitely many elements, whether it is aperiodic or not.

A shortcut is to observe that $aa \approx_L \varepsilon$ but $a \not\approx_L \varepsilon$ and that hence, for all $n \geqslant 0$ we have $[a]^{2n} = [\varepsilon]$ and $[a]^{2n+1} = [a]$ and therefore $M(L) = {}^{A^*}\!/_{\approx_L}$ is not aperiodic.

The rest of this section is devoted to the proof of Schützenberger's theorem. The implication from 2 to 3 follows directly from the fact that $M(L)$ recognises L. For the implication from 3 to 2 assume that L is recognised by a finite aperiodic monoid M. Then, by Theorem 2.5, $M(L) \preccurlyeq M$ so, by the above observations, also $M(L)$ is aperiodic. So the proof consists essentially of bridging the gap between star-freeness and aperiodicity. The direction from star-freeness to aperiodicity is rather straightforward and will be finished in the below Lemma 2.13. The other direction will need considerably more work.

Lemma 2.13. *Let $L \subseteq A^*$ be a star-free language. Then $M(L)$ is finite and aperiodic.*

Proof. Since every star-free language is regular, $M(L)$ is finite. For aperiodicity it suffices to show that there is an $n \in \mathbb{N}$ s.t. for all $w, u, v \in A^*$:

$$uw^n v \in L \Leftrightarrow uw^{n+1}v \in L,$$

i.e., $w^n \approx_L w^{n+1}$. Because then for all $[w] \in M(L) = {}^{A^*}\!/_{\approx_L}$ we have $[w]^n = [w^n] = [w^{n+1}] = [w]^{n+1}$ thus $M(L)$ is aperiodic.

Let E be a star-free A-regular expression. We show by induction on E that there is $n(E)$ s.t. for all $w \in A^*$: $w^{n(E)} \approx_L w^{n(E)+1}$. For $E = \varnothing$ let $n(\varnothing) = 0$, then $uw^0v \notin \varnothing$ and $uw^1v \notin \varnothing$. For $E = x \in A$, let $n(x) = 2$. Then $uw^2v = x$ implies $w = \varepsilon$ and $uv = x$ which in turn implies $uw^3v = x$ and vice versa. For $E = \varepsilon$, let $n(\varepsilon) = 1$. Then $uw^1v = \varepsilon$ iff $u = v = w = \varepsilon$ iff $uw^2v = \varepsilon$.

If $E = E_1 \cup E_2$, let $n(E) = \max\{n(E_1), n(E_2)\}$. Then

$$\begin{aligned}
&uw^{n(E)}v \in L(E)\\
\text{iff } &uw^{n(E)}v \in L(E_1) \text{ or } uw^{n(E)}v \in L(E_2)\\
\text{iff } &uw^{n(E)+1}v \in L(E_1) \text{ or } uw^{n(E)+1}v \in L(E_2)\\
\text{iff } &uw^{n(E)+1}v \in L(E).
\end{aligned}$$

If $E = E_0^c$, let $n(E) = n(E_0)$. Then $uw^{n(E)}v \in L(E)$ iff $uw^{n(E)}v \notin L(E_0)$ iff $uw^{n(E)+1}v \notin L(E_0)$ iff $uw^{n(E)+1}v \in L(E)$.

If $E = E_1E_2$ let $n(E) = n(E_1) + n(E_2) + 1$. Let $u, v, w \in A^*$, then we have

$$\begin{aligned}
&uw^{n(E_1)+n(E_2)+1}v \in L(E_1E_2)\\
\text{iff } &uw^{n(E_1)}v' \in L(E_1) \text{ and } v'' \in L(E_2) \text{ s.t. } v'v'' = w^{n(E_2)+1}v \text{ or}\\
&u'w^{n(E_2)}v \in L(E_2) \text{ and } u'' \in L(E_1) \text{ s.t.} u''u' = uw^{n(E_1)+1}\\
\text{iff } &uw^{n(E_1)+1}v' \in L(E_1) \text{ with } v'' \text{ as above or}\\
&u'w^{n(E_2)+1}v \in L(E_2) \text{ with } u'' \text{ as above}\\
\text{iff } &uw^{n(E_1)+n(E_2)+2}v \in L(E_1E_2).
\end{aligned}$$

□

2.3.2 Local divisors

Lemma 2.14. *Let M be a monoid and $k \in M$. Then $M_k = (kM \cap Mk, \circ, k)$ with $xk \circ ky := xky$ is a monoid and a divisor of M.*

For example, if $A = \{a, b\}$, then A^*_{aba} is the set of words that start and end with aba. In particular, also $aba, ababa \in A^*_{aba}$ which shows that, in general, $kM \cap Mk \neq kMk$. In A^*_{aba} we have, e.g., $abacbaba \circ ababa = abacbababa$.

Proof. The operation is well-defined since $x_1k = x_2k$ and $y_1k = y_2k$ implies $x_1ky_1 = x_1ky_2 = x_2ky_2$. Furthermore, M_k is closed under $\circ$ since $xky = kx'y$ and $xky = xy'k$. The operation is associative since

$$\begin{aligned}
&(xk \circ ky) \circ kz = xky \circ kz = xy'k \circ kz = xy'kz = xkyz, \text{ and}\\
&xk \circ (ky \circ kz) = xk \circ (y'k \circ kz) = xk \circ y'kz = xk \circ kyz = xkyz.
\end{aligned}$$

Also k is a unit element since $k \circ kx = kx$ and $kx \circ k = x'k \circ k = x'k = kx$.

Let $M' = \{x \in M \mid kx \in Mk\}$. Then M' is a submonoid of M, because if $x \in M$ with $kx \in Mk$ and $y \in M$ with $ky \in Mk$, then $xy \in M$ and $kxy = x'ky = x'y'k \in Mk$. Moreover, $\varphi : M' \to M_k, x \mapsto kx$ is a homomorphism because $\varphi(1) = k$ and $\varphi(x)\varphi(y) = kx \circ ky = x'k \circ ky = x'ky = kxy = \varphi(xy)$. It remains to show that φ is surjective: let $z \in M_k$, then there are $x, x' \in M$ s.t. $z = kx = x'k$. Now $\varphi(x) = kx = z$, $x \in M$ and $kx = x'k \in Mk$ hence $x \in M'$. Therefore M_k is a divisor of M. □

M_k is called *local divisor of M at k*.

Lemma 2.15. *Let M be aperiodic and $x_1, \ldots, x_k \in M$. Then $x_1 \cdots x_k = 1$ iff $x_i = 1$ for all $i \in \{1, \ldots, k\}$.*

Proof. The right to left direction is trivial. For the left to right direction, assume $xy = 1$. Then $1 = xy = xxyy = \ldots = x^n y^n = x^{n+1} y^n = x \cdot 1 = x$. Analogously one can show that $xy = 1$ also implies $y = 1$. The result then follows by induction. □

Lemma 2.16. *If M is a finite aperiodic monoid and $k \in M \backslash \{1\}$, then M_k is aperiodic and $|M_k| < |M|$.*

Proof. We first show $(kx)^i = kx^i$ for all $i \geqslant 0$ in M_k by induction. The induction base $i = 0$ is trivial, for the induction step we have

$$(kx)^{i+1} = kx \circ (kx)^i \overset{\text{IH}}{=} kx \circ kx^i = x'k \circ kx^i = x'kx^i = kx^{i+1}.$$

Therefore $x^n = x^{n+1}$ in M implies $(kx)^n = (kx)^{n+1}$ in M_k. Furthermore, $1 \notin kM \cap Mk$, for suppose $1 = kx$ for some $x \in M$, then by Lemma 2.15, $k = 1$ which contradicts the assumption. Therefore $|M_k| < |M|$. □

2.3.3 Schützenberger's theorem

Before we prove the main lemma, we need one more simple result about star-free languages.

