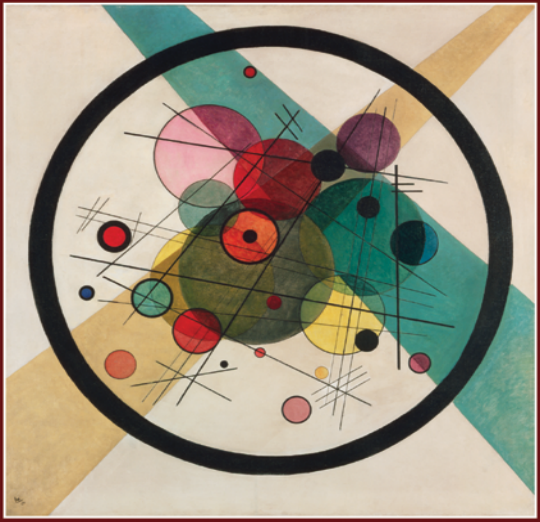
ELEMENTS OF AUTOMATA THEORY



JACQUES SAKAROVITCH

CAMBRIDGE

CAMBRIDGE

more information – www.cambridge.org/9780521844253

ELEMENTS OF AUTOMATA THEORY

This page intentionally left blank

ELEMENTS OF AUTOMATA THEORY

JACQUES SAKAROVITCH
Centre National de la Recherche Scientifique (CNRS),
and École Nationale Supérieure des Télécommunications (ENST), Paris

Translated by

REUBEN THOMAS

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521844253

Originally published in French as *Éléments de théorie des automates* by Vuibert, Paris, 2003.
© Vuibert, Paris 2003

First published in English by Cambridge University Press 2009
English translation © Cambridge University Press 2009

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-84425-3 hardback

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party Internet websites referred to
in this publication, and does not guarantee that any content on such
websites is, or will remain, accurate or appropriate.

To the memory of my father,

Jean Sakarovitch,
architect,

who gave me a taste for figures and from whom I have perhaps inherited a certain tendency to return to them again and again.

To the memory of his grandson,

Mathias,

my nephew, gone so soon.

This page intentionally left blank

TABLE OF CONTENTS

Preface to the English edition xv

Preface xvii

M. Pascal's division machine 1

0 Fundamental structures 7
- 1 Relations . . . 9
- 2 Monoids . . . 14
- 3 Words and languages . . . 18
- 4 Free monoids . . . 24
- 5 Semirings . . . 27
- 6 Matrices . . . 30
- 7 Lexicon of graph theory . . . 33
- 8 Complexity and decidability . . . 34
- Solutions to the exercises . . . 39
- Notes & references . . . 46

The three stages of rationality

I The simplest possible machine... 49
- 1 What is an 'automaton'? . . . 51
  - 1.1 First definitions – first examples . . . 51
    States, transitions, etc. – Computations, recognised languages etc. – Transposition and left–right duality
  - 1.2 Basic constructions, basic properties . . . 60
    Union – Cartesian product – Quotient (of a language)
  - 1.3 The graph perspective . . . 66
    Trim automata – The empty and the infinite – Criteria for recognisability
  - 1.4 Some supplementary definitions . . . 74
    Unambiguous automata – Complete automata – Deterministic automata – Automata with spontaneous transitions
- 2 Rational languages . . . 82
  - 2.1 Rational operations . . . 82
    Product of languages – Star of a language – Rational operations
  - 2.2 Rational languages . . . 86

2.3 Rational is recognisable 87
Normalised automata – Closure under product and star – Standard automata
2.4 Recognisable is rational 94
The McNaughton–Yamada algorithm, or algorithm MNY– The state elimination method – Solving equations
3 The functional perspective 101
3.1 From transitions to the transition function 102
3.2 Deterministic automata 104
Reformulation of the definition – Determinisation – The case of one-letter alphabets – Complement of recognisable languages
3.3 Minimisation 111
The automaton of quotients of a language. . .– . . .is minimal – Computation of the minimal automaton – Another minimisation method
3.4 Return to the Star Lemma 118
Block iteration and block simplification – Ramsey's Theorem – Proof of Theorem 3.3
4 Rational expressions 123
4.1 Rational expressions and languages 124
Rational expressions over an alphabet – Rational expressions over a set of variables
4.2 Rational identities 128
Classical identities – A formal computation
4.3 Expressions for the behaviour of a finite automaton 133
The state elimination and equation solution methods – The BMC and MNY algorithms, identical orders – The BMC and MNY algorithms, distinct orders
4.4 Derivation of expressions 138
Derivatives of an expression – A theorem of J. Brzozowski – Derivative automaton
5 From expressions to automata 145
5.1 The standard automaton of an expression 145
Direct construction – Thompson's construction
5.2 The derived term automaton 149
Derived terms – A theorem of V. Antimirov
5.3 String matching 152
Automaton for finding a word – Searching by sliding window – Implementation with a default successor
6 Star height 157
6.1 Two heights and a degree 158
Star height of an expression – Star height of a language – Loop complexity of an automaton
6.2 Eggan's Theorem 162
From expressions to automata – From automata to expressions: calculating the index– Not so fast
6.3 An infinite hierarchy 167
6.4 Generalised star height 170
7 A field of automata 171

7.1 The Rabin–Scott model 171
7.2 Two-way automaton 172
7.3 Moore and Mealy machines 174
8 A crop of properties 175
Solutions to the exercises 179
Notes & references 214

II The power of algebra 217
1 Automata and rational sets 219
1.1 Automata over a monoid 219
1.2 Rational sets 220
The semiring $\mathfrak{P}(M)$ – Rational operations and subsets – Rational expressions – Image under morphism – Intersection and inverse morphism
1.3 Behaviour of finite automata 225
1.4 Unambiguous rational sets 228
Definitions – The family URat
2 Actions and recognisable sets 231
2.1 Actions on a set 232
Definition – Matrix representation of actions – Subsets recognised by an action
2.2 Recognisable here, recognisable there 238
Consistency – Kleene's Theorem – Automaton of an action
2.3 Elementary operations on recognisable subsets 243
Boolean operations – Inverse morphism – Quotient – Morphism and product
2.4 Minimisation 246
Action morphisms – Minimal action – Syntactic congruence and monoid
2.5 Algebra at work 251
Two examples – Recognisable subsets included in a product
3 Morphisms and coverings 255
3.1 Automata morphisms 255
Definitions and examples – Conformal morphisms – Local properties
3.2 Quotients of automata 261
Out-surjective morphisms – Totally surjective morphisms – Moore's algorithm
3.3 Automata coverings 264
From local to global – Product of an automaton with an action – The Coloured Transition Lemma
3.4 The Schützenberger covering 270
4 Universal automaton 273

4.1 Factorisations . . . 275
2-factorisations – Sub-factorisations and factorisations – Morphisms and factorisations
4.2 Universal automata of a subset . . . 279
Definitions and examples – Properties – Universal automaton relative to a generating set – Universality of universal automata
4.3 Construction of the universal automaton . . . 286
Expansion of a deterministic automaton – Extraction of the universal automaton
4.4 Language approximations . . . 291
5 The importance of being well ordered . . . 293
5.1 Well quasi-orderings . . . 293
5.2 Derivations . . . 297
Preparations – Proof of Theorem 5.4
6 Rationals in the free group . . . 301
6.1 Recognisable and rational in groups . . . 301
Recognisable subsets – Rational subgroups – Fatou property
6.2 Description of the free group . . . 305
Dyck congruence and Dyck words – Shamir congruence and parenthetic words – Simplifications – Reduction associated with a simplification – Unambiguous factorisation induced by a reduction
6.3 Rationals of the free group . . . 314
Rationals of simplification monoids – Return to the free group
6.4 Büchi systems . . . 319
7 Rationals in commutative monoids . . . 323
7.1 The natural order on $A^{\oplus}$ . . . 323
The free commutative monoid – Dickson's Lemma
7.2 The lexicographic order on $\mathbb{N}^k$ . . . 326
Congruences of $\mathbb{N}^k$ – Lexicographic decomposition
7.3 Subtractive submonoids and affine sets . . . 330
7.4 Semi-linear and semi-simple sets . . . 333
7.5 Rationals of $\mathbb{N}^k$ . . . 335
The Freedom Lemma – Positive solutions of Diophantine linear systems – Semi-simple subsets of $\mathbb{Z}^k$ – Proof of Theorems 7.3 and 7.4
7.6 Rationals of commutative monoids . . . 341
8 Star height of group languages . . . 342
Solutions to the exercises . . . 348
Notes & references . . . 372

III The pertinence of enumeration 375
1 Formal power series on a graded monoid . . . 379
1.1 Formal power series over M with coefficients in $\mathbb{K}$ . . . 379
Operations on $\mathbb{K}\langle\langle M\rangle\rangle$ – Support of a series – characteristic series – Hadamard product – Scalar product
1.2 Graded monoids . . . 383

1.3 Topology on $\mathbb{K}\langle\langle M\rangle\rangle$ . . . 385
Distance – Distance on $\mathbb{K}\langle\langle M\rangle\rangle$ – Summable families – Continuous morphisms
2 $\mathbb{K}$-automata and $\mathbb{K}$-rational power series . . . 392
2.1 Star of a power series . . . 393
Star in a topological semiring – Star of a proper series – Star of an arbitrary series
2.2 $\mathbb{K}$-rational series . . . 398
$\mathbb{K}$-rational operations – Rational $\mathbb{K}$-expressions – Star of a matrix
2.3 Weighted automaton in a semiring . . . 402
$\mathbb{K}$-automaton over M – Behaviour of a $\mathbb{K}$-automaton – Notes – Some other definitions and examples
2.4 The Fundamental Theorem of finite automata . . . 409
Proper automata – proper families of series – Statement and proof – Notes and corollaries
2.5 $\mathbb{K}$-coverings – $\mathbb{K}$-quotients . . . 416
From coverings to $\mathbb{K}$-coverings – Matrix description – Co-$\mathbb{K}$-covering – Minimal $\mathbb{K}$-quotient
3 $\mathbb{K}$-recognisable series . . . 424
3.1 $\mathbb{K}$-representations . . . 424
3.2 Products . . . 426
Tensor product of $\mathbb{K}$-representations – Hadamard product – Tensor product of series – Shuffle product
3.3 The Kleene–Schützenberger Theorem . . . 433
4 Series on a free monoid . . . 438
4.1 A characterisation of recognisable series . . . 438
Quotients of series – Stable modules – The Kleene–Schützenberger Theorem revisited
4.2 Derivation of rational $\mathbb{K}$-expressions . . . 443
Polynomials of $\mathbb{K}$-expressions – $\mathbb{K}$-derivatives of a $\mathbb{K}$-expression – Derived terms – The derived term automaton
4.3 Series on a field . . . 451
Rank of a series – Reduced representation – Linear recurrence – Effective computations
4.4 Rational series and their supports . . . 463
Rationality of supports – The Rational Skimming Theorem, I – Undecidable questions
5 Series on an arbitrary monoid . . . 470
5.1 Complete semirings, continuous semirings . . . 470
5.2 Star of a series . . . 472
5.3 $\mathbb{K}$-rational series . . . 474
6 Rational subsets in free products . . . 476
6.1 Free product of monoids . . . 476
6.2 Bipartite automaton over a free product . . . 478
6.3 Bipartite deterministic automaton . . . 482
6.4 Minimal deterministic bipartite automaton . . . 484
7 A non-commutative linear algebra primer . . . 488
Solutions to the exercises . . . 498

Notes & references 519

Rationality in relations

IV The richness of transducers 523
1 Rational relations: an introduction 525
1.1 Rational relations 525
Rational relations between free monoids – Rational relations between arbitrary monoids
1.2 Realisation by automata 529
1.3 Realisation by morphisms 531
Realisation – Evaluation Theorem – Composition Theorem – Star Lemma
1.4 Recognisable relations 539
1.5 Realisation by representation 540
Real-time transducers – From real-time transducers to representations – Theorem of evaluation and composition of representations
1.6 The Rabin–Scott model 545
2 $\mathbb{K}$-relations 546
2.1 Definitions 548
The canonical isomorphism – $\mathbb{K}$-relations – Support of relations – characteristic relations – Continuity
2.2 Composition 553
2.3 Multiplicative $\mathbb{K}$-relations 555
3 Rational $\mathbb{K}$-relations 557
3.1 Reasonable semirings 558
Image of series under continuous morphisms – Image of seres under projections – $\mathbb{K}$-intersections
3.2 Realisation of rational $\mathbb{K}$-relations 561
Realisation by $\mathbb{K}$-automaton – Realisation by $\mathbb{K}$-representation – Realisation by morphisms
3.3 Evaluation and Composition Theorems 564
Using recognition by morphisms – Using recognition by representation
4 Equivalence of finite $\mathbb{K}$-transducers 568
4.1 Equivalence of $\mathbb{B}$-transducers, general case 569
4.2 Equivalence of $\mathbb{B}$-transducers, case of small alphabets 571
4.3 Equivalence of $\mathbb{N}$-transducers 574
5 Deterministic rational relations 577
5.1 Transducers with an endmarker 577
5.2 Deterministic transducers 578
Definition – Uniqueness of computations – Almost an action
5.3 Deterministic relations 584
Definitions – Complement – Iteration Lemma
5.4 Geography of Rat $A^* \times B^*$ I 588