Lemma 2.17. *Let A and B be alphabets, let $X \subseteq A^+$ and $\varphi : X^* \to B^*$ be a homomorphism s.t. 1. for all $u \in X$: $\varphi(u) \in B$ and 2. for all $b \in B$: $\varphi^{-1}(b) \in \mathrm{SF}(A)$. Let $L \in \mathrm{SF}(B)$, then $\varphi^{-1}(L) \in \mathrm{SF}(A)$.*

This lemma is shown, essentially, by replacing in a star-free regular expression for L over B each letter $b \in B$ by a star-free regular expression for $\varphi^{-1}(b)$ over A. The result is a star-free regular expression for $\varphi^{-1}(L)$ over A.

Proof. We proceed by induction on the structure of a star-free regular expression which defines $L \in \mathrm{SF}(B)$. If $L = \varnothing$, then $\varphi^{-1}(\varnothing) = \varnothing$. If $L = \{\varepsilon\}$, then $\varphi^{-1}(\{\varepsilon\}) = \{\varepsilon\}$ by assumption 1. on φ. If $L = \{b\}$ for a $b \in B$, then, by assumption 2. on φ, $\varphi^{-1}(\{b\}) \in \mathrm{SF}(A)$.

If $L = L_1 \cup L_2$ for $L_1, L_2 \in \mathrm{SF}(B)$, then $\varphi^{-1}(L) = \varphi^{-1}(L_1 \cup L_2) = \varphi^{-1}(L_1) \cup \varphi^{-1}(L_2)$, and thus $\varphi^{-1}(L) \in \mathrm{SF}(B)$ by induction hypothesis.

If $L = L_0^{\mathrm{c}}$ for $L_0 \in SF(B)$, then

$$\varphi^{-1}(L) = \varphi^{-1}(L_0^{\mathrm{c}}) = \{w \in A^* \mid \varphi(w) \notin L_0\} = \{w \in A^* \mid w \notin \varphi^{-1}(L_0)\} = A^* \backslash \varphi^{-1}(L_0) = \varphi^{-1}(L_0)^{\mathrm{c}}.$$

and thus $\varphi^{-1}(L) \in \mathrm{SF}(A)$ by induction hypothesis.

If $L = L_1L_2$ for $L_1, L_2 \in \mathrm{SF}(B)$, we claim that $\varphi^{-1}(L_1L_2) = \varphi^{-1}(L_1)\varphi^{-1}(L_2)$. For the right-to-left direction, let $w \in \varphi^{-1}(L_1)\varphi^{-1}(L_2)$, then $w = w_1w_2$ with $w_1 \in \varphi^{-1}(L_1)$ and $w_2 \in \varphi^{-1}(L_2)$, i.e., $\varphi(w_1) \in L_1$ and $\varphi(w_2) \in L_2$, so $\varphi(w_1w_2) \in L_1L_2$, i.e., $w = w_1w_2 \in \varphi^{-1}(L_1L_2)$. For the left-to-right direction, let $w \in \varphi^{-1}(L_1L_2)$. We have $w = u_1 \cdots u_n \in X$, so $\varphi(w) = \varphi(u_1)\cdots\varphi(u_n) \in L_1L_2$ and thus, by assumption 1. on φ, there is a $k \in \{0, \ldots, n\}$ s.t. $\varphi(u_1)\cdots\varphi(u_k) \in L_1$ and $\varphi(u_{k+1})\cdots\varphi(u_n) \in L_2$. Letting $w_1 = u_1 \cdots u_k$ and $w_2 = u_{k+1} \cdots u_n$, we have $w = w_1w_2$, $\varphi(w_1) \in L_1$, and $\varphi(w_2) \in L_2$. Therefore $w \in \varphi^{-1}(L_1)\varphi^{-1}(L_2)$. Thus, $\varphi^{-1}(L) = \varphi^{-1}(L_1L_2) = \varphi^{-1}(L_1)\varphi^{-1}(L_2) \in \mathrm{SF}(A)$ by induction hypothesis. □

The lexicographic order $<$ on $\mathbb{N} \times \mathbb{N}$ is defined by:

$$(m_1, n_1) < (m_2, n_2) \quad \text{iff} \quad \begin{array}{l} m_1 < m_2 \text{ or} \\ m_1 = m_2 \text{ and } n_1 < n_2 \end{array}$$

The following is the main lemma:

Lemma 2.18. *Let A be an alphabet, M a finite aperiodic monoid and let $\varphi : A^* \to M$ be a homomorphism. Then for all $p \in M$ we have $\varphi^{-1}(p) \in \mathrm{SF}(A)$.*

Proof. We proceed by induction on the lexicographic order on $(|M|, |A|)$. If $A = \varnothing$, then $A^* = \{\varepsilon\}$ and clearly every subset of $\{\varepsilon\}$ is a star-free language. For the case $p = 1$ we claim that $\varphi^{-1}(1) = \{x \in A \mid \varphi(x) = 1\}^*$. To show this, let $w = x_1 \cdots x_n$ with $x_i \in A$. For the left-to-right direction assume $w \in \varphi^{-1}(1)$. Then $\varphi(w) = \varphi(x_1)\cdots\varphi(x_n) = 1$ and so, by Lemma 2.15, $\varphi(x_i) = 1$ for all $i \in \{1, \ldots, n\}$. For the right-to-left direction we immediately obtain $\varphi(w) = \varphi(x_1)\cdots\varphi(x_n) = 1$. So by Lemma 2.12 we have $\varphi^{-1}(1) \in \mathrm{SF}(A)$. This covers both the case $|M| = 1$ and $\varphi(x) = 1$ for all $x \in A$.

So, for the induction step, let $c \in A$ with $\varphi(c) \neq 1$. Let $B = A \backslash \{c\}$ and $\varphi_c : B^* \to M$ be the restriction of φ to B^*. We claim that

$$\varphi^{-1}(p) = \varphi_c^{-1}(p) \cup \bigcup_{p = p_1p_2p_3} \varphi_c^{-1}(p_1)\big(\varphi^{-1}(p_2) \cap cA^* \cap A^*c\big)\varphi_c^{-1}(p_3)$$

The right-to-left inclusion is straightforward since $\varphi_c^{-1}(q) \subseteq \varphi^{-1}(q)$ and $\varphi^{-1}(p_1)\varphi^{-1}(p_2)\varphi^{-1}(p_3) \subseteq \varphi^{-1}(p)$ if $p = p_1p_2p_3$. For the left-to-right inclusion, let $w \in \varphi^{-1}(p)$. If w does not contain c, then $w \in \varphi_c^{-1}(p)$. If w contains c, then $w = w_1w_2w_3$ with $w_1, w_3 \in B^*$ and $w_2 \in cA^* \cap A^*c$. Therefore $\varphi(w) = \varphi(w_1)\varphi(w_2)\varphi(w_3) = \varphi_c(w_1)\varphi(w_2)\varphi_c(w_3)$ and letting $p_1 = \varphi_c(w_1)$, $p_2 = \varphi(w_2)$, and $p_3 = \varphi_c(w_3)$ we have $w \in \varphi_c^{-1}(p_1)(\varphi^{-1}(p_2) \cap cA^* \cap A^*c)\varphi_c^{-1}(p_3)$.