5.5 Matrix representations 590
Representation of a deterministic transducer – Representation of a deterministic relation
5.6 An example: the map equivalence of a morphism 592
6 Synchronisation of transducers 595
6.1 Rational relations of bounded length discrepancy 596
Definitions, notation and conventions – Characterisation of rational bld-relations – Translation into automata theoretic terms, and corollaries
6.2 Transducers of bounded lag 602
Lag in a computation or transducer – Resynchronisation algorithm for transducers – Composition of letter-to-letter transducers
6.3 Synchronous relations 609
Another family of rational relations – Determinisation and minimisation – Geography of $\mathrm{Rat}\, A^* \times B^*$ *II*
7 Malcev–Neumann series 616
7.1 Order on the free group 617
On ordered groups – Representation of the free group – A detour via ordered rings – Order on the free group
7.2 Series on an ordered group 622
The semiring $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ *– Ordered semigroups – The field* $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ *– A last inclusion*
Solutions to the exercises 627
Notes & references 641

V The simplicity of functional transducers 643
1 Functionary 645
1.1 Deciding functionality 645
An effective characterisation of functionality – Equivalence of rational functions
1.2 Sequential functions 651
Some unconventional terminology – Dual definitions – Composition
1.3 Pure sequential functions 658
1.4 Local functions 661
2 Uniformisation of rational relations 664
2.1 Proof of Theorem 2.1 (transducer version) 666
2.2 Proof of Theorem 2.1 (representation version) 667
Represent of S-immersions of an automaton – Semi-monomial matrices – Representation of S-uniformisations
2.3 Decomposition of rational functions 673
The Weak Decomposition Theorem – The Strong Decomposition Theorem
2.4 The Rational Skimming Theorem II 677
3 Cross-section of rational functions 679

3.1 The rational cross-section property 680
The Rational Cross-Section Theorem – The rational cross-section property for a monoid – Return to simplification monoids
3.2 Choosing the uniformisation (or the cross-section) 685
Uniformisation of synchronous relations – Uniformisation of deterministic relations – Th. 3.3 back on the loom
4 Sequential functions 692
4.1 Two characterisations 692
Translations of a function – A functional characterisation – A quasi-topological point of view
4.2 Deciding sequentiality 699
4.3 Minimisation 704
Conjugation – Blockage of a sequential transducer – Reduction – Effective computation
4.4 The (Great) Sequentiality Theorem 711
Differential of a function – Proof of Theorem 4.5 iii) $\Rightarrow$ i) *– Proof of Theorem 4.5* ii) $\Rightarrow$ iii) *– Return to the Sequentiality Theorem*
4.5 Pure sequential functions and local functions 717
Solutions to the exercises 719
Notes & references 737

Bibliography 739

Index 749

Preface to the English edition

As this English edition is, or was intended to be, the direct translation of the French one, although already few years old, there is not much to say in this preface. Nothing but to express my gratitude to all those who have made the volume possible and helped in its realisation.

First, to David Tranah from Cambridge University Press, whom I first met more than twenty years ago and who showed interest in the book on automata I dreamt of writing. He was encouraging when I eventually engaged in that project and supportive when I was in the throes of completing the French edition. He was quick to welcome its publication by Cambridge University Press, and of infinite patience when waiting for the final version of the manuscript.

Under the seal of secrecy, I am glad to confess that I am grateful to Wolfgang Thomas who, in a highly confidential review, warmly recommended that CUP have the French edition translated and published.

I was very lucky, thanks to James Martin, to meet Reuben Thomas, who agreed to translate the book into English. The reader will appreciate, probably even better than I, the fluidity of his English. Even more striking, and I am still the best witness to that, was his talent and eagerness not only to translate the words but also to convey the style in which I had written them. I did not think I could ever find someone not only technically qualified for such a specialised text, but able to render with such virtuosity the literary puns and misquoted poetry with which I could not help peppering the French edition, in order to escape for a second from arid formalism.

I am also grateful to all the colleagues who have had the kindness to send me notes and remarks on the numerous mistakes that, in spite of all my efforts during the long years of writing, still riddle the French edition. The errata are now almost twenty pages long and all the corresponding corrections have been incorporated in the English edition: this is the only difference, in substance, between the two editions.

I want to thank especially Olivier Carton and Christian Choffrut who have been particularly attentive readers, and my student Pierre-Yves Angrand who pointed out two serious errors. Pierre-Jean Bourcet, Sandrine Bouthémy, Marie-Eve Harvie, Sylvain Lombardy, Bruno Petazzoni, Rodrigo de Souza, Nicolas Stroppa, and Reuben Thomas have kindly commented on the text, and Jeffrey Shallit always answered my questions about translations of mathematical terms quickly and cordially.

*

Finally, I would like to mention in this page of gratitude, all that I owe to the loving support of Michèle Lévy-Vicas during these endless years when I was writing the French edition and correcting the English one. When we met, she was unaware that mathematicians are much worse companions than artists: not only are they as busy and absent-minded as the latter but there is nothing in their production that one can appreciate. Her stoicism touches the sublime.

J. S.

Paris, November 2008

Preface

He who kisses many keeps few.

Popular adage

One cannot woo all women,
but one should try.

Another adage

For a long time I would go through, head whirling, the writing of this preface. I would rattle off whole sentences to give myself the heart to work. Now that I have to do it for real, I understand that the task is no easier than the heart of the work. How to justify writing a book on automata theory? Another one! and so thick! Justify? One can always dream; present, perhaps.

*

A shining light of computer science research in the nineteen-sixties, a compulsory part of instruction in the discipline in the seventies and eighties, automata theory seems to have disappeared from lecture theatre and conference hall. Nonetheless, we find it, explicitly or implicitly, in the essence or the premises of a number of subjects in computer science which are currently new or fashionable. As a possible explanation, I suggest that automata theory is the *linear algebra* of computer science. I mean this in two ways. Properly speaking, automata theory *is* non-commutative linear algebra, or can be viewed as such: the theory of matrices with coefficients in suitable algebras. I am more interested, however, in the figurative sense: automata theory as a basic, fundamental subject, known and used by everyone, which has formed part of the intellectual landscape for so long that it is no longer noticed. And yet, there it is, structuring it, organising it: and knowing it allows us to orient ourselves.

Finite automata are the simplest model of machines, so simple that they take forms, appear in contexts, and hide in applications as numerous as they are diverse. I shall not describe the many faces of the model, *nor the applications* of automata theory – at the very most I shall mention a few. I would like to present this theory *for its own sake* and I shall try to show its richness. Nonetheless, this book's girth is not only, or even mostly, due to the problems posed by finite automata, or to the number of results to which they give rise – of which in any case I am far from giving a comprehensive account – but because I wanted to give each property the most direct explanation whilst still placing it in the most general possible context. Simple properties are proved simply, then interpreted as special cases of global propositions

expressed in more abstract terms, which the simpler forms help us to understand. This is why, in the first part of this book, organised around notions of *rationality* and *recognisability*, I tell the same story three times, each time with a different point of view and theoretic apparatus. The material takes on a richer perspective when viewed from these different angles.

The second part deals with relations between words realised by finite automata. This subject exemplifies automata theory, both the variety of its methods and its fields of application. Automata with output are susceptible to an elementary presentation and yet some of their properties involve deep algebraic methods. Their study illustrates the usefulness of each aspect of the theory developed in the first part.

Map

After Chapter 0, which gathers together the definitions of the structures that will be used throughout the book, Chapter I presents a 'naive theory' of finite automata, as taught in all works on automata theory, and which begins – and often ends – with the elementary version of the result called 'Kleene's Theorem'. I forbade myself the use of any other structure than that of the free monoid: not the least morphism or smallest finite semigroup. I present the notion of a rational expression and tackle, as a natural extension of Kleene's Theorem, the problem of the transformation of an expression into an automaton, or, to use a more marketable formulation, searching for a pattern in a text. Elementary does not mean simplistic, and this theory is already bountiful. I have included in this chapter two beautiful combinatorial properties: the necessary and sufficient version of the Star Lemma and the proof that the 'star height' of a rational language can be arbitrarily large.

Chapter II starts the subject anew with algebra as a guide, mainly in the form of the idea of a *morphism* (of monoids or automata). The first consequence of this point of view is the distinction between *action* and *automaton*, between a *recognisable* set and a *rational* set of an arbitrary monoid, which sheds a new light on both the notion of automata and on Kleene's Theorem. The horizon opens so wide that we have to make some choices, and in particular, I do not tackle 'variety theory', which has already been treated in several works. On the other hand, I develop two ideas that are more original: first the notion of a morphism on automata, which enables me to define what I call the 'Schützenberger covering' of an automaton and which I use several times later on; and next the definition of what I call the 'universal automaton' of a language, which is a new presentation of a construction due to Conway and which, among other things, enables me to give a novel treatment of McNaughton's Theorem on the star height of group languages at the end of the chapter. A section is devoted to the *partial well ordering* structure, which is fundamental to the whole of theoretical computer science, and which will be used at two crucial junctures.

I then study the family of rational sets in two structures: the *free group* and the *free commutative monoid*. In both cases this family is an effective Boolean algebra and this explains, at least in part, why it is often in such a setting that finite automata enable us to describe the behaviour of processes whose set of sets is infinite (like

Petri nets, timed automata *etc.*). The free group is likewise the algebraic structure underlying the behaviour of *pushdown automata*, all subjects which are outside the scope of this work.

Chapter III goes back to the beginning once again. The aim is less to generalise even further than to add a dimension by taking into account the *weights* of computations: '(formal) languages' become *formal series*, actions become (matrix) representations. The substance thus given to the results helps us better understand their deep nature, and connects them with more classical mathematical domains and their powerful methods. Even if I seek to give the most general propositions and deal with series on *graded monoids*, it is the case of series on the free monoid, with coefficients in a field, which remains the richest and in which it is possible to present the theory of reduction of representations (due to Schützenberger).

The two last chapters are devoted to relations and functions realised by finite automata. Chapter IV deals with relations in general: first studied in a naive manner – that is, using the results of Chapters II and III as little as possible; then more generally – which poses the delicate problem of the definition and treatment of relations with weights. In place of the solution due to Jacob, one used by other authors dealing with this subject, and which restricts us to so-called 'regulated' relations, I propose to restrict weights to semirings which I call 'reasonable', which does not rule out any of the usual semirings, and which is enough to prove the two 'pivotal' theorems of the theory. The chapter continues with the problem of equivalence, which is undecidable even in the case where the output alphabet is unary (a theorem due to Ibarra and, independently, Lisovik), but is decidable if we take weights into account, a result due to Harju and Karhumäki. This last rests on the construction of the field of series called 'Malcev–Neumann' series, a result of which I endeavour to give a complete proof. Two families of relations are then described because of the variety of situations in which they appear: the *deterministic* relations and, above all, the *synchronous* relations.

Chapter V deals with functional relations realised by finite automata. The two hypotheses of functionality and rationality together give some remarkable structural results: in particular, the theorem of Elgot and Mezei, proved here with the Schützenberger covering construction. The chapter ends with the study of sequential transducers – recently restored to relevance by researchers in the area of the computational treatment of natural languages – and the characterisation, due to Choffrut, of sequential functions.

*

A book may be described as well by what it omits. I do not deal with *trees*, nor therefore with tree automata, even if a large part of the theory extends naturally to them and even though these objects are commonplace in computer science. Nor do I deal with *infinite words* to which my colleagues and friends Dominique Perrin and Jean-Eric Pin have just devoted an entire book. These choices were deliberate. I regret, on the other hand, not having treated the links between *logic* and finite automata (and hence, secondarily, alternating automata), a chapter which would naturally have

found its place in the first part; but one must stop writing at some point!

The absence of variety theory has already been mentioned; the study of pushdown automata, that of weighted automata in semirings of type $(\mathsf{max}, +)$, and the links between numeration systems and finite automata are also subjects with which I have not dealt, but sometimes mention or touch upon in an example or exercise, and which could be developed on the basis of the notions put in place in this work. If there is a sequel, they will doubtless occupy chapters there.