Since $(|M|, |B|) < (|M|, |A|)$ we can apply the induction hypothesis to $\varphi_c : B^* \to M$ to obtain $\varphi_c^{-1}(q) \in \mathrm{SF}(B)$ and hence, by Lemma 2.12, $\varphi_c^{-1}(q) \in \mathrm{SF}(A)$ for $q \in \{p, p_1, p_3\}$. Since $\mathrm{SF}(A)$ is closed under union and concatenation, it suffices to show that

$$\varphi^{-1}(p) \cap cA^* \cap A^*c \in \mathrm{SF}(A) \text{ for all } p \in \varphi(c)M \cap M\varphi(c).$$

The rest of this proof is devoted to doing this. Let $T = \varphi_c(B^*)$, then T is a submonoid of M. We use T as alphabet and consider the free monoid T^*. We also consider the submonoid $(B^*c)^*$ of A^* and define

$$\tau : (B^*c)^* \to T^*, v_1c \cdots v_kc \mapsto \varphi_c(v_1)\cdots\varphi_c(v_k)$$

for $k \geqslant 0$ and $v_i \in B^*$. Note that τ is a homomorphism since

$$\tau(v_1c \cdots v_ic)\tau(v_{i+1}c \cdots v_kc) = \varphi_c(v_1)\cdots\varphi_c(v_i)\varphi_c(v_{i+1})\cdots\varphi_c(v_k) = \tau(v_1c \cdots v_kc)$$

$$\begin{tikzcd}
(B^*c)^* \arrow[r, "\tau"] \arrow[d, "\varphi"'] & T^* \arrow[d, "\psi"] \\
M\varphi(c) \cup \{1\} \arrow[r, "x \mapsto \varphi(c)x"] & M_{\varphi(c)}
\end{tikzcd}$$

Figure 2.1: Lemma 2.18, induction step

Furthermore, define

$$\psi : T^* \to M_{\varphi(c)} \text{ as the unique homomorphic extension of } \varphi_c(v) \mapsto \varphi(cvc).$$

This function is well-defined since $\varphi_c(v_1) = \varphi_c(v_2)$ implies $\varphi(v_1) = \varphi(v_2)$ and hence also $\varphi(cv_1c) = \varphi(cv_2c)$, see Figure 2.1. Let $w = v_1c \cdots v_kc$ for $k \geqslant 0$ and $v_i \in B^*$. Then, in $M_{\varphi(c)}$, we have

$$\begin{aligned}
\psi(\tau(w)) &= \psi(\varphi_c(v_1) \cdots \varphi_c(v_k)) \\
&= \varphi(cv_1c) \circ \cdots \circ \varphi(cv_kc) \\
&= \varphi(c)\varphi(v_1)\varphi(c) \circ \cdots \circ \varphi(c)\varphi(v_k)\varphi(c) \\
&= \varphi(c)\varphi(v_1) \cdots \varphi(c)\varphi(v_k)\varphi(c) \\
&= \varphi(cw).
\end{aligned}$$

Therefore $cw \in \varphi^{-1}(p)$ iff $w \in \tau^{-1}(\psi^{-1}(p))$ for all $p \in M_{\varphi(c)}$, so the diagram in Figure 2.1 commutes. This shows $\varphi^{-1}(p) \cap cA^* \cap A^*c = c \cdot \tau^{-1}(\psi^{-1}(p))$ for all $p \in \varphi(c)M \cap M\varphi(c)$.

So it suffices to show $\tau^{-1}(\psi^{-1}(p)) \in \mathrm{SF}(A)$ for all $p \in M_{\varphi(c)}$. By Lemma 2.16, the monoid $M_{\varphi(c)}$ is aperiodic and $|M_{\varphi(c)}| < |M|$ so we can apply the induction hypothesis to $\psi : T^* \to M_{\varphi(c)}$ to obtain $\psi^{-1}(p) \in \mathrm{SF}(T)$. Now, observe that $\tau(w) \in T$ for all $w \in B^*c$ and that, for $t \in T$,

$$\tau^{-1}(t) = \{v_1c \cdots v_kc \in (B^*c)^* \mid \varphi_c(v_1) \cdots \varphi_c(v_k) = t\} = \{v_1c \in B^*c \mid \varphi_c(v_1) = t\} = \varphi_c^{-1}(t)c.$$

By induction hypothesis applied to φ_c we have $\varphi_c^{-1}(t) \in \mathrm{SF}(B) \subseteq \mathrm{SF}(A)$ and thus $\tau^{-1}(t) \in \mathrm{SF}(A)$. Therefore Lemma 2.17 can be applied to yield $\tau^{-1}(\psi^{-1}(p)) \in \mathrm{SF}(A)$. □

Theorem 2.6 (Schützenberger)**.** *Let $L \subseteq A^*$. Then the following are equivalent:*

1. *L is star-free.*
2. *$M(L)$ is a finite aperiodic monoid.*
3. *L is recognised by a finite aperiodic monoid.*

Proof. 1. ⇒ 2. is Lemma 2.13. 2. ⇒ 3. follows from $M(L)$ recognising L. For 3. ⇒ 1. assume that M recognises L, i.e., that there is a homomorphism $\varphi : A^* \to M$ and a $P \subseteq M$ s.t. $\varphi^{-1}(P) = L$. Then $L = \bigcup_{p \in P} \varphi^{-1}(p)$ and hence, by Lemma 2.18, L is star-free. □

Corollary 2.2. *There is an algorithm which, given a regular language L as input, e.g., as a DFA, determines whether L is star-free.*

Proof. The algorithm first computes the minimal automaton $D(L)$ from L and then the transition monoid $M(D(L))$ from $D(L)$. Since $M(L) \simeq M(D(L))$, checking $M(D(L))$ for aperiodicity yields the required result. Aperiodicity of a finite monoid is decidable. □

To this day it is unknown whether there exists a regular language of star height 2 or higher. This is known as the (generalised) star height problem.

2.4 The variety theorem

Schützenberger's theorem is a characterisation of a class of languages by the class of monoids which recognise them. There are more abstract algebraic reasons for the possibility of such a characterisation: in this section we will consider *varieties of (finite) monoids* and *varieties of regular languages* and show that "being recognised by" is a one-to-one correspondence between these (Eilenberg's variety theorem). Thus one can build a dictionary of monoid properties and classes of regular languages and translate back and forth between them. Schützenberger's result in this frame shows that the star-free languages correspond to aperiodic monoids. There are many other such correspondences. We will, for example, characterise the languages recognised by finite commutative groups in the exercises.

Definition 2.22. A class $\mathcal{M}$ of finite monoids is a *variety* if $\mathcal{M}$ satisfies the following conditions:

1. If $M \in \mathcal{M}$ and N is a submonoid of M, then $N \in \mathcal{M}$.
2. If $M \in \mathcal{M}$ and N is a quotient of M, then $N \in \mathcal{M}$.
3. If $M_1, \ldots, M_n \in \mathcal{M}$, then $\prod_{i=1}^{n} M_i \in \mathcal{M}$.

It is straightforward to show that, in the above definition, conditions 1. and 2. can be replaced by the following condition

4. If $M \in \mathcal{M}$ and $N \preccurlyeq M$, then $N \in \mathcal{M}$.

Example 2.17. The finite aperiodic monoids form a variety: if M is aperiodic and N is a submonoid of M, then N is also aperiodic. If M is aperiodic and $\varphi : M \to N$ is surjective, then N is aperiodic, since, for every $y \in N$ there is an $x \in M$ s.t. $\varphi(x) = y$ and thus $y^n = \varphi(x)^n = \varphi(x^n)$ and thus $x^{k+1} = x^k$ implies $y^{k+1} = y^k$. If $M_1, \ldots, M_n$ are aperiodic, then $\prod_{i=1}^{n} M_i$ is aperiodic too.

Note that we are considering only finite monoids in the above definition. Therefore, and in contrast to the notion of variety considered in Birkhoff's theorem in universal algebra, we only demand closure under finite products. As in the case of Birkhoff's theorem, varieties in our sense also permit a characterisation in terms of (a certain kind of) equations but we will not go into this topic here.

Definition 2.23. A class of regular languages is a function $\mathcal{L}$ which maps each alphabet A to a set $\mathcal{L}_A$ of regular languages.

Definition 2.24. A variety of regular languages is a class $\mathcal{L}$ of regular languages which satisfies the following conditions for all alphabets A and B:

1. $\mathcal{L}_A$ is a closed under finite union, finite intersection and complement.
2. For every $L \in \mathcal{L}_A$ and every $x \in A$: $x^{-1}L \in \mathcal{L}_A$ and $Lx^{-1} = \{w \in A^* \mid wx \in L\} \in \mathcal{L}_A$.
3. For every $L \in \mathcal{L}_A$ and every homomorphism $\varphi : B^* \to A^*$: $\varphi^{-1}(L) \in \mathcal{L}_B$.

From now on we simply speak about a "variety of languages" instead of the longer "variety of regular languages" and a "variety of monoids" instead of the longer "variety of finite monoids".

Definition 2.25. For a variety of monoids $\mathcal{M}$ and an alphabet A we define

$$\Phi(\mathcal{M})_A = \{L \subseteq A^* \mid M(L) \in \mathcal{M}\}$$

We have thus defined a mapping Φ from the varieties of monoids to classes of regular languages. The main result of this section is:

Theorem 2.7 (Eilenberg)**.** *Φ is a bijection between the varieties of monoids and the varieties of languages.*

Before we start to prove this result we make some preliminary observations.

Example 2.18. Let $\mathcal{AP}$ be the variety of aperiodic monoids, then, by Schützenberger's theorem, $\Phi(\mathcal{AP})$ is the class of star-free languages, i.e., for every alphabet A, $\Phi(\mathcal{AP})_A$ is the set of star-free languages over A.

Lemma 2.19. *Let $\mathcal{M}$ be a variety of monoids and let A be an alphabet. Then*

$$\Phi(\mathcal{M})_A = \{L \subseteq A^* \mid \textit{there is a monoid } M \in \mathcal{M} \textit{ which recognises } L\}.$$

Proof. If $L \in \Phi(\mathcal{M})_A$ then $M(L) \in \mathcal{M}$ and, by Corollary 2.1, $M(L)$ recognises L. On the other hand, if there is an $M \in \mathcal{M}$ which recognises L, then, by Theorem 2.5, $M(L) \preccurlyeq M$ and therefore $M(L) \in \mathcal{M}$ by the closure properties of a variety. Hence $L \in \Phi(\mathcal{M})_A$. □

Lemma 2.20. *Let $\mathcal{M}$ be a variety of monoids, then $\Phi(\mathcal{M})$ is a variety of languages.*

Proof. Fix an alphabet A. Let $L_1, L_2 \in \Phi(\mathcal{M})_A$, let $\eta_1 : A^* \to M(L_1)$ and $\eta_2 : A^* \to M(L_2)$ be the syntactic homomorphisms of L_1 and L_2, then $L_1 = \eta_1^{-1}(\eta_1(L_1))$ and $L_2 = \eta_2^{-1}(\eta_2(L_2))$. Define $M = M(L_1) \times M(L_2)$ and let $\eta : A^* \to M, w \mapsto (\eta_1(w), \eta_2(w))$. Then

$$L_1 \cap L_2 = \eta^{-1}(\eta_1(L_1) \times \eta_2(L_2)).$$

Let $L \in \Phi(\mathcal{M})_A$, let $\eta : A^* \to M(L)$ be the syntactic homomorphism, then $\eta^{-1}(\eta(L)) = L$. Then $\eta^{-1}(M(L)\backslash\eta(L)) = A^*\backslash L$ so $M(L) \in \mathcal{M}$ recognises $A^*\backslash L$. Let $x \in A$ and define $P = \{m \in M(L) \mid \eta(x)m \in \eta(L)\}$. Then

$$\begin{aligned}\eta^{-1}(P) = \{w \in A^* \mid \eta(w) \in P\} &= \{w \in A^* \mid \eta(x)\eta(w) \in \eta(L)\} = \\ = \{w \in A^* \mid \eta(xw) \in \eta(L)\} &= \{w \in A^* \mid xw \in L\} = x^{-1}L,\end{aligned}$$

so $M(L)$ recognises $x^{-1}L$. The proof for Lx^{-1} is analogous. Let B be an alphabet and $\varphi : B^* \to A^*$ be a homomorphism. Then $\psi = \eta \circ \varphi : B^* \to M(L)$ and $\psi^{-1}(\eta(L)) = \varphi^{-1}(\eta^{-1}(\eta(L))) = \varphi^{-1}(L)$, so $M(L)$ recognises $\varphi^{-1}(L)$. □

The above lemma shows that Φ, as a mapping from the varieties of monoids to the *varieties* of languages is well-defined. Before proving injectivity of Φ, we need some lemmas.

Lemma 2.21. *Let M be a monoid, let $\approx_1$ and $\approx_2$ be congruences on M s.t. $\approx_1$ is a refinement of $\approx_2$. Then ${}^{M}/_{\approx_2}$ is a quotient of ${}^{M}/_{\approx_1}$.*

Proof. The homomorphism $\varphi : {}^{M}/_{\approx_1} \to {}^{M}/_{\approx_2}, [m]_{\approx_1} \mapsto [m]_{\approx_2}$ is surjective. □

Lemma 2.22. *Let M be a monoid and $(\approx_i)_{i\in I}$ be a family of congruences on M. Define $m \approx n$ by $m \approx_i n$ for all $i \in I$. Then ${}^{M}/_{\approx}$ is isomorphic to a submonoid of $\prod_{i\in I} {}^{M}/_{\approx_i}$.*

Proof. For $i \in I$ let $\pi_i : M \to {}^{M}/_{\approx_i}, m \mapsto [m]_{\approx_i}$. Moreover, let $\pi : M \to \prod_{i\in I} {}^{M}/_{\approx_i}, m \mapsto (\pi_i(m))_{i\in I}$. Then $m \approx n$ iff $m \approx_\pi n$. Therefore ${}^{M}/_{\approx} = {}^{M}/_{\approx_\pi} \simeq \pi(M)$ which is a submonoid of $\prod_{i\in I} {}^{M}/_{\approx_i}$. □

Lemma 2.23. *Let $\mathcal{M}$ be a variety of monoids and let $M \in \mathcal{M}$. Then there is an alphabet A and languages $L_1, \ldots, L_n \in \Phi(\mathcal{M})_A$ s.t. $M \preccurlyeq \prod_{i=1}^n M(L_i)$.*

Proof. Let $A = M$ and $\varphi : M^* \to M$ be the homomorphism induced by the identity mapping. For $m \in M$ the language $L_m = \varphi^{-1}(m)$ is recognised by M and thus $L_m \in \Phi(\mathcal{M})_M$. Let $n = |M|$ and $L_1, \ldots, L_n$ be the languages L_m for $m \in M$.