*

I would not want to end this presentation without mentioning this book's debt to the teachings and writings of the magisterial Prof. Marcel-Paul Schützenberger, a universal and paradoxical mind, who had no model and remains inimitable. I have drawn on many other sources, and benefited from the suggestions and advice of many colleagues, and acknowledge just as willingly the relationship of this book to Samuel Eilenberg's *Automata, Languages and Machines*,[1] but the influence of Schützenberger is of another order. It seems to me that, at least implicitly, Schützenberger developed in his articles, from *A remark on finite transducers* (1961) to *Une propriété de Hankel des relations rationnelles*[2](1977), a way of seeing automata – he might have said an 'Automatenanschauung', pretending to make fun of you – which he illustrated in his lectures and above all in his conversations: putting the automaton at the centre of the design, bringing other concepts to it, expressing the many propositions in this framework, and finally treating automata principally in terms of their matrix representation. Furthermore, Schützenberger explicitly set out[3] objectives for this algebraic theory of automata: to classify problems, bring out concepts, unify arguments, and apply these last to the fundamental results of mathematics. I think I have tried to take up these objectives; the reader will judge to what extent I have met them.

I must finally acknowledge that I have referred, mostly in the supplementary sections, to the best part of my own work. If I do not do it myself, who will? and if not now, when will I do it? The reader will admit however that I have not, by a long chalk, restricted myself to my own works. Furthermore, these works correspond to questions which have attracted me during the long preparation of this tome.

Key

There is no map without a key, a guide to convention which it is useful to assimilate before starting the journey.

On numbering. The basic unit is the *section*, subdivision of the chapter, itself divided into *subsections* which are sometimes broken down into *subsubsections*. Since

[1] Which besides also leads back to Schützenberger since Eilenberg ends his preface by thanking him thus: 'every phase of the development was endlessly discussed with him.'

[2] *Note added in translation: A Hankel property for rational relations.*

[3] In his lecture to the IFIP congress held in New York in 1965.

"subsubsection" is a horrible word, these are usually cross-referenced with a paragraph symbol, §.

The label of each numbered item is formed from the number of its section and its ordinal within its section. Propositions, lemmas and corollaries form a single class, and other items, theorems, definitions, properties, remarks, examples, exercises, figures and equations each have a class to themselves.[4] According to current usage, the items are referred to within their chapter by their label, and when they are referred to in another chapter, by their label prefixed by the number of their chapter, written in Roman numerals.

The division into sections of each chapter is replicated when we come to the solutions to exercises. The labels of those, rare, items that need to be identified there are prefixed with the initials SE.

On delimiters. The statements of theorems, propositions, lemmas and corollaries are in italics. Definitions, remarks and examples are not italicised, and thus the end of each is marked with a □.

The proofs of statements begin with '*Proof.*' and their end is marked with a ■ at the end of the line. When a proposition is not followed by its proof, whether because it is the conclusion of a chain of reasoning that has already been explained, or because this proof is so simple, or deferred to an exercise, the statement itself is terminated with a ■.

On notes. The marginal notes give precise references to propositions that are referred to in the text, while footnotes are reserved for comments on the terminology and notation adopted. At the end of each chapter is a brief section of historical notes and bibliographic references.

On exercises. I have put some exercises at the end of most subsections and have given the solution, or answer, to a good many of them, marked in this case with a • in the margin. There are three basic classes of exercises (which are not marked as such).

The first are exercises *for the sake of exercise.* I believe that one cannot understand a proposition unless one can perform, at least for a particular case, the computations which correspond to its proof in the general case, and that conversely such a feat is a powerful aid to understanding such comprehension. Furthermore, on many occasions I have taken advantage of these warm-ups to anticipate future computations or examples.

[4]This choice, like all choices, can be debated, and criticised inasmuch as a single ordering makes finding items easier; this criticism is alleviated by the fact that marginal notes give a page number for each cross-reference.

A second category of exercises consists of the proof of certain propositions left to the reader. I have indicated them systematically both to encourage the reader to do them and because it is necessary to reference those whose proof is given in the solutions.

The exercises of the third family allow us to explore areas that I did not want to cover, at least not with the same care as the body of the text. When solutions are given, they form an integral part of the chapter, which is why the solutions are not grouped at the end of the book. In more than one case, they are referred to in a later chapter.

Acknowledgements

Books, especially scientific books, are like expeditions: even if one rows alone, one does not make land, and certainly does not begin the adventure, without being sure of outside help.

The team... I was lucky enough to belong to the Laboratoire Informatique Théorique et Programmation, which for twenty-five years has been one of the most active centres of research in the world in the field of automata theory, and all the aspects of the theory that I have tried to cover were represented there. Not only were my colleagues at the laboratory a source of valuable conversation, but the fame of the LITP enabled me to meet visiting colleagues from around the world.

Since undertaking to write this book, I have not stopped asking them for advice, to clarify some point, to explain a difficulty, or to obtain a reference. For their welcome, patience, knowledge, answers, and also for their questions and encouragement, I warmly thank: F. Bassino, M.-P. Béal, J. Berstel, L. Boasson, J. Brzozowski, Ch. Choffrut, M. Crochemore, Z. Ésik, Ch. Frougny, P. Gastin, J. von zur Gathen, R. Gilman, S. Grigorieff, P.-C. Héam, J. van der Hoeven, J. Karhumäki, J. Mairesse, A. Mandel, S. Margolis, Y. Métivier, M. Nivat, K. Peeva, M. Pelletier, D. Perrin, J.-E. Pin, Ch. Prieur, A. Restivo, G. Sénizergues, I. Simon, J. Siquier, M. Tanaka, M. Waldschmidt and P. Weil.

Apologies to those whom I may have forgotten.

Several colleagues took the trouble to read most attentively an earlier (French) edition finished in March 2001. Their remarks and comments have been a valuable help and powerful encouragement. May A. Arnold, J. Berstel, J. Brzozowski, S. Crespi-Reghizzi, G. Duchamp, G. Jacob, J. Justin, J. Néraud, B. Petazzoni, Ch. Reutenauer, P. Simonnet, W. Thomas, R. Tijdeman and F. Yvon find here the expression of my thanks.

I would like finally to give a special mention to O. Carton and V. Bruyère who did me the kindness of entrusting me with then-unpublished manuscripts, and Sylvain Lombardy, who was my PhD student, and with whom I have worked in constant collaboration for the past four years.

This book was also nourished by experience acquired from the courses I have had the opportunity to give throughout this time, as part of the DEA, which was

successively called *Languages, Algorithms and Programming*; *Theoretical Computer Science, Computation and Programming*; and finally *Algorithmics*.[5] I have also had occasion to teach this material abroad, at the invitation of professors G. Hotz in Saarbrücken, A. Restivo in Palermo, I. Simon in São Paulo, J. Karhumäki in Turku and N. Sabadini in Como, whom I have here the pleasure of thanking, and finally at the ENST. My DEA students of the last two years have greatly helped to correct and improve the first edition; I am happy to thank particularly: L. Danthony, B. Durak, L. Fousse, S. Oudot, and Phan Duong Hieu.

According to the established formula, the errors and misprints which remain, and I know, statistically, that there are many, are all mine. I am grateful in advance to readers who will be kind enough to tell me of any they find.[6] I will try to keep an up-to-date list of errata, accessible from my web page.[7]

the equipment... Seated at the computer, the scientist has become author, typist and typographer all rolled into one. I am not among those who rejoice in this development, but since it is now the norm, with no discussion or revolt possible, I may as well thank those who have enabled me, thankfully, to have tools of such quality at my disposal.

At the head, D. Knuth for TEX and L. Lamport for LATEX. I used P. Keleher's remarkable text editor, Alpha, T. Kiffe's CMacTEX distribution, and T. van Zandt's diabolical PSTricks package for the figures.

But it is not true that all this works by itself! I am most indebted to those who, throughout this time, have guided me in choosing programs, supplied me with working versions, and have helped me out of difficulty. I thank M. Bovani, R. Card, A. Demaille, P. Gastin, J.-L. Mounier, S. Tardieu, and M. Zeitoun, for their help and patience, and spare a special thought for my old friend Éric Picheral's ready support.

A very early version of Chapters 0 and I appeared in a textbook[8] published by Ellipses whom I thank for allowing me to reuse the material. Finally, it is a pleasure to thank Jean-Philippe Moreux of Vuibert for his patience and competence when faced with my tortured author's delays, my apprentice typographer's whims, and the LATEX demons that I endlessly awoke.

...and the camp. B. Robinet and J.-P. Tubach warmly welcomed me to the ENST six years ago, as a member of of the Information Processing and Communication Laboratory; the successive directors of the Department of Computer Science and Networks, U. Finger and M. Riguidel, have seen to it that I lacked for nothing. I thank all of them sincerely.

---

[5] *Note added in translation:* The 'Diplôme d'Études Avancées' is the French equivalent of a Master's degree.

[6] For example by sending me an email at `sakarovitch@enst.fr`.

[7] `www.enst.fr/~jsaka`

[8] P. Bellot and J. Sakarovitch. *Logique et automates.* Ellipses, 1998.

I would like finally to thank the Centre National de la Recherche Scientifique (CNRS), a remarkable institution, often decried in France but which the whole world envies us.[9] Without the freedom given me by my position as a researcher, I do not think I could have done the work necessary to produce the material for this book. By way of thanks, I hope I shall have been able, if my peers should find merit in what I have done, and to finish at last in a less austere tone, to demonstrate that *recherche* is not *temps perdu*.

J. S.

Paris, September 2003

[9] *Note added in translation:* Thanks to the policies of the new French government, the world will soon have nothing to envy us (Paris, November 2008).

PROLOGUE

M. Pascal's division machine

Division? you say to yourself, Do you not rather mean the Pascaline, the adding machine which the young Blaise built to relieve his father from tiresome calculations and which once and for all set France in the firmament of computer-building nations?

— No indeed, I assure you, it is of *division* that I want to speak to you; but your surprise is not misplaced, and Pascal himself would be intrigued that we speak of a *machine*.

We can read[1] however in his complete works an original article[2] in which the mathematician–philosopher analyses the mechanism of division. Let us give him the floor:

> *Nihil tritius est apud arithmeticos quam...*

On second thoughts, let us turn instead to his translator:

> *Nothing in arithmetic is better known than the proposition according to which any multiple of* 9 *is composed of digits whose sum is itself a multiple of* 9. *[...] In this little treatise [...], I shall also set out a general method which allows one to discover, by simple inspection of its digits, whether a number is divisible by an arbitrary other number; this method applies not only to our decimal system of numeration (which system rests on a convention, an unhappy one besides, and not on a natural necessity, as the vulgar think), but it also applies without fail to every system of numeration having for base whatever number one wishes, as may be discovered in the following pages.*

Having stated the result:

UNIQUE PROPOSITION

> *From the sum of its digits alone may be deduced the remainder of division of a given arbitrary number by another fixed integer.*

Pascal proceeds in two steps. First, he observes that given an integer k and a base b, the sequence of remainders of division by k of successive powers of the base b is eventually periodic. This follows from the fact that the remainder (of division by k)

[1] I am grateful to Christiane Frougny for having brought this text to my attention.

[2] *De Numeribus Multiplicibus...* in B. Pascal, *Œuvres complètes*, see [180, pp. 84–89].

of the product of two numbers p and q is equal to the remainder of the product of the remainders of p and q respectively; this is written formally as:

$$p \equiv r \bmod k \ , \qquad q \equiv s \bmod k \quad \Longrightarrow \quad p\,q \equiv r\,s \bmod k \ ,$$

which for the case we are interested in gives

$$b \equiv r \bmod k \quad \Longrightarrow \quad b^2 \equiv b\,r \equiv r^2 \bmod k \ , \tag{$*$}$$

and is read: 'b^2 is congruent to r^2 modulo k' (to be congruent modulo k means to have the same remainder after division by k). There is only a finite number of possible remainders: $\{0, 1, \ldots, k-1\}$, and from the above we deduce that when we obtain a particular remainder *for the second time*, then we shall repeat indefinitely the sequence starting with this remainder. For example, if $k = 7$ and $b = 10$, the usual base, we have:

$$\begin{array}{lll} 10^0 = 1 \equiv 1 \mod 7 \ , & 10^1 \equiv 3 \mod 7 \ , & 10^2 \equiv 9 \equiv 2 \mod 7 \ , \\ 10^3 \equiv 6 \mod 7 \ , & 10^4 \equiv 4 \mod 7 \ , & 10^5 \equiv 5 \mod 7 \ , \\ 10^6 \equiv 1 \mod 7 \ , & 10^7 \equiv 3 \mod 7 \ , & \ldots \end{array}$$

and the sequence repeats: $1, 3, 2, 6, 4, 5, 1, 3, 2, 6, 4, 5, 1, 3, 2, \ldots$

Starting with this observation, Pascal constructs the following table: on the first row, in increasing order, *but from right to left*, the sequence of natural numbers, an abbreviation for the sequence of powers of the base; on the second row, the sequence of remainders that we have just calculated. This gives, for our example:

$$\begin{array}{ccccccccccccc} \ldots & n & \ldots & 9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 \\ \ldots & r_n & \ldots & 6 & 2 & 3 & 1 & 5 & 4 & 6 & 2 & 3 & 1 \end{array} \tag{$**$}$$

Pascal's second step is to propose an algorithm based on the table he has just constructed. Following the custom of the time he describes his algorithm using an example, and gets involved personally:

Suppose the remainder of division by 7 *of some number,* 548 *for example, is to be calculated.*

I take the first digit from the right, 8, *which I multiply by* 1 *(which is the* rightmost *figure on the second row of the table), that is to say:* 8.

To 8 *I add the second digit of the number under consideration,* 4, *multiplied by the second figure,* 3, *in the second row of the table:* $8 + 12 = 20$.

To 20, *I add* 5 *multiplied by* 2, *that is:* $20 + 10 = 30$.

The remainder of the division by 7 *of* 548 *is the same as that of the division of* 30, *that is:* 2.

Why? Because $548 = 5 \times 10^2 + 4 \times 10^1 + 8 \times 10^0$ and we use the fact that, like multiplication, the addition of natural numbers carries over to that of numbers modulo 7:

$$p \equiv r \bmod k \ , \qquad q \equiv s \bmod k \quad \Longrightarrow \quad p + q \equiv r + s \bmod k \ . \tag{$***$}$$

If we want to compute the remainder of large numbers, Pascal suggests we use his algorithm iteratively: for 389 265 978 412 for example, a first pass will enable us to compute that this number has the same remainder as 262 and a second pass that 262 (and hence 389 265 978 412) has the same remainder as 24, that is: 3 .

But it is instructive to rewrite ancient texts with the benefit of modern insights. A *simple but essential modification* of Pascal's method will allow us to implement this calculation of the remainder not only in a single pass, but also *using only numbers which are bounded independently of the size of the number whose remainder we are calculating.* Thus transformed, Pascal's algorithm becomes a 'machine' for computing the remainder.

Suppose the remainder of division by 7 *of* 548 *is to be calculated.*

I take the first digit from the right, 8, *which I multiply by* 1 *(which is the* rightmost *figure on the second row of the table), that is to say:* 8 *;* from 8, I subtract 7 as many times as possible; there remains: 1 .

To 1, *I add the second digit of the number under consideration,* 4, *multiplied by the second figure,* 3, *in the second row of the table:* $1 + 12 = 13$, from which I subtract 7 as many times as possible; there remains: 6 .

To 6, *I add* 5 *which I multiply by* 2, *or* $6 + 10 = 16$, from which I subtract 7 as many times as possible; there remains: 2 .

The remainder of the division by 7 *of* 548 *is:* 2 .

That this algorithm indeed computes the remainder follows obviously from repeated application of $(***)$. What has changed radically is the *information* which is transmitted from one step to the next of the algorithm. It consists of the remainder computed up to this step, a number between 0 and 6 for our example, and of the position in the table $(**)$. This table is infinite, so we might think we have not yet reached our goal. However, what counts is that, as we have already seen, the second row of this table is, after some point less than k, periodic with period less than k – in our example, of period 6, starting from the beginning.

*

We could explain how we turn this algorithm into a model of a machine with an internal memory of 42 (6×7) locations and which, by reading numbers written in base 10 from right to left, indicates at the end of reading the remainder of division by 7 of the number read. However, 42 positions is a lot, and the lengthy description would end up confusing the reader who has made the effort to follow us this far. Thus, at the risk of tiring the poor reader, we shall instead look at another, simpler example, which would have been too simple to illustrate the mechanism of division, but which will be perfect to describe the transformation from algorithm to machine.

We now consider numbers *written in base* 2 and seek to compute their *remainder after division by* 3. We have:

$$2^0 = 1 \equiv 1 \mod 3 \ , \quad 2^1 \equiv -1 \mod 3 \ , \quad 2^2 \equiv 1 \mod 3 \ ,$$
$$2^3 \equiv -1 \mod 3 \ , \quad 2^4 \equiv 1 \mod 3 \quad \dots \ ,$$

and the table (∗∗) becomes:

$$\begin{array}{cccccc} \dots & 4 & 3 & 2 & 1 & 0 \\ \dots & 1 & -1 & 1 & -1 & 1 \end{array} \qquad (**')$$

Pascal's algorithm (modified) becomes:

Suppose the remainder of division by 3 *of some number,* 1101 *for example (*13 *written in binary) is to be calculated.*

I take the first digit from the right, 1*, which I multiply by* 1 *(which is, in the table (∗∗′), the rightmost figure on the second row), that is:* 1 .

To 1*, I add the second digit of the number under consideration,* 0*, multiplied by the second digit,* -1*, of the second row of the table, that is:* $1 + 0 = 1$.

To 1*, I add* 1 *multiplied by* 1*, obtaining:* 2 .

To 2*, I add* 1 *multiplied by* -1*, which gives:* 1 .

The remainder of division by 3 *of* 1101 *is:* 1 .

Note that we rediscover in this procedure the rule, almost as well known as 'casting out nines', that a number is divisible by 11 (3 written in binary) if and only if the sum of its odd-ranked digits subtracted[3] from the sum of its even-ranked digits is divisible by 11.

At each step of the algorithm, we must 'remember' two pieces of information: the remainder of the division by 3 of the number which we have read up to this step, which we can decide to encode by one of the three numbers 0, 1 or 2; and the parity of the position of the digit which we are about to read, which we can decide to encode by $+1$ (even position) or -1 (odd position).

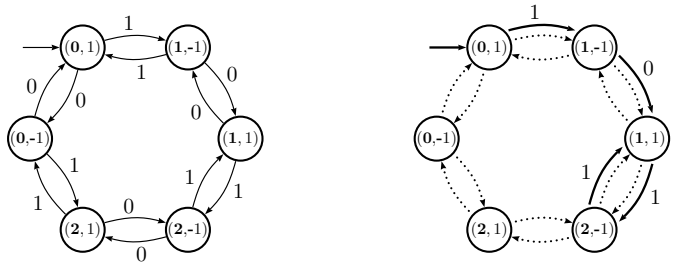
We will thus imagine a machine equipped with a memory that can take six different positions, encoded by the pairs $(0, +1)$, $(0, -1)$, $(1, +1)$, $(1, -1)$, $(2, +1)$ and $(2, -1)$, which we call the *state* of the machine. We imagine that this machine will read successively, from right to left, the digits of the number which it is processing. We do not worry *how* it reads these digits, but we shall specify that the *effect* of this reading translates the step of the algorithm corresponding to the digit read and the current state of the machine, *by changing the state* in a suitable manner. Thus, for our example, if the machine is in the state $(1, -1)$ and reads a 1, it will *change to the state* $(0, +1)$ since the algorithm consists of adding to the current remainder 1 the product of the digit read, 1, by the coefficient, -1, given by the position of the digit.

More generally, each digit read will change the second component of the state, $+1$ becoming -1 and *vice versa*; the first component will be left unchanged by reading

[3]It was to make this result appear that, subtly, we wrote that 2^1, 2^3, *etc.* are congruent to -1 (and not to 2) modulo 3, in the table (∗∗′).

a 0, while reading a 1 will increment it or decrement it by 1 (modulo 3), according to whether the machine is in a state whose second component is $+1$ or -1.

The complete device can be represented as in Figure P.1 (a). The six circles represent the states of the machine. Each arrow that goes from one state to another, marked with a digit, symbolises the change of state induced by reading that digit. Finally, the little arrow pointing at the state $(0, +1)$ indicates the state of the machine at the start of the computation.



(a) The machine (b) The computation of the remainder of 13

Figure P.1 : The divider by 3 which Pascal could have built

The computation of the remainder of division of a number n by k can thus be represented, as in Figure P.1 (b), by a succession of arrows, the first starting at the state $(0, +1)$, and each subsequent arrow starting where the last one ended; the succession of digits attached to these arrows gives the written form of n *from right to left*, and the state where the last arrow ends gives the *result* of the computation, which allows us to know the remainder.

*

A little reflection shows that the second part of Pascal's method, as modified by us, can in fact do without the first (the construction of the table $(**)$) provided that *we change the direction in which the number is read* and go from left to right, which is anyway the natural direction when performing division.

Suppose that we have already computed the remainder r of the division by k of a number n which is written, in base b, as a sequence of digits f. The number m which is written fc (f followed by c) where c is a digit, is equal to $n\,b+c$ and the remainder of the division of m by k is equal to the remainder of the division of $r\,b+c$, again by applying $(***)$. Let us use this idea on our example:

Suppose the remainder of division by 3 *of* 1101 *is to be calculated.*

I take the first digit from the left, 1*; the remainder of* 1 *divided by* 3 *is:* 1 .

I multiply this remainder 1 by 2, obtaining 2; *to* 2*, I add the second digit* from the left *of the number being considered,* 1*, which gives* 3*; the remainder of* 3 *divided by* 3 *is:* 0 .

I multiply this remainder 0 *by* 2*, giving* 0*, and add the third digit* 0*, which gives:* 0 .

I multiply the remainder 0 *by* 2, *giving* 0, *and add the fourth digit* 1, *giving:* 1 .
The remainder of the division by 11 *of* 1101 *is:* 1 .

We observe that at each step, the only datum we need to know is the remainder of division by 3 of the number which we have already read. Following the method above, we can turn this algorithm into a model of a machine with an internal memory of 3 positions. This machine, which reads numbers written in base 2 *from left to right* and computes the remainder of division by 3, will have 3 states, corresponding to the possible remainders 0, 1 and 2. From each state there will be two outgoing arrows, one corresponding to reading a 0, the other to reading a 1. The destination of these arrows is computed in the following table:

If the remainder of the number written f is:	0	1	2
then the remainder of the number written $f\,0$ is:	0	2	1
and the remainder of the number written $f\,1$ is:	1	0	2

The machine starts computing in state 0. All of this is shown in Figure P.2 .

(a) The machine

(b) The computation of the remainder of 13

Figure P.2 : The divider by 3 which Pascal could have built, if he had thought of the foregoing.

*

Here we have an example of a non-trivial computation which can be performed with a fixed quantity of memory, independent of the data being processed. Aside from its anecdotal quality (it is always exciting to find the germ of a modern idea in old texts) it enabled us to describe some of constituent elements of the structure 'finite automaton' which is the object of our study.

Despite its elementary aspect, this example raises in particular one of the questions that we shall try to resolve with the notions that we shall develop and the results we shall prove. By two different methods we have constructed two distinct machines which compute the same thing. If these two machines were given to us, from outside, as the result of some process over which we had no control, would we be able to see that they computed the same thing?

This question is fundamental. It comes down to knowing to what extent these machines, which are entirely defined by a *finite quantity of information*, faithfully represent, and in a useful manner, *infinite sets* (in the example, numbers divisible by 3). The answer is positive. This is one of the reasons that finite automata are a useful model and worthy of study.

Chapter 0

Fundamental structures

As its number indicates, this chapter was not written to be read. Here will be found reminders of more or less standard notions and structures which are used in this book, *with their notation*. The intention is that the reader should come here when the need arises.

This reader will however note that Sections 3 and 4 deal with *words*, a notion which it would be dishonest to brand 'standard' in the same sense as, for example, set union or the product of two matrices. In fact, and unlike the usual mathematical point of view which deals with numbers – a measure of *continuous magnitudes* – and with the related notions of functions and functionals, computer science, or at least that part which abstracts from the physical realisation of computers and concentrates on the problems of information processing, deals with *sequences*, usually *finite*, of *symbols*. These are *words*. This most general notion is not intended to conceal a vacuous concept; on the contrary, the variety of situations that it encompasses gives it its richness. It is well worth some definitions, and, for neophytes, the corresponding sections merit a detour.

Sections 7 and 8 are also rather non-standard outside the computer science community. Section 7 recalls some basic definitions of graph theory. It is not a preliminary to automata theory (all of whose definitions are given in Chapter I); rather, it allows us to refer, in some results on automata, to corresponding graph-theoretic results. Section 8 tackles more fundamental subjects, sketching the two notions of the complexity of a procedure and of the decidability of a problem. Even if the subject is not developed further here, these two notions are intrinsically linked to automata theory and will be referred to many times throughout this work.

This chapter contains, like all the rest, many exercises, which may seem to contradict the earlier injunction not to spend time on it. On the contrary: besides the exercises intended to familiarise the reader with new notions, most of the others allow me to state, and to prove in the solutions, quasi-ancillary properties which will be used later, sometimes much later, in the body of the text: it would be a bad idea to burden the reader's memory with them from the start. The marginal cross-references link to them immediately when needed.

Contents of the chapter

1. Relations
2. Monoids
3. Words and languages
4. Free monoids
5. Semirings
6. Matrices
7. Lexicon of graph theory
8. Complexity and decidability

Solutions to the exercises

Notes & references

1 Relations

In this first section, we will revise the basic mathematical definitions and notations that we shall need later, mainly connected with the notion of *relations*.