For $v, w \in M^*$ define $v \approx w$ by $v \approx_{L_m} w$ for all $m \in M$. Then $\approx$ is a congruence relation. If $v \approx w$, then $v \approx_{L_{\varphi(v)}} w$ so $\varepsilon v \varepsilon \in L_{\varphi(v)}$ iff $\varepsilon w \varepsilon \in L_{\varphi(v)}$. Since $v \in L_{\varphi(v)} = \varphi^{-1}(\varphi(v))$, $v \approx w$ implies $w \in L_{\varphi(v)} = \varphi^{-1}(\varphi(v))$, i.e., $\varphi(v) = \varphi(w)$, i.e., $v \approx_\varphi w$.

Now $\varphi(M^*) = M \simeq {}^{M^*}/_{\approx_\varphi}$. Since $\approx$ is a refinement of $\approx_\varphi$, ${}^{M^*}/_{\approx_\varphi}$ is a quotient of ${}^{M^*}/_{\approx}$ by Lemma 2.21. Moreover, by Lemma 2.22, ${}^{A^*}/_{\approx}$ is isomorphic to a submonoid of $\prod_{m\in M} {}^{M^*}/_{\approx_{L_m}}$ which is, by definition, $\prod_{m\in M} M(L_m)$. Thus we have obtained $M \preccurlyeq \prod_{i=1}^n M(L_n)$. □

Lemma 2.24. *Let $\mathcal{M}$ and $\mathcal{N}$ be varieties of finite monoids. Then $\mathcal{M} \subseteq \mathcal{N}$ iff, for every alphabet A, $\Phi(\mathcal{M})_A \subseteq \Phi(\mathcal{N})_A$. In particular, $\mathcal{M} = \mathcal{N}$ iff, for every alphabet A, $\Phi(\mathcal{M})_A = \Phi(\mathcal{N})_A$.*

Proof. If $\mathcal{M} \subseteq \mathcal{N}$ then $\Phi(\mathcal{M})_A \subseteq \Phi(\mathcal{N})_A$ by definition. For the other direction, suppose that $\Phi(\mathcal{M})_A \subseteq \Phi(\mathcal{N})_A$ for every alphabet A and let $M \in \mathcal{M}$. Then, by Lemma 2.23, there is an alphabet A and languages $L_1, \ldots, L_n \in \Phi(\mathcal{M})_A \subseteq \Phi(\mathcal{N})_A$ s.t. $M \preccurlyeq \prod_{i=1}^n M(L_i)$. Then $M(L_1), \ldots, M(L_n) \in \mathcal{N}$ and thus $M \in \mathcal{N}$. □

This shows injectivity of Φ. We now turn to showing surjectivity of Φ.

Lemma 2.25. *Let $\mathcal{L}$ be a variety of languages, let A be an alphabet, let $L \in \mathcal{L}_A$ and let $\eta : A^* \to M(L), w \mapsto [w]_{\approx_L}$. Then, for every $m \in M(L)$ we have $\eta^{-1}(m) \in \mathcal{L}_A$.*

Proof. For $w \in A^*$ define

$$C(w) = \{(u,v) \in A^* \times A^* \mid uwv \in L\} = \{(u,v) \in A^* \times A^* \mid w \in u^{-1}Lv^{-1}\}.$$

Then we have

$$w \approx_L w' \text{ iff } \forall u, v \in A^* : (uwv \in L \Leftrightarrow uw'v \in L) \text{ iff } C(w) = C(w').$$

Therefore

$$[w]_{\approx_L} = \bigcap_{(u,v)\in C(w)} u^{-1}Lv^{-1} \cap \bigcap_{(u,v)\notin C(w)} (u^{-1}Lv^{-1})^{\mathrm{c}}.$$

Since $L \in \mathcal{L}_A$ also $u^{-1}Lv^{-1} \in \mathcal{L}_A$. Since L is recognisable, $\approx_L$ has finite index, so there are only finitely many sets of the form $u^{-1}Lv^{-1}$ for $u, v \in A^*$. Since $\mathcal{L}_A$ is closed under Boolean operations we have $[w]_{\approx_L} \in \mathcal{L}_A$.

Let $m \in M(L)$ and $w \in A^*$ with $\eta(w) = m$, then $\eta^{-1}(m) = \eta^{-1}(\eta(w)) = [w]_{\approx_L} \in \mathcal{L}_A$. □

Lemma 2.26. *For every variety of languages $\mathcal{L}$ there is a variety of monoids $\mathcal{M}$ s.t. $\Phi(\mathcal{M}) = \mathcal{L}$.*

Proof. Let $\mathcal{L}$ be a variety of languages. Let $\mathcal{M}$ be the variety of monoids generated by $\{M(L) \mid L \in \mathcal{L}_A, A \text{ alphabet}\}$. We will show that $\Phi(\mathcal{M})_A = \mathcal{L}_A$ for all alphabets A. If $L \in \mathcal{L}_A$, then $M(L) \in \mathcal{M}$ and thus $L \in \Phi(\mathcal{M})_A$ by definition.

For the other direction let $L \in \Phi(\mathcal{M})_A$, then $M(L) \in \mathcal{M}$, so, by Lemma 2.23, there is an alphabet B and languages $L_1, \dots, L_n \in \Phi(\mathcal{M})_B$ s.t. $M(L) \preccurlyeq \prod_{i=1}^{n} M(L_i) =: M$. By Theorem 2.5 the monoid M recognises L, i.e., there is a homorphism $\varphi : A^* \to M$ and a $P \subseteq M$ s.t. $L = \varphi^{-1}(P)$. Let $\pi_i : M \to M(L_i), (m_1, \dots, m_n) \mapsto m_i$ and let $\varphi_i = \varphi \circ \pi_i$. Let $\eta_i : B^* \to M(L_i)$ be the syntactic homomorphism of L_i. Since η_i is surjective, there is a homomorphism $\psi_i : A^* \to B^*$ s.t. $\varphi_i = \psi_i \circ \eta_i$. Therefore the following diagram commutes:

$$\begin{tikzcd}
A^* \arrow[r, "\psi_i"] \arrow[d, "\varphi"'] \arrow[dr, "\varphi_i"] & B^* \arrow[d, "\eta_i"] \\
M \arrow[r, "\pi_i"] & M(L_i)
\end{tikzcd}$$

We want to show that $L \in \mathcal{L}_A$. We have $L = \bigcup_{m \in P} \varphi^{-1}(m)$ and, since $\mathcal{L}_A$ is closed under union, it suffices to show $\varphi^{-1}(m) \in \mathcal{L}_A$ for all $m \in M$.

Let $m = (m_1, \dots, m_n)$, then

$$w \in \varphi^{-1}(m) = \varphi^{-1}((m_1, \dots, m_n)) \Leftrightarrow \forall i \in \{1, \dots, n\}\, w \in \varphi_i^{-1}(m_i) \Leftrightarrow w \in \bigcap_{i=1}^{n} \varphi_i^{-1}(m_i)$$

so $\varphi^{-1}(m) = \bigcap_{i=1}^{n} \varphi_i^{-1}(m_i)$ and, since $\mathcal{L}_A$ is closed under intersection, it suffices to show $\varphi_i^{-1}(m) \in \mathcal{L}_A$ for all $m_i \in M(L_i)$.