Sets

A *set* is a collection of objects described either by simple enumeration:

$$A = \{a, b, c, d, e\}$$

is the set of the first five letters of the Roman alphabet; or by the definition of a property satisfied by all the objects in the collection:

$$P = \{n \in \mathbb{N} \mid n = 2p \ , \quad p \in \mathbb{N}\}$$

is the set of even numbers. We will confine ourselves to this naive approach, borrowing the vocabulary and notation of set theory's elementary definitions: union, intersection, subset, membership and their notation are already familiar. As far as possible, I will use the following conventions: the elements (of a set) are denoted by lower case letters, sets by upper case letters, and functions and relations on sets by Greek letters. The number of elements, or *cardinal*, of a *finite* set E is written[1] $\|E\|$.

If E is infinite, its cardinal, a generalisation of the number of elements, is still written $\|E\|$. Most of the *infinite sets* which we will consider will be *denumerable* – that is to say that they can be enumerated, or, more precisely, put in bijection with $\mathbb{N}$, the set of non-negative integers. These are in a sense the smallest possible infinite sets. On the other hand, infinite sets of higher cardinality, which cannot be put in bijection with $\mathbb{N}$, are called *non-denumerable*.

The *cartesian product*, or more simply *product*, of two sets E and F is the set of pairs of elements, the first taken from E and the second from F:

$$E \times F = \{(e, f) \mid e \in E \, , \ f \in F\} \ .$$

The cardinal of the product of two *finite* sets is the product of the cardinals of the sets. The set of subsets of a set E is written $\mathfrak{P}(E)$; its cardinal, if E is finite, is $2^{\|E\|}$. If E is infinite, the cardinal of $\mathfrak{P}(E)$ is still written $2^{\|E\|}$ and is strictly greater than $\|E\|$: if E is infinite, $\mathfrak{P}(E)$ is non-denumerable.

The complement of a subset X of E is written $\complement X$ (and sometimes $\overline{X}$ for brevity). The *difference* of two subsets X and Y of E is written $X \setminus Y$ and we have $X \setminus Y = X \cap \complement Y$.

A *partition* of E is a set of pairwise disjoint subsets of E whose union is equal to E. A subset of $\mathfrak{P}(E)$, that is, a set of subsets of E, is a *Boolean algebra* if it is closed under union and complement (and thus under intersection and difference) and if it contains the empty set $\emptyset$ (and thus E).

[1] The notation $|E|$ would be more usual, but it the same as that used for the *length of words* (see Section 3). The risk of confusion is minimal until Chapter V, where the two notations would conflict.

The set of non-negative integers is written[2] $\mathbb{N}$; the set $\{1, 2, \ldots, n\}$ is written $[n]$: $[0]$ is the empty set.[3] In the rest of this chapter, $E, F, \ldots$ denote sets.

Functions and maps

A *function* α from E to F, written $\alpha\colon E \to F$, is a procedure which associates an element e of E with a value in F, called in general the *image of* e *(under* α*)*, and written

$$e\alpha \qquad \text{or alternatively} \quad (e)\alpha \ .$$

The justification of this postfix notation, a mostly North American usage, will become clear in the course of this book: the ease of writing composition (see below), and consistency with the natural notation for *actions* which will be one of our central objects of study (see Chapter II). It may seem cumbersome at first, but as we go along it will start to suggest ways of writing things which help our intuition. We will still use a prefix notation for certain classic cases:[4] $d(P)$ for the degree of a polynomial, $\mathsf{supp}(m)$ for the support of a matrix or series, $\mathfrak{P}(E)$, as we have just seen, for the set of subsets of a set, $\mathsf{min}(a, b)$ for the minimum of two integers, *etc.*

For some elements of E the procedure may fail, so that the function α does not associate any value with these elements; we say that α *is not defined on* these elements. We call the set of elements of E for which α is defined the *domain* of α, written $\mathsf{Dom}\,\alpha$, and the set of elements of F which are the image of one or more elements of E is called the *image* of α, or *codomain* of α, written $\mathsf{Im}\,\alpha$.

A function $\alpha\colon E \to F$ whose domain is all of E is called *total*; otherwise it is called *partial*. We will call a function whose definition implies that it is total – for example the morphisms of the next section – a *map*.

A function α from E to F is *injective* if any two distinct elements of E have distinct images, *surjective* if every element of F is the image of at least one element of E, that is, $(E)\alpha = F$, and *bijective* if it is both injective and surjective. This idea really makes sense only for total functions, and a bijective map is called a *bijection*.

Relations

A relation α from E to F is defined by a set of pairs (e, f), where e in E and f in F are the elements 'put into relation' by α, and again written $\alpha\colon E \to F$. This subset of $E \times F$ is called the *graph* of α and is written $\widehat{\alpha}$. If e is in E,

$$e\alpha = \{f \in F \mid (e, f) \in \widehat{\alpha}\} \subseteq F$$

is again called the *image of* e *in* α. The *domain* of α is the set of elements of E whose image is non-empty; the *image* of α (or *codomain* of α) is the set of elements of F

[2] *Note added in translation:* In other texts, $\mathbb{N}$ often excludes 0, and the non-negative integers are written $\mathbb{Z}_+$ or $\mathbb{N}_0$.

[3] In many texts, $[n]$ denotes $\{0, 1, 2, \ldots, n-1\}$; I have adopted the convention that seemed most convenient, betraying my tendency to start indices at 1 rather than at 0.

[4] And where the result is less often the argument of another function.

which are the image of some element of E; that is

$$\mathsf{Dom}\,\alpha = \{e \in E \mid e\alpha \neq \emptyset\} = \{e \in E \mid \exists f \in F \quad (e,f) \in \widehat{\alpha}\} \ ,$$

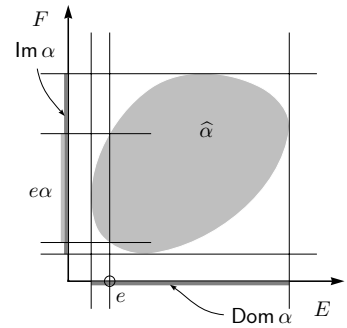
and $$\mathsf{Im}\,\alpha = \{f \in F \mid \exists e \in E \quad f \in e\alpha\} \qquad = \{f \in F \mid \exists e \in E \quad (e,f) \in \widehat{\alpha}\} \ .$$

The *inverse* of the relation α is the relation from F to E which has the same graph as α (modulo the natural identification between $F \times E$ and $E \times F$); it is written α^{-1}. It follows that

$$\forall f \in F \qquad f\alpha^{-1} = \{e \in E \mid f \in e\alpha\} = \{e \in E \mid (e,f) \in \widehat{\alpha}\} \ .$$

We obviously have

$$\mathsf{Dom}\,\alpha^{-1} = \mathsf{Im}\,\alpha \quad \text{and} \quad \mathsf{Im}\,\alpha^{-1} = \mathsf{Dom}\,\alpha$$



A function can be viewed as a relation such that $\|e\alpha\| = 1$ for all e in $\mathsf{Dom}\,\alpha$; such a relation is called *functional*. It is in this sense that we can speak of the *inverse of a function*.

A relation $\alpha\colon E \to F$ can also be viewed as a *map* from E to $\mathfrak{P}(F)$: for every e in E, the image of e in α is indeed an element of $\mathfrak{P}(F)$ and $e\alpha = \emptyset$ if and only if e is not a member of $\mathsf{Dom}\,\alpha$. Conversely, every map from E to $\mathfrak{P}(F)$ defines a relation from E to F. Following the same idea, every relation $\alpha\colon E \to F$ can be extended *by (complete) additivity* to a *map*, usually also written α, from $\mathfrak{P}(E)$ to $\mathfrak{P}(F)$:

$$\forall P \subseteq E \qquad P\alpha = \bigcup_{e \in P} e\alpha \ . \tag{1.1}$$

We can thus write $\mathsf{Im}\,\alpha = E\alpha$ and $\mathsf{Dom}\,\alpha = F\alpha^{-1}$. A map from $\mathfrak{P}(E)$ to $\mathfrak{P}(F)$ corresponds in this sense to a relation from E to F only if it is *additive*, that is if and only if it satisfies (1.1), and if it sends the empty set (element of $\mathfrak{P}(E)$) to the empty set (element of $\mathfrak{P}(F)$).

Additive extension allows us to express the *composition* of relations simply. If α is a relation from E to F, and β is a relation from F to G, then the *compositional product*, or simply *product*, of α by β (or of α *and* β, providing we remember that the order matters), written $\alpha \circ \beta$, or simply $\alpha\beta$, is the relation from E to G which associates an element e of E with the image in β of the image of e in α:

$$(e)\,[\alpha \circ \beta] = \big((e)\alpha\big)\beta = \{g \in G \mid \exists f \in F \quad f \in e\alpha \quad \text{and} \quad g \in f\beta\} \ .$$

Proposition 1.1 *Let $\alpha\colon E \to F$ and $\beta\colon F \to G$ be two relations. Then*

$$[\alpha \circ \beta]^{-1} = \beta^{-1} \circ \alpha^{-1} \ .$$

Proof. For every g in G, we have

$$\begin{aligned}(g)[\alpha \circ \beta]^{-1} &= \{e \in E \mid g \in e(\alpha \circ \beta)\} = \{e \in E \mid \exists f \in F \quad f \in e\alpha\,,\ g \in f\beta\} \\ &= \{e \in E \mid \exists f \in F \quad f \in g\beta^{-1}\,,\ e \in f\alpha^{-1}\} \\ &= \big((g)\beta^{-1}\big)\alpha^{-1} = (g)\,\big[\beta^{-1} \circ \alpha^{-1}\big] \ .\end{aligned}$$

■

Henceforth we will use the notation $\alpha\beta$ except when it would be cryptic.

Equivalences and orders

A relation α from E to itself is:

- *reflexive* if, for every e in E, $e \in e\alpha$, that is, if $\{(e,e) \mid e \in E\} \subseteq \widehat{\alpha}$;
- *symmetric* if, for every e and f in E, $f \in e\alpha \Rightarrow e \in f\alpha$, that is, if $\alpha = \alpha^{-1}$;
- *antisymmetric* if, on the contrary, for every e and f in E,
$$(f \in e\alpha \quad \text{and} \quad e \neq f) \Rightarrow e \notin f\alpha\,;$$
- *transitive* if, for every e, f and g in E, $(f \in e\alpha \quad \text{and} \quad g \in f\alpha) \Rightarrow g \in e\alpha$.

An *equivalence relation* is a relation that is reflexive, symmetric and transitive; a *partial order* is a relation that is reflexive, antisymmetric and transitive.

Let $\alpha\colon E \to E$ be an equivalence relation; if $f \in e\alpha$, then we also write
$$e \equiv f \bmod \alpha \qquad \text{or} \qquad e \equiv f \quad [\alpha]\ .$$
The subset $e\alpha$ is the *(equivalence) class* of e modulo α. The set of classes is called the *quotient set* of E by α, and is written E/α.

The relation whose graph is all of $E \times E$ is called the *universal relation* (on E), and is an equivalence relation. The relation of equality is called the *identity* (on E). This is the relation whose graph is the set of pairs (e,e) for every e in E, that is, the *diagonal* of $E \times E$.

cf. Ex.1.5 If $\alpha\colon E \to F$ is a function, then $\alpha\alpha^{-1}\colon E \to E$ is an equivalence relation, called the *map equivalence* of α.

When α is a partial order on E, we usually write $e\,\alpha\,f$ for $f \in e\alpha$ and such a relation is usually denoted by a symbol other than a Greek letter: $\leqslant$, $\preccurlyeq$, $\sqsubseteq$, *etc.* Notation of this sort is also used for equivalence relations. However, it is useful to bear in mind that equivalence relations and partial orders, like all relations, correspond to subsets of $E \times E$.

A partial order $\leqslant$ on E invites the use of conventional terminology: we say '*e is smaller than f*' if $e \leqslant f$. An order $\leqslant$ on E is *total* if $e \leqslant f$ or $f \leqslant e$ for every pair of elements e and f in E. A total order on E is a *well ordering* if every non-empty subset of E contains a smallest element.[5]

See also Sec. II.5, p. 293

Inclusion and intersection

We often implicitly identify a relation with its graph, and speak of the *inclusion* of two relations: $\alpha \subseteq \beta$ if $\widehat{\alpha} \subseteq \widehat{\beta}$. If $\alpha \subseteq \beta$ then, for every e in E, $e\alpha \subseteq e\beta$, and, for every f in F, $f\alpha^{-1} \subseteq f\beta^{-1}$.