We have $\varphi_i^{-1}(m_i) = \psi_i^{-1}(\eta_i^{-1}(m_i))$ and, since $\psi_i : A^* \to B^*$ and $\mathcal{L}$ is closed under inverse homomorphism, it suffices to show that $\eta_i^{-1}(m_i) \in \mathcal{L}_B$. This follows immediately from Lemma 2.25.

□

Proof of Theorem 2.7. Φ is well-defined by Lemma 2.20, injective by Lemma 2.24, and surjective by Lemma 2.26. □

Example 2.19. Let $\mathcal{CG}$ be the class of finite commutative groups. It is easy to verify that $\mathcal{CG}$ is a variety of monoids. So $\Phi(\mathcal{CG})$ is a variety of languages. For an alphabet A, $x \in A$, $m \geqslant 1$, and $0 \leqslant k < m$ define

$$L(x, k, m) = \{w \in A^* \mid n_x(w) \equiv k \pmod{m}\}.$$

One can show that the variety of languages $\Phi(\mathcal{CG})_A$ is obtained from taking all Boolean combinations of all languages of the form $L(x, k, m)$.

Chapter 3

Further topics

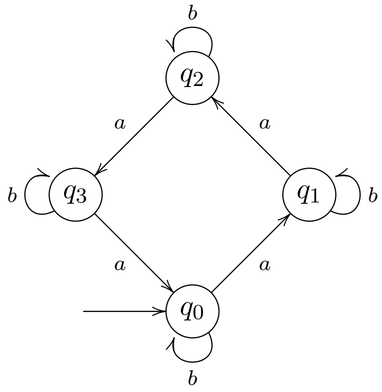
3.1 Automatic sequences

3.1.1 DFAs with output

Definition 3.1. A *deterministic finite automaton with output (DFAO)* is a tuple $D = \langle Q, \Sigma, \cdot, q_0, \Delta, \tau\rangle$ where $Q, \Sigma, \cdot, q_0$ is defined as for a DFA, Δ is the finite *output alphabet* and $\tau : Q \to \Delta$ is the *output function.*

For a DFAO $D = \langle Q, \Sigma, \cdot, q_0, \Delta, \tau\rangle$ and $x \in \Delta$ we define $L_x(D) = \{w \in \Sigma^* \mid \tau(q_0 \cdot w) = x\}$.

Example 3.1. Consider the following automaton:



We have $q_0 \cdot w = q_i$ iff $n_a(w) \equiv i \pmod 4$. Let $\Delta = \{e, 1, 3\}$ and $\tau(q_0) = \tau(q_2) = e$, $\tau(q_1) = 1$, and $\tau(q_3) = 3$. Then, given a $w \in \{a, b\}^*$, this DFAO determines whether the number of a's in w is even, congruent 1 modulo 4 or congruent 3 modulo 4.

Every DFA can be considered a DFAO with $\Delta = \{Y, N\}$ and $\tau : Q \to \Delta$ defined as

$$\tau(q) = \begin{cases} Y \text{ if } q \text{ is a final state} \\ N \text{ otherwise} \end{cases}$$

Then $L(D) = \{w \in \Sigma^* \mid \tau(q_0 \cdot w) = Y\}$. A DFA hence corresponds to a partition of Σ^* into two regular languages: $L(D)$ and $L(D)^c$. In general, a DFAO $D = \langle Q, \Sigma, \cdot, q_0, \Delta, \tau\rangle$ induces the partition $\Sigma^* = \biguplus_{x\in\Delta} L_x(D)$. A finite partition $L_1 \uplus \cdots \uplus L_n = \Sigma^*$ is called *regular* if all L_i are regular languages.

Theorem 3.1. *A finite partition of Σ^* is regular iff it is induced by a DFAO.*

Proof. For the right-to-left direction, fix $x \in \Delta$ and consider the DFA $D = \langle Q, \Sigma, \cdot, q_0, F_x \rangle$ with $F_x = \{q \in Q \mid \tau(q) = x\}$. Then $L(D) = \{w \in \Sigma^* \mid q_0 \cdot w \in F_x\} = \{w \in \Sigma^* \mid \tau(q_0 \cdot w) = x\}$. Therefore each of the classes and hence the partition is regular.

For the left-to-right direction let $L_1 \uplus \cdots \uplus L_n = \Sigma^*$ be a regular partition. Then for all $i \in \{1, \ldots, n\}$ there is a DFA $D_i = \langle Q_i, \Sigma, \cdot, q_{i,0}, F_i \rangle$ with $L(D_i) = L_i$. W.l.o.g. we assume the Q_i to be disjoint. Let $Q = Q_1 \times \cdots \times Q_n$, let $q_0 = (q_{1,0}, \ldots, q_{n,0})$ and for $x \in \Sigma$ and $(q_1, \ldots, q_n) \in Q$ define

$$(q_1, \ldots, q_n) \cdot x = (q_1 \cdot x, \ldots, q_n \cdot x).$$

Let $Q' = \{q \in Q \mid \exists w \in \Sigma^* \text{ s.t. } q_0 \cdot w = q\}$. Now we claim that for every $(q_1, \ldots, q_n) \in Q'$ there is exactly one i s.t. $q_i \in F_i$. There is at least one because $(q_1, \ldots, q_n)$ is reachable by a $w \in \Sigma^*$ and w must be in one of the L_i. On the other hand w can also be in at most one of the L_i since the L_i are disjoint. Let $\Delta = \{1, \ldots, n\}$ and define $\tau : Q' \to \Delta$ by letting $\tau((q_1, \ldots, q_n))$ be this unique i. We define the DFAO $D = \langle Q', \Sigma, \cdot, q_0, \Delta, \tau \rangle$. Then for all $i \in \{1, \ldots, n\}$ we have

$$L_i(D) = \{w \in \Sigma^* \mid \tau(q_0 \cdot w) = i\} = \{w \in \Sigma^* \mid (q_0 \cdot w)_i \in F_i\} = \{w \in \Sigma^* \mid q_{i,0} \cdot w \in F_i\} = L(D_i) = L_i.$$

□

3.1.2 k-automatic sequences

First we need to fix some notation about the base-k representation of natural numbers. It is well known that, for fixed $k \geqslant 2$, every $n \in \mathbb{N}$ can be written as

$$n = \sum_{i=0}^{r} a_i k^i \text{ with } 0 \leqslant a_i < k.$$

Defining $\Sigma_k = \{0, \ldots, k-1\}$ we can consider $a_r \cdots a_0$ as a word in Σ_k^*. For $w = a_r \cdots a_0 \in \Sigma_k^*$ we write $[w]_k$ for the natural number n defined as above. Note that the base k representation is not unique due to the possibility of adding leading zeros, i.e., the function $[\cdot]_k$ is not injective. However, each n permits a unique representation without leading zeros, i.e., for every $n \in \mathbb{N}$ there is exactly one representation of the form

$$n = \sum_{i=0}^{r} a_i k^i \text{ with } 0 \leqslant a_i < k \text{ and } a_r \neq 0.$$

For $n \in \mathbb{N}$ we write $(n)_k$ for the unique word $w = a_r \cdots a_0 \in \Sigma_k$ with $[w]_k = n$ and $a_r \neq 0$. Note that $n = 0$ is represented by the empty word $\varepsilon \in \Sigma_k^*$ since the empty sum is 0. Hence $(0)_k = \varepsilon$.

We will consider infinite words, i.e., infinite sequences over a (finite) alphabet:

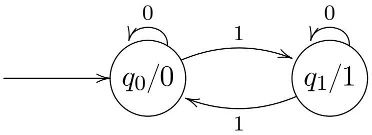
Definition 3.2. For an alphabet Σ, define $\Sigma^\omega = \{(a_n)_{n \geqslant 0} \mid a_n \in \Sigma\}$.