Remark 1.1 The above identification is allowed inasmuch as the two notions of 'relation' and 'graph' correspond to a single concept;[6] it is often taken to the point at

[5] *Note added in translation:* We use 'order' and ' ordering' as synonyms for 'order relation'.

[6] Unless it is the concepts of relation and graph that correspond to the same notion.

which the same symbol is used for both. *For our purposes*, it would be a mistake to lose the ability to make a distinction. In fact, we will find ourselves (in Chapters IV and V) in a situation where we will define a *product* for graphs (subsets) distinct from the compositional product defined on relations. □

We define the *union* and *intersection* of two relations, or of *any family of relations* in the same way. All equivalence relations contain the identity and are contained in the universal relation. If α is a relation from E to F and A a subset of E, the *restriction* of α to A is the intersection of α with the relation whose graph is $A \times F$.

Transitive closure

The intersection of any family of transitive relations (from E to itself) is a transitive relation. The intersection of *all* the reflexive and transitive relations which contain a relation α – there is at least one because the universal relation is reflexive and transitive – is *the smallest* reflexive and transitive relation that contains α; it is written[7] $\overset{\star}{\alpha}$ and is called the *transitive closure* of α. It can also be defined extensionally:

Proposition 1.2 *Let α be a relation from E to itself and $\overset{\star}{\alpha}$ its transitive closure. Then, $e \overset{\star}{\alpha} f$ if and only if there exists a sequence $e = e_0\,,\ e_1\,,\ \ldots,\ e_n = f$ of elements of E such that, for all i in $[n]$, we have $e_{i-1}\,\alpha\, e_i$.*

Proof. The relation β defined by:

$$e\,\beta\,f \quad \Leftrightarrow \quad \exists e_0, e_1, \ldots, e_n \in E\,,\ e_0 = e\,,\ e_n = f\,,\ \forall i \in [n] \qquad e_{i-1}\,\alpha\, e_i$$

is transitive, and hence contains $\overset{\star}{\alpha}$. Furthermore, it is obviously contained in every transitive relation which contains α and hence in $\overset{\star}{\alpha}$: $\beta = \overset{\star}{\alpha}$. ■

The transitive closure of a symmetric and reflexive relation is an equivalence relation.

Exercises

1.1 Let α be a relation from E to F, A and B two subsets of E, and P and Q two subsets of F. Show:

(a) $(A \cup B)\alpha = A\alpha \bigcup B\alpha$; (b) $(A \cap B)\alpha \subseteq A\alpha \bigcap B\alpha$;

(c) $(P \cup Q)\alpha^{-1} = P\alpha^{-1} \bigcup Q\alpha^{-1}$;

1.2 Suppose moreover that α is functional. Show:

(a) $(P \cap Q)\alpha^{-1} = P\alpha^{-1} \bigcap Q\alpha^{-1}$; (b) $(\complement P)\alpha^{-1} = \mathsf{Dom}\,\alpha \setminus P\alpha^{-1}$.

(c) $P\alpha^{-1} \subseteq Q\alpha^{-1} \Leftrightarrow (P \cap \mathsf{Im}\,\alpha) \subseteq (Q \cap \mathsf{Im}\,\alpha)$.

Deduce from the last equivalence that if α is surjective, we have $P\alpha^{-1} \subseteq Q\alpha^{-1} \Leftrightarrow P \subseteq Q$.

[7] This unusual notation reveals my difficulty: the common usage is to write the transitive closure of α as α^*, but I want to keep the superscript $*$ for another purpose, while keeping the custom of writing transitive closure with a star...

• 1.3 With the same notation as above, suppose that α is functional and that $A\alpha = Q$. Show that $P \bigcap Q = (P\alpha^{-1} \cap A)\alpha$.

1.4 **Additive extension of relations.** (a) Verify that the notion of additive extension is consistent with that of composition product (that is, the composition of the extensions of two relations is equal to the extension of the composition of the relations).

(b) Observe that the same *does not hold for inversion* (that is, the inverse of the extension of a relation is not the extension of the inverse of that relation).

(c) Every relation $\alpha\colon E \to F$ can also be seen as a *map* from E to $\mathfrak{P}(F)$; the (additive) extension of this last is then a map from $\mathfrak{P}(E)$ to $\mathfrak{P}(\mathfrak{P}(F))$. What is the relationship between this extension and that of the relation α?

• 1.5 Let $\alpha\colon E \to F$ be a function. Characterise the relation $\alpha\alpha^{-1}$ in such a way that it becomes obvious that it is an equivalence relation on E.

• 1.6 Let $\alpha\colon E \to F$ be a relation. Verify that α is a (partial) function if and only if $\alpha^{-1}\alpha$ is the restriction of the identity on $\mathsf{Im}\,\alpha$.

1.7 Let $\alpha\colon E \to F$ and $\beta\colon F \to G$ be two functions. Verify that if $\alpha\beta$ is injective, α is injective, and that if $\alpha\beta$ is surjective, β is surjective.

1.8 We will say that a relation is *injective* if the images of two distinct elements are disjoint (this definition is obviously consistent with that of an injective function). Show that a relation is injective if and only if its inverse is a function.

• 1.9 **Generated Boolean algebras.** We say that an element of a Boolean algebra is an *atom* if it contains no other element of the algebra (which implies that it is either contained in or disjoint from every other element of the algebra).

Let E be a set and $\mathcal{F}$ a set of subsets of E: $\mathcal{F} \subseteq \mathfrak{P}(E)$. The *Boolean algebra generated by* $\mathcal{F}$ is the smallest set of subsets of E which contains $\mathcal{F}$ and which is a Boolean algebra.

(a) Characterise the atoms of the Boolean algebra generated by a set $\mathcal{F}$.

(b) If $\mathcal{F}$ is finite and contains n subsets of E, what is the largest possible number of elements (that is, of subsets of E) of the Boolean algebra generated by $\mathcal{F}$?

2 Monoids

A *monoid* is a set M, that is equipped with an *associative binary operation* and has a *neutral element* for this operation. The operation is usually called *multiplication* and written as a dot: $m \cdot n$, or, most often, by concatenation of the operands: mn. The neutral element is written 1_M, or simply 1 if there is no risk of ambiguity. Thus, by definition, if M is a monoid, we have:

$$\forall m, n, p \in M \qquad (mn)p = m(np), \qquad \text{and} \qquad m\,1_M = 1_M\,m = m\,.$$

The second equation implies that the neutral element is unique.

When the operation is *commutative*, that is, $mn = nm$ for all m and n, it is often called *addition* rather than multiplication and written as such; the neutral element is then written 0_M (or 0), and we say that the monoid M is *commutative*.

Submonoids

A *submonoid* S of a monoid M is a subset of M that is *stable* (or closed) under multiplication *and contains the neutral element.*

The intersection of an arbitrary number of submonoids of M is a submonoid. If P is an arbitrary subset of M, the *submonoid generated* by P, written $\langle P \rangle$, is the smallest submonoid which contains P; it is the intersection of all the submonoids which contain P.

A subset G of M such that $\langle G \rangle = M$ is a *set of generators* for M, or *generates* M; the monoid M is *finitely generated* if there exists such a finite G; we also say that M is *of finite type.*

Semigroups

We also use the *semigroup* structure: a semigroup is a monoid which does not necessarily have a neutral element; that is, a set equipped with an associative binary operation. The notions of *sub-semigroup*, of *generated sub-semigroup*, and of *set of generators* are defined as for monoids.

Every semigroup S can be transformed into a monoid, written S^1, by adding a neutral element unless S already has a neutral element.

Zero

An element z of a monoid, or of a semigroup, M (necessarily written multiplicatively) is a *zero* if, for every element m of M, we have:

$$z\,m = m\,z = z\ .$$

This equation implies that a zero, if it exists, is unique.

Inverses and groups

An element m' of a monoid M is an *inverse* of an element m of M if:

$$m'\,m = m\,m' = 1_M\ \ .$$

Associativity implies that this inverse is *unique.* A *group* is a monoid where every element has an inverse. It is a much richer (and more thoroughly studied) algebraic structure than monoids and semigroups.

Examples 2.1 (i) The set $\mathbb{N}$ of non-negative integers, equipped with the usual addition operation, is a (commutative) monoid. Likewise, $\mathbb{Z}$, the set of integers, is a monoid for addition; $\mathbb{N}$ is a submonoid of $\mathbb{Z}$. The singleton $\{1\}$ generates $\mathbb{N}$; the pair $\{+1, -1\}$ generates $\mathbb{Z}$. The set of (strictly) positive integers, written[8] $\mathbb{N}_*$, is a sub-semigroup of $\mathbb{N}$ (and of $\mathbb{Z}$).

[8]The notation $\mathbb{N}_+$ would not be consistent with that of $\mathbb{Q}_+$ or $\mathbb{R}_+$ which we will introduce later; $\mathbb{N}^*$, which one sometimes sees, is impossible because of the role assigned to the star exponent later.

(ii) The same sets of integers, equipped with ordinary multiplication, are also monoids, whose neutral element is 1; and 0 is the *zero* of $\mathbb{N}$ (and of $\mathbb{Z}$). The set P of prime numbers is a (minimal) set of generators for $\mathbb{N}_*$; $\{0\} \cup P$ generates $\mathbb{N}$, and $\{-1, 0\} \cup P$ generates $\mathbb{Z}$.
(iii) The set $\{0, 1\}$ is a (multiplicative) submonoid of $\mathbb{N}$; we write it U_2.
(iv) The set of relations (resp. of maps) from a set E to itself, equipped with the composition operation, is a monoid, written $\mathfrak{R}_E$ (resp.[9] $\mathfrak{T}_E$; sometimes $\mathfrak{T}(E)$). □

In the rest of this chapter, M and N denote two monoids.

Morphisms and congruences

A *map* $\alpha\colon M \to N$ from M to N is a *morphism (of monoids)* if it sends the neutral element of M to that of N and if it respects multiplication, that is, if:

$$(1_M)\alpha = 1_N \qquad \text{and} \qquad \forall p, q \in M \qquad (p \cdot q)\alpha = p\alpha \cdot q\alpha \ .$$

Note that in the equation above, the dot on the left hand side represents multiplication in M and the dot on the right hand side represents multiplication in N. A *bijective morphism* is also called an *isomorphism*.

An equivalence relation ε on M is a *congruence* if it is regular for multiplication in M; that is, if the product of two elements is equivalent to the product of two of their arbitrary equivalents:

$$\forall p, q, r, s \in M \qquad \left.\begin{array}{r} p \equiv q \bmod \varepsilon \\ r \equiv s \bmod \varepsilon \end{array}\right\} \quad \Longrightarrow \quad p\,r \equiv q\,s \bmod \varepsilon \ .$$

Morphisms and congruences are in fact one and the same thing, but described from two different points of view. Let $\alpha\colon M \to N$ be a morphism. The map equivalence of α, $\alpha\,\alpha^{-1}$, is a congruence; in fact, if

$$p \equiv q \bmod \alpha\,\alpha^{-1} \qquad \text{and} \qquad r \equiv s \bmod \alpha\,\alpha^{-1}, \qquad \text{we have} \quad p\,r \equiv q\,s \bmod \alpha\,\alpha^{-1} \ ,$$

since $(p\,r)\alpha = p\alpha \cdot r\alpha = q\alpha \cdot s\alpha = (q\,s)\alpha$. Conversely, if ε is a congruence on M we can equip the set M/ε of classes modulo ε with a multiplication, and the map which relates each element m of M to its class $[m]_\varepsilon$ is a (surjective) morphism.

Let G be a generating set of M; a morphism $\alpha\colon M \to N$ is entirely defined by the image of G under α. Conversely, it is not in general true that an arbitrary map from G to N can be extended into a morphism from M to N; this will be the characteristic property of the *basis* of a *free monoid*.

Generated congruence

The intersection of an arbitrary family of congruences (on M) is a congruence. Let R be a set of pairs (u, v) of elements of M; the intersection of *all* congruences which

[9] A gothic 'T'.

contain R – there exists at least one since the universal relation is a congruence – is the *finest* congruence which contains R; it is written $[R]$ and is called the *congruence generated by* R.