Definition 3.3. A sequence $(a_n)_{n \geqslant 0} \in \Delta^\omega$ is called *k-automatic* if there is a DFAO $D = \langle Q, \Sigma_k, \cdot, q_0, \Delta, \tau \rangle$ s.t. $a_n = \tau(q_0 \cdot w)$ for all $n \geqslant 0$ and w with $[w]_k = n$.

Example 3.2. The Thue-Morse sequence $(t_n)_{n \geqslant 0}$ is defined by letting t_n be the number of 1's modulo 2 in the binary representation of n. Its first few elements are:

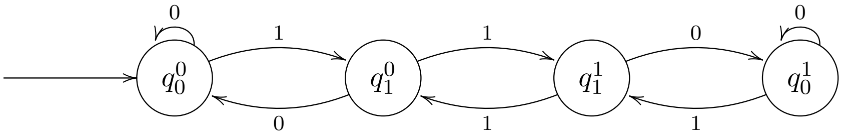
n	$=$	0	1	2	3	4	5	6	7	8	9	10	$\cdots$
t_n	$=$	0	1	1	0	1	0	0	1	1	0	0	$\cdots$

The Thue-Morse sequence is 2-automatic since it is generated by the following DFAO:



where the notation q/i represents $\tau(q) = i$. The Thue-Morse sequence is one of the most famous automatic sequences and has numerous interesting properties. For example, it is overlap-free (i.e. it does not contain a subword of the form $xwxwx$ for $x \in \{0,1\}, w \in \{0,1\}^*$) which can be used for constructing a square-free word over $\{0,1,2\}$.

Example 3.3. Define $e_{2;11}(n)$ as the number of (possibly overlapping) occurrences of the word 11 in the binary representation of n. For example, the word 1101110 contains three occurrences of the word 11. Then the Rudin-Shapiro sequence $(r_n)_{n\geqslant 0}$ is defined as $r_n = (-1)^{e_{2;11}(n)}$. The Rudin-Shapiro sequence is 2-automatic since it is generated by the following DFAO:



with $\tau(q_i^j) = (-1)^j$. Here the subscript denotes the last letter seen and the superscript denotes the number of occurrences of 11 modulo 2. As one can easily verify, the transitions preserve these properties. An interesting property of r_n is that it defines a space-filling lattice walk. Let

$$d_{n+1} = \begin{cases} \text{R} & \text{if } r_{n+1}r_n = (-1)^n \\ \text{L} & \text{if } r_{n+1}r_n = (-1)^{n+1} \end{cases}$$

The first few elements of $(r_n)_{n\geqslant 0}$ and $(d_n)_{n\geqslant 0}$ are:

$$\begin{array}{rcccccccccccccc} n & = & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & \cdots \\ r_n & = & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & \cdots \\ d_{n+1} & = & \text{R} & \text{L} & \text{L} & \text{R} & \text{R} & \text{R} & \text{L} & \text{L} & \text{R} & \text{L} & \cdots & \end{array}$$

Define a walk in $\mathbb{Z} \times \mathbb{Z}$ by starting at $(0,0)$ making a first step to $(0,1)$ and then turning left or right according to the value of d_n. The result is a walk that contains all points in the first quadrant which are above (or on) the diagonal.

A k-automatic sequence corresponds to a partition of Σ_k^* in the following sense:

Proposition 3.1. *A sequence $(a_n)_{n\geqslant 0} \in \Delta^\omega$ is k-automatic iff $\biguplus_{d\in\Delta}\{w \in \Sigma_k^* \mid a_{[w]_k} = d\}$ is a regular partition of Σ_k^*.*

Proof. $(a_n)_{n\geqslant 0}$ is k-automatic iff there is a DFAO $D = \langle Q, \Sigma_k, \cdot, q_0, \Delta, \tau\rangle$ s.t. $a_n = \tau(q_0 \cdot w)$ for all $w \in \Sigma_k^*$ with $[w]_k = n$. In other words we require that $L_d(D) = \{w \in \Sigma_k^* \mid a_{[w]_k} = d\}$. Then Theorem 3.1 entails that $(a_n)_{n\geqslant 0}$ is k-automatic iff $\biguplus_{d\in\Delta} = \{w \in \Sigma_k^* \mid a_{[w]_k} = d\}$ is a regular partition of Σ_k^*. □

For defining the notion of k-automatic sequence, it suffices to consider number representations without leading zeros.

Lemma 3.1. *Let $(a_n)_{n\geqslant 0} \in \Delta^\omega$ and let $D = \langle Q, \Sigma_k, \cdot, q_0, \Delta, \tau\rangle$ s.t. $a_n = \tau(q_0 \cdot (n)_k)$ for all $n \geqslant 0$, then there is a DFAO $D' = \langle Q', \Sigma_k, \cdot, q_0', \Delta, \tau'\rangle$ s.t. $a_n = \tau(q_0' \cdot w)$ for all $w \in \Sigma_k^*$ with $[w]_k = n$ and all $n \geqslant 0$. Moreover, we have $q_0' \cdot 0 = q_0'$.*

Proof. Let q_0' be a new state and define $Q' = Q \cup \{q_0'\}$ and $q_0' \cdot 0 = q_0'$ and $q_0' \cdot x = q_0 \cdot x$ for all $x \in \Sigma_k \backslash \{0\}$. The output function τ' is defined by $\tau'|_Q = \tau$ and $\tau'(q_0') = \tau(q_0)$.

Then we claim that $\tau'(q_0' \cdot 0^i(n)_k) = a_n$ for all $i \geqslant 0$. First note that we have $\tau'(q_0' \cdot 0^i(n)_k) = \tau'(q_0' \cdot (n)_k)$ because $q_0' \cdot 0 = q_0'$. If $n = 0$, then $\tau'(q_0' \cdot (n)_k) = \tau'(q_0') = \tau(q_0) = a_0$. If $n \neq 0$, then $(n)_k$ starts with an $x \in \Sigma_k \backslash \{0\}$ and hence $\tau'(q_0' \cdot (n)_k) = \tau'(q_0 \cdot (n)_k) = \tau(q_0 \cdot (n)_k) = a_n$. □

Proposition 3.2. *A sequence $(a_n)_{n \geqslant 0}$ is k-automatic iff there is a DFAO $D = \langle Q, \Sigma_k, \cdot, q_0, \Delta, \tau \rangle$ s.t. $a_n = \tau(q_0 \cdot (n)_k)$ for all $n \geqslant 0$.*

Proof. The left-to-right direction is trivial. The right-to-left direction follows from Lemma 3.1. □

3.1.3 Morphic sequences

A homomorphism $\varphi : \Sigma^* \to \Delta^*$ can be extended to $\varphi : \Sigma^\omega \to \Delta^\omega$ by defining

$$\varphi(a_0 a_1 a_2 \cdots) = \varphi(a_0)\varphi(a_1)\varphi(a_2) \cdots .$$

Note that, since $\varphi(vw) = \varphi(v)\varphi(w)$ every way of splitting $(a_n)_{n \geqslant 0}$ into finite words will lead to the same value for $\varphi((a_n)_{n \geqslant 0})$. A sequence $s \in \Sigma^\omega$ is called *fixed point* of a homomorphism $\varphi : \Sigma^* \to \Sigma^*$ if $\varphi(s) = s$. Let $\varphi : \Sigma^* \to \Sigma^*$ be a homomorphism. If there is an $x \in \Sigma$ s.t. $\varphi(x) = xw$ for some $w \in \Sigma^{k-1}$, then φ is called *prolongable on x*.