Again for a set R of pairs of elements of M, we write $\underset{R}{\vdash\!\dashv}$ for the symmetric and regular closure of R, so that for all m and n in M, we have

$$m \underset{R}{\vdash\!\dashv} n \quad \Longleftrightarrow \quad \exists (u,v) \in R\,,\ \exists h,k \in M$$
$$(m = h\,u\,k \quad \text{and} \quad n = h\,v\,k) \quad \text{or} \quad (m = h\,v\,k \quad \text{and} \quad n = h\,u\,k)\ .$$

We can then define $[R]$ extensionally:

Proposition 2.1 *The congruence $[R]$ generated by a set R of pairs (u,v) of elements of M is equal to $\overset{\star}{\underset{R}{\vdash\!\dashv}}$, the reflexive and transitive closure of $\underset{R}{\vdash\!\dashv}$.*

Proof. The reasoning is the same as for the proof of Proposition 1.2. The relation $\overset{\star}{\underset{R}{\vdash\!\dashv}}$ contains R and is a congruence: it contains $[R]$. Furthermore, it is clearly contained in every congruence which contains R and hence in $[R]$. ■

Exercises

- 2.1 What is the other monoid with two elements?
- 2.2 How many elements do $\mathfrak{R}_E$ and $\mathfrak{T}_E$ have, if $\|E\| = n$?
- 2.3 Verify that $\mathfrak{P}(M)$ equipped with the operation $P \cdot Q = \{p \cdot q \mid p \in P, q \in Q\}$ is a monoid. What is its neutral element? Does it have a zero?
- 2.4 Show that an equivalence relation ε on M is a congruence if and only if

$$\forall p,q,r \in M \qquad p \equiv q \bmod \varepsilon \quad \Longrightarrow \quad \{\, p\,r \equiv q\,r \bmod \varepsilon \quad \text{and} \quad r\,p \equiv r\,q \bmod \varepsilon \,\}\ .$$

2.5 Let $\alpha\colon M \to N$ be a morphism. Show that:

(a) if $P \subseteq M$ is a submonoid, $P\alpha$ is a submonoid of N;

(b) if $Q \subseteq N$ is a submonoid, $Q\alpha^{-1}$ is a submonoid of M;

(c) if $P \subseteq M$, $\langle P \rangle \alpha = \langle P\alpha \rangle$.

2.6 Let $\alpha\colon M \to N$ be a morphism.

(a) Show that the additive extension $\alpha\colon \mathfrak{P}(M) \to \mathfrak{P}(N)$ is a morphism of monoids.

(b) Let P and Q be in N; show that $P\alpha^{-1}Q\alpha^{-1} \subseteq (P\,Q)\alpha^{-1}$.

2.7 **Ideals.** A non-empty subset I of M is a *right ideal* if $I\,M = I$, a *left ideal* if $M\,I = I$, and a *two-sided ideal*, or simply an *ideal*, if it is both a right ideal and a left ideal; that is, if $M\,I\,M = I$.

(a) Verify that the family of right (resp. left, two-sided) ideals is closed under arbitrary union; what about intersection?

(b) Show that the inverse image, under a morphism of monoids, of a right (resp. left) ideal is a right (resp. left) ideal.

• 2.8 **Product.** (a) Verify that the direct product $M \times N$ is 'naturally' a monoid. What is its neutral element? Under what conditions does it have a zero?

(b) Show that for every monoid T and every pair of morphisms $\varphi\colon T \to M$ and $\psi\colon T \to N$ there exists a unique morphism, written $\varphi \times \psi$: $\varphi \times \psi\colon T \to M \times N$, which makes the adjacent diagram commute (where we write π_M and π_N for the projections from $M \times N$ to M and N respectively).

Characterise $\varphi \times \psi$.

$$\begin{tikzcd} & M \\ T \arrow[ur, "\varphi"] \arrow[r, "\varphi\times\psi"] \arrow[dr, "\psi"'] & M\times N \arrow[u, "\pi_M"'] \arrow[d, "\pi_N"] \\ & N \end{tikzcd}$$

(c) Show that conversely every monoid with this property with respect to M and N is isomorphic to $M \times N$.

• 2.9 In general, which submonoids of $M \times M$ are isomorphic to M?

2.10 **Group of units.** Show that the set of elements of a monoid M which have an inverse is a group (for the multiplication induced by that of M).

• 2.11 **Normal subgroups.** A subgroup H of a group G is called *normal* if for every g in G we have: $g\,H = H\,g$. Verify that the congruences of G are in bijection with its normal subgroups.

3 Words and languages

Here are the *words* promised earlier. The terminology, at least partly, starting with the word 'word', is borrowed from linguistics, a sign of the interest which linguists had, and still have, in automata theory.

Alphabet and words

An *alphabet* is a non-empty set, *a priori* finite, which is a set of symbols, usually denoted A, or at any rate, by an upper-case Roman letter. Logically, the elements of A – the symbols – are called *letters*, and finite sequences of letters are called *words*. A word f is thus written

$$f = (a_1, a_2, \ldots, a_n) \ ,$$

with the a_i in A. The set of words, sequences of letters of A, is written[10] A^* .

Examples 3.1 (i) The first example that comes to mind, prompted by the terminology, is that of the Roman alphabet, the usual set of letters. The words are sequences of letters, not necessarily found in the dictionary: 'ant', 'abracadabra' and 'xwoqz' are words.

(ii) We saw in the prologue that we can take a set of digits as an alphabet: for example, $\{0,1\}$, or $\{0,1,2,3,4,5,6,7,8,9\}$, or $\{0,1,2,3,4,5,6,7,8,9,\mathsf{A},\mathsf{B},\mathsf{C},\mathsf{D},\mathsf{E},\mathsf{F}\}$,

[10] An unhappy notation – we shall see why in the next chapter – but solidly established and naturally induced by the notation $a_1 a_2 \cdots a_n$ for the elements. In fact, it does not usually create annoying ambiguity. In Chapters II and IV we shall be led to distinguish carefully a sequence from the product of its elements and will write $\mathrm{V}(X)$ for the set of words over X (*cf.* also the solution to Exercise 4.10, p. 27).

and the words are representations of integers in the base corresponding to the chosen alphabet: 1101, 013 and D are *three representations* of the *number* 13 in bases 2, 10 and 16 respectively.

(iii) The Morse alphabet $\{\cdot, -, \sqcup\}$ is formed from the letters 'dot', 'dash' and 'space'. The words over this alphabet are messages: $\cdots\sqcup---\sqcup\cdots\sqcup$, for example.[11]

(iv) In some cases, the alphabet will be considerably larger. We can for example take as an alphabet the English *lexicon*: the 'words' of this special alphabet are sequences of words in the English language, which we usually call 'sentences', but they are not necessarily correct or meaningful; in this context,

beautiful him duke to die antidisestablishmentarianism

is a word over the *Oxford English Dictionary* alphabet. We can allow ourselves to be carried along by the evocative power of the terminology, but we must not be taken in by it.

(v) Finally, we take the alphabet made from sequences of characters which end in a 'carriage return', that is, the lines of a file; the words, which are sequences of lines, are exactly files. In this case, if we do not place a limit on the number of characters in a line, the alphabet is potentially infinite. □

Product

The set of words – that is, of sequences – is naturally equipped with a binary operation, that of *concatenation*:

$$(a_1, a_2, \ldots, a_n) \cdot (b_1, b_2, \ldots, b_p) = (a_1, a_2, \ldots, a_n, b_1, \ldots, b_p) \ .$$

This operation is clearly associative, as it has a neutral element: the *empty sequence* or *empty word*. Thus, A^* is a monoid; following the last section, concatenation is called *product* and the empty word is written 1_{A^*}, or even 1 when this does not create ambiguity.[12] The set of non-empty words is written

$$A^+ = A^* \setminus 1_{A^*} \ .$$

The qualifier 'proper' is often used as a synonym for *different from the empty word* or *does not contain the empty word*.

The product on words *is not commutative* when A contains at least two letters. The definition of product implies that the word f can be written

$$f = (a_1, a_2, \ldots, a_n) = (a_1) \cdot (a_2) \cdot \ \cdots \ \cdot (a_n) \ .$$

[11] 'SOS', of course.

[12] In most works in this area, especially in English, the most often used notation for the empty word is ε, (or λ, when the alphabet is written Σ). We prefer 1_{A^*}, which is more obviously different from a letter symbol, and which emphasises the nature of the neutral element for multiplication. This notation is obviously less well suited to the case where A is an alphabet of digits, as in Example 3.1 (ii) above; we shall then write the empty word ε. In fact, we shall often prefer to write numbers with alphabets of letters which will then be interpreted as digits (*cf.* Exam. I.1.2, p. 52 and III.2.5, p. 406).

Thus, every word is the product of the letters which compose it: A^* is *generated by* A. By identifying the sequence (a) with the letter a and by deleting, as usual, the explicit symbol for product, we thus write the word f in the form

$$f = a_1 a_2 \cdots a_n \quad ,$$

and the product of two words f and g in the form $f\,g$. We return to the product of words in the next section, and confine ourselves for the rest of this section to purely descriptive notions of words viewed as sequences of letters.

Length

The *length* of a word is naturally the length of the sequence; that is, the number of letters the word contains. The length of $f = a_1 a_2 \cdots a_n$ is n; it is written $|f|$. We write $|f|_a$ for the number of occurrences of the letter a in f. If f is a word in A^*, we then have

$$|f| = \sum_{a \in A} |f|_a \quad .$$

Thus, for $f = a\,b\,a\,a\,b\,a\,b\,a$, $|f| = 8$, $|f|_a = 5$ and $|f|_b = 3$.

The empty word 1_{A^*} is of length 0 and is the only such word in A^*. All the letters of the alphabet have length 1 and, for all words f and g in A^*, and every letter a in A, we have

cf. Exam. 4.1, p. 25

$$|f\,g| = |f| + |g| \qquad \text{and} \qquad |f\,g|_a = |f|_a + |g|_a \quad . \tag{3.1}$$

Factors

Let f, g, h and u be words in A^*; g is a *left factor* or *prefix* of f if there exists h such that $f = g\,h$; g is a *proper left factor* or *proper prefix* if h is other than the empty word; h is a *right factor* or *suffix* of f if there exists g such that $f = g\,h$; h is a *proper right factor* or *proper suffix* if g is other than the empty word; u is a *factor* of f if there exist g and h such that $f = g\,u\,h$; u is a *proper factor* if g and h are not both equal to the empty word.[13]

In other words, g is a prefix of f if f 'starts' with g, h is a suffix if f 'ends' in h, and u is a factor if f 'contains' u (that is, contains the letters of u consecutively and in the right order). Every factor of f is the suffix of a prefix of f, and also a prefix of a suffix of f.

Example 3.2 Again let $f = a\,b\,a\,a\,b\,a\,b\,a$; $a\,b\,a$ is a prefix, suffix and factor of f, $a\,b\,a\,a$ is a prefix but not a suffix, $b\,a\,b\,a$ is a suffix but not a prefix, $a\,a\,b\,a$ is a factor which is neither a prefix nor a suffix, and $a\,a\,a$ is not a factor of f. □

[13] *Note added in translation:* Once again, we depart from the most common English usage, where factors are called 'subwords' and what we call subwords are called 'sparse subwords', 'scattered subwords' or 'subsequences'. Our choice emphasises the multiplicative structure: a word is a product of (some of its) factors.

Note that the empty word is a prefix, suffix and factor of every word. A word of length n has exactly $n+1$ *distinct* prefixes (n proper prefixes) and $n+1$ *distinct* suffixes (n proper suffixes), but we do not know *a priori* the number of distinct factors of a word of given length. We write $\mathsf{Pre}(f)$, $\mathsf{Suf}(f)$ and $\mathsf{Fac}(f)$ for the sets of prefixes, suffixes and factors of f respectively. A most natural way to express conditions on words is to put them in terms of a prefix, suffix or factor; for example, that a word starts, or does not start, in such and such a way. *cf. Exer. 3.3, p. 23*

Example 3.3 Let $B = \{0, 1\}$ be the alphabet of digits 0 and 1. Each word of B^* is the binary representation of a positive integer and each strictly positive integer[14] has a binary representation, unique if we specify that it *does not start with* 0. The integers divisible by 8 are characterised by the fact that their binary representations *end in* 000. □

The relation 'is a prefix of' is an order relation on A^*: we write $g \leqslant f$ if g is a prefix of f (N.B. $g \geqslant f$ means then that f is a prefix of g, not that it is a suffix).

A *factorisation* of a word f is a sequence $(g_1, g_2, \ldots, g_n)$ of words such that $f = g_1 g_2 \cdots g_n$ (we shall allow ourselves to write: 'let $f = g_1 g_2 \cdots g_n$ be a factorisation of f'); this factorisation is *proper* if no g_i is equal to the empty word (and if n is greater than 1).

Subwords

Since a word f is a *sequence* (of letters), we call a subsequence of the sequence f a *subword*. Every factor of f is a subword of f, but the converse is false, and the two notions should not be confused:[15] $aaaa$, bb and $babaa$ are subwords of $f = abaababa$ without being factors. In other words, g is a subword of f if f contains the letters of g in the right order but not necessarily consecutively.

Example 3.4 The utility program `diff` compares two files and computes to that end the longest subword common to these files considered as words written in the alphabet of lines (which is why we chose to give Example 3.1 (v)). □

Mirror image

If $f = a_1 a_2 \cdots a_n$ is a word in A^*, the *mirror image* of f (or *transpose* of f) is the word f^{t}:

$$f^{\mathsf{t}} = a_n a_{n-1} \cdots a_1 \quad .$$

Thus, $(abaababa)^{\mathsf{t}} = ababaaba$. The prefixes of f^{t} are the transposes of the suffixes of f and *vice versa*. The factors of f^{t} are the transposes of the factors of f.