Example 3.4. Define $\varphi : \Sigma_2^* \to \Sigma_2^*$ by $\varphi(0) = 01$ and $\varphi(1) = 10$. Then φ is prolongable on both 0 and 1. We have

$$\begin{aligned}
\varphi(0) &= 01 \\
\varphi^2(0) = \varphi(01) &= 0110 \\
\varphi^3(0) = \varphi(0110) &= 01101001 \\
\varphi^4(0) = \varphi(01101001) &= 0110100110010110 \\
&\vdots
\end{aligned}$$

Note that the positions already computed do not change. As we will see now this can be generalised to yield a mechanism for defining an infinite sequence.

Proposition 3.3. *Let $\varphi : \Sigma^* \to \Sigma^*$ be a homomorphism which is prolongable on $a \in \Sigma$, i.e., $\varphi(a) = aw$. Then $\varphi^\omega(a) = aw\varphi(w)\varphi^2(w)\cdots$ is the unique fixed point of φ in Σ^ω that starts with a.*

Proof. First observe that $\varphi^\omega(a)$ is indeed a fixed point of φ since

$$\varphi(\varphi^\omega(a)) = \varphi(aw\varphi(w)\varphi^2(w)\cdots) = \varphi(a)\varphi(w)\varphi^2(w)\varphi^3(w)\cdots = aw\varphi(w)\varphi^2(w)\cdots = \varphi^\omega(a).$$

Furthermore, assume that $s \in \Sigma^\omega$ is a fixed point of φ which starts with a. We claim that s must start with $a\varphi^0(w)\varphi^1(w)\cdots\varphi^l(w)$ for all $l \geqslant -1$. Since this determines all finite prefixes of s it uniquely determines $s = \varphi^\omega(a)$. We proceed by induction on l. The case $l = -1$ follows immediately from the assumption that s starts with a. For the induction step, assume that s starts with $a\varphi^0(w)\varphi^1(w)\cdots\varphi^l(w)$ for some $l \geqslant -1$, then

$$\begin{aligned}
s = \varphi(s) &= \varphi(a\varphi^0(w)\varphi^1(w)\cdots\varphi^l(w)t) \\
&= \varphi(a)\varphi(w)\varphi^2(w)\cdots\varphi^{l+1}(w)\varphi(t) \\
&= a\varphi^0(w)\varphi(w)\cdots\varphi^{l+1}(w)\varphi(t).
\end{aligned}$$

□

In the other direction, note that, if $(s_n)_{n\geqslant 0}$ is a fixed point of a homomorphism φ, then φ must be prolongable on s_0 since $\varphi(s_0s_1s_2\cdots) = \varphi(s_0)\varphi(s_1s_2\cdots) = s_0s_1s_2\cdots$.

Definition 3.4. A homomorphism $\varphi : \Sigma^* \to \Delta^*$ is called k-uniform if $|\varphi(x)| = k$ for all $x \in \Sigma$. A 1-uniform homomorphism is also called a *coding*.

If $\varphi : \Sigma^* \to \Sigma^*$ is a k-uniform homomorphism and $a \in \Sigma$ s.t. φ is prolongable on a, then $\varphi^\omega(a)$ is called *pure morphic sequence*. If, in addition, τ is a coding, then $\tau(\varphi^\omega(a))$ is called *morphic sequence*.

Lemma 3.2. *Let $s = (s_n)_{n\geqslant 0}$ be fixed point of a k-uniform homomorphism φ. Then $\varphi(s_n) = s_{kn}s_{kn+1}\cdots s_{kn+k-1}$ for all $n \geqslant 0$.*

Proof. Since φ is k-uniform, we have $|\varphi(s_0\cdots s_i)| = k(i+1)$ and since $\varphi(s) = s$, we have $\varphi(s_0\cdots s_i) = s_0\cdots s_{ki+k-1}$. Now we have

$$\begin{aligned}\varphi(s_0\cdots s_n) &= s_0\cdots s_{kn+k-1},\\ \varphi(s_0\cdots s_{n-1}) &= s_0\cdots s_{kn-1}, \text{ and}\\ \varphi(s_0\cdots s_n) &= \varphi(s_0\cdots s_{n-1})\varphi(s_n)\end{aligned}$$

and therefore $\varphi(s_n) = s_{kn}\cdots s_{kn+k-1}$. □

Theorem 3.2. *A sequence is k-automatic iff it is morphic sequence of a k-uniform homomorphism.*

Proof. For the left-to-right direction, assume that $(a_n)_{n\geqslant 0}$ is k-automatic. Then there is a DFAO $D = \langle Q, \Sigma_k, \cdot, q_0, \Delta, \tau\rangle$ s.t. $a_n = \tau(q_0 \cdot (n)_k)$. By Lemma 3.1 we can assume that $q_0 \cdot 0 = q_0$. We consider Q as alphabet and define a homomorphism $\varphi : Q^* \to Q^*$ by

$$\varphi(q) = (q\cdot 0)(q\cdot 1)\cdots(q\cdot(k-1)) \qquad \text{for each } q \in Q.$$

Now φ is k-uniform and prolongable on q_0, so by Proposition 3.3 we know that $(s_n)_{n\geqslant 0} = \varphi^\omega(q_0)$ is a fixed point of φ. We claim that $s_{[w]_k} = q_0 \cdot w$ for all $w \in \Sigma_k^*$. We proceed by induction on $|w|$: if $w = \varepsilon$, then $q_0 \cdot w = q_0 = s_{[\varepsilon]_k} = q_0$ since s starts with q_0. For the induction step, let $w = vx$ with $x \in \Sigma_k$. Then

$$q_0 \cdot w = q_0 \cdot vx =^{\text{IH}} s_{[v]_k} \cdot x =^{\text{Def. } \varphi} \varphi(s_{[v]_k})_x =^{\text{Lem. 3.2}} s_{k[v]_k+x} = s_{[vx]_k} = s_{[w]_k}.$$

Therefore $s_n = q_0 \cdot (n)_k$ and hence $\tau(s_n) = \tau(q_0 \cdot (n)_k) = a_n$.

For the right-to-left direction, let $\varphi : Q^* \to Q^*$ be a k-uniform homomorphism, let $s \in Q^\omega$ with $s = \varphi(s)$, $\tau : Q \to \Delta$ be a coding and let $a = \tau(s)$, i.e., $a_n = \tau(s_n)$ for all $n \geqslant 0$. Define the DFAO $\langle Q, \Sigma_k, \cdot, s_0, \Delta, \tau\rangle$ where $q \cdot i$ is the i-th letter of $\varphi(q)$. We claim that $s_0 \cdot (n)_k = s_n$ for all $n \geqslant 0$. We proceed by induction on n: if $n = 0$, then $s_0 \cdot (n)_k = s_0 \cdot \varepsilon = s_0$. For the induction step, let $n = kn' + d$ with $0 \leqslant d < k$. Then

$$s_0 \cdot (n)_k = (s_0 \cdot (n')_k) \cdot d =^{\text{IH}} s_{n'} \cdot d =^{\text{Def. } D} \varphi(s_{n'})_d =^{\text{Lem. 3.2}} s_{kn'+d} = s_n.$$

Therefore $a_n = \tau(s_n) = \tau(s_0 \cdot (n)_k)$. □