Examples 3.5 (i) We remarked that to test the divisibility of a number, a word written in the alphabet of digits of some base, by a given integer, we can proceed to

[14] The *integer* 'zero' is usually represented by the *word* 0 and does not satisfy the condition.

[15] See Note 13 above.

read the word *from right to left* (the method suggested by Pascal) or *from left* to right. We see that reading a word f from right to left is the same as reading the transpose of f from left to right.

(ii) A word which is equal to its mirror image is called a *palindrome*. In English, 'ewe', and 'radar' are well-known palindromes. We denote the set of palindromes of A^* by $\mathsf{Pal}_{\mathsf{A}}$, or if there is no ambiguity about the alphabet, simply Pal:

$$\mathsf{Pal} = \{f \in A^* \mid f = f^{\mathsf{t}}\} \ . \qquad \square$$

Languages

A *language over* A, or *language of* A^*, is any set of words written in the alphabet A.

In other words, a language of A^* is a subset of A^*; that is, an element of $\mathfrak{P}(A^*)$, the set of all the subsets of A^*. We can thus naturally define for languages all the usual operations on the subsets of a set: union, intersection, complement and difference, with the usual notation.

Example 3.6 Let $A = \{a, b\}$ and $Z_1 = \{f \in A^* \mid |f|_a = |f|_b\}$, the set of words in A^* which contain the same number of a's and b's, $L_2 = \{f \in A^* \mid |f| \equiv 1 \bmod 2\}$, the set of words of odd length and $K_1 = \{a^n b^m \mid n, m \in \mathbb{N}\}$, the set[16] of words comprised of a sequence of a's *followed* by a sequence of b's. We have (the reader will be easily convinced)

$$Z_1 \cap L_2 = \emptyset \ , \quad Z_1 \cap K_1 = \{a^n b^n \mid n \in \mathbb{N}\} \ ,$$
$$A^* \setminus K_1 = \{f \in A^* \mid \ f \text{ contains a factor } ba \ \} \ . \qquad \square$$

All the functions which we define on words extend, *additively*, to languages. Thus, for prefixes, suffixes and factors:

$$\mathsf{Pre}(L) = \bigcup_{f \in L} \mathsf{Pre}(f) \ , \quad \mathsf{Suf}(L) = \bigcup_{f \in L} \mathsf{Suf}(f) \ , \quad \text{and} \qquad \mathsf{Fac}(L) = \bigcup_{f \in L} \mathsf{Fac}(f) \ .$$

We also have $\mathsf{Fac}(L) = \mathsf{Pre}(\mathsf{Suf}(L)) = \mathsf{Suf}(\mathsf{Pre}(L))$. Likewise for the mirror image:

$$L^{\mathsf{t}} = \{f^{\mathsf{t}} \mid f \in L\} \ .$$

We then have $\mathsf{Pal} = \mathsf{Pal}^{\mathsf{t}}$, $Z_1 = Z_1^{\mathsf{t}}$, and $L_2 = L_2^{\mathsf{t}}$.

A language is called *prefix* (resp. *suffix*) if it does not contain two words one of which is a proper prefix (resp. suffix) of the other;[17] that is,

$$L \subseteq A^* \quad \text{prefix} \qquad \Longleftrightarrow \qquad \forall f, g \in L \qquad g \leqslant f \quad \Rightarrow \quad g = f \ .$$

Prefix (and suffix) languages are closed under intersection. The transpose of a prefix language is suffix and *vice versa*.

[16] The examples are usually written with symbols that correspond to their type (for languages and sets of words, upper-case Roman letters) indexed by integers. As we are trying to keep the same notation for each example throughout the book, we are led to choices which may seem bizarre; some will be clarified later (such as Z_1), others are made to avoid usurping a later use of the same notation (such as L_2 rather than the expected L_1).

[17] This terminology is somewhat bizarre, but well established; some authors, however, say more explicity 'prefix-free'.

A language is called *prefix-closed* (resp. *suffix-closed*, *factor-closed*) if it contains the prefixes (resp. the suffixes, the factors) of all its elements; that is, if $\mathsf{Pre}(L) = L$ (resp. $\mathsf{Suf}(L) = L$, $\mathsf{Fac}(L) = L$). Prefix-closed (and suffix-closed and factor-closed) languages are closed under intersection. The transpose of a prefix-closed language is suffix-closed and *vice versa*; the transpose of a factor-closed language is factor-closed.

A combinatorial property

The combinatorics of words is a rich and subtle field; one could devote entire works[18] to it without exhausting it. It is a powerful tool for constructing examples (and counter-examples). If $A = \{a, b\}$, the only words which do not contain a square are: a, b, ab, ba, aba and bab. Hence the significance of the following result, which we shall not prove.[19]

Theorem 3.1 [**Thue**] *Over an alphabet of three letters, there exists an infinity of words that contain no factor which is a square.* ■

Exercises

• 3.1 How many letters has the alphabet in which English words are written?

• 3.2 If the alphabet A has k letters, how many words of length n are there in A^*? How many of length n which contain l occurrences of a given letter in A?

• 3.3 How many distinct factors of length 3, 4 and 5 in the following three words of length 20?

(a) $f = abaababaabaababaabab$; (b) $g = aaababbbbaabbabaaaab$;

(c) $h = abaaabaaabaaabaaabaa$.

3.4 By making a and b play asymmetric roles, verify that every word f in $\{a, b\}^*$ can be written uniquely in the form

$$f = a^{k_0} b a^{k_1} b \cdots a^{k_{n-1}} b a^{k_n} \quad ,$$

where the k_i are non-negative integers.

3.5 Let A be a finite alphabet. Show that A^* is a denumerable set. What if A is a denumerable set itself?

3.6 **Fine & Wilf's Theorem.** If we view the words in A^* as *functions* from an interval $[1, n]$ of $\mathbb{N}$ to A (which is not a bad idea), then

$$f(i) = a_i \qquad \text{when} \qquad f = a_1 a_2 \cdots a_n \quad ,$$

and we naturally say the *period* of f is an integer p such that for all i, $1 \leqslant i \leqslant n - p$, we have $f(i + p) = f(i)$.

Let f be a word which has two periods p and q and let $d = \gcd(p, q)$. Show that if $|f| \geqslant p + q - d$ then d is a period of f. Show that the condition given on the length f is optimal.

---

[18] See for example the two works published by colleagues and friends under the collective *nom de plume* M. Lothaire: *Combinatorics on Words* [158] and *Algebraic Combinatorics on Words* [159]. *Added in translation:* And even a third one: *Applied Combinatorics on Words* [160].

[19] A. Thue's original proof is more accessible in [158, Ch. 2].

- 3.7 What is $\mathsf{Pre}(Z_1)$?
- 3.8 Describe $\mathsf{Pal} \cap Z_1 \cap a^*b^*a^*b^*a^*$.
- 3.9 Let $B_1 = \{a\,b, a\,b\,a^2b, a\,b\,a^2b\,a^3b, \ldots, a\,b\,a^2b \cdots b\,a^{n-1}b\,a^nb, \ldots\}$. Describe $A^*b \setminus B_1$.
- 3.10 **Lexicographic order.** Let $A = \{a_1, a_2, \ldots, a_n\}$ be a *totally ordered* alphabet: $a_1 < a_2 < \cdots < a_n$. We define the relation $\preccurlyeq$ on A^* by:

$$f \preccurlyeq g \Longleftrightarrow \begin{cases} g = f\,h & \text{with} \quad h \in A^* \quad \text{or} \\ f = u\,a_i v\,, \quad g = u\,a_j w & \text{with} \quad i < j \ . \end{cases}$$

(a) Verify that $\preccurlyeq$ is a total order, which we call the *lexicographic order*, but is not a well ordering.

(b) Explain why the words in an English dictionary are *not* in lexicographic order. Describe the order.

4 Free monoids

The set A^* of words over A is equipped with product, concatenation, and a neutral element for product, the empty word 1_{A^*}: it is a monoid, called the *free monoid generated by A*. The definition of A^* as a set of sequences implies the fundamental property, from which everything follows, that every word has a *unique written form* as the product of letters of A:

$$a_1a_2 \cdots a_n = b_1b_2 \cdots b_m \quad \Longrightarrow \quad \begin{cases} n = m \\ \text{and} \quad a_i = b_i \quad \text{for all } i,\ 1 \leqslant i \leqslant n. \end{cases} \tag{4.1}$$

Universal property

The free monoid A^* is also *characterised* by the following property:[20]

Property 4.1 *Every map γ from A to an arbitrary monoid M* extends to a unique morphism $\breve{\gamma}$ *from A^* to M:*

$$\forall f = a_1a_2 \cdots a_n \in A^* \qquad (f)\breve{\gamma} = (a_1)\gamma \cdot (a_2)\gamma \cdot \ \cdots \ \cdot (a_n)\gamma$$

where the dot on the right hand side represents the product in M.

Proof. The unique written representation of f as a sequence of letters in A ensures both the definition of $\breve{\gamma}$, the fact that it is a morphism, and its uniqueness.

If F is a monoid which satisfies this universal property (for A), we obtain, by taking A^* as the monoid M, that F and A^* are isomorphic. ∎

It follows, among other things, that every morphism $\alpha\colon A^* \to M$ is entirely determined by the image of A in M under α. For simplicity, we shall from now on use

[20] It is in fact this property, called *universal* from a category-theoretic point of view, which gives A^* its status as a *free monoid*.

the same symbol, γ for example, for a map from A to M and the morphism from A^* to M that this map defines, instead of $\breve{\gamma}$ as above. Along the same lines, we have the following property, which is also characteristic of free monoids. *cf. Exer. 4.10*

Lemma 4.1 *Let M and S be two monoids. Let $\alpha\colon A^* \to M$ be a morphism, and $\sigma\colon S \to M$ a* surjective *morphism. Then, there exists a morphism $\theta\colon A^* \to S$ such that $\alpha = \theta \circ \sigma$.*

Proof. For each a in A, let s_a be an arbitrary but fixed element of $a\alpha\sigma^{-1}$, which is non-empty since σ is surjective. The morphism θ defined by $a\theta = s_a$ fits the bill. ■

$$\begin{tikzcd} & A^* \arrow[dl, dashed, "\theta"'] \arrow[d, "\alpha"] \\ S \arrow[r, "\sigma"'] & M \end{tikzcd}$$

Morphisms

A morphism $\alpha\colon A^* \to B^*$ is called *continuous*[21] if no letter in A is sent by α to the empty word; that is, $A\alpha \subseteq B^+$, or $|f| \leqslant |f\alpha|$ for every f in A^*. The morphism α is *alphabetic* if the image of every letter is either a letter, or the empty word; that is, $A\alpha \subseteq B \cup \{1_{B^*}\}$, or $|f\alpha| \leqslant |f|$ for every f in A^*, and *strictly alphabetic* if it is a both alphabetic and continuous. If A is the disjoint union of B and C, then the *projection* of A^* on to B^* is the morphism which is the identity on B and which sends each letter of C to the empty word.

Example 4.1 The 'length' map from A^* to $\mathbb{N}$ is a morphism since

$$|f\,g| = |f| + |g| \qquad \text{and} \qquad |1_{A^*}| = 0 \ .$$

Likewise, each of the maps $f \mapsto |f|_a$, for a in A, is a morphism from A^* to $\mathbb{N}$.

If the alphabet A has a single letter ($A = \{a\}$), then $A^* = a^*$ and the length morphism is *injective* (it is also *surjective* as in all cases). The classic formula

$$a^n \cdot a^m = a^{n+m}$$

is the expression of *isomorphism* between a^* and $\mathbb{N}$. □

As in the case of arbitrary monoids, the product of words extends to a *product of languages*. This operation is too important to us to be hidden in these preliminaries, and we shall come back to it later. *cf. Exer. 2.3, p. 17* *cf. Sec. I.2.1, p. 82*

A combinatorial lemma

The obvious relations between the various prefixes, suffixes and factors of a word are formalised in the following proposition. This simple lemma is the basis of other less trivial results (*cf.* Exer. 4.6, 4.7 and 4.8).

[21] The terminology is justified in Chapter IV.

Lemma 4.2 *Let f, g, k and l be four words of A^* such that $f\,g = k\,l$. We have:*

(i) *if $|f| = |k|$, then $f = k$ and $g = l$;*

(ii) *if $|f| > |k|$, there exists a unique word h such that $f = k\,h$ and $l = h\,g$;*

(iii) *if $|f| < |k|$, there exists a unique word h such that $k = f\,h$ and $g = h\,l$.*

Proof. (i) is obvious. (ii) If $|f| > |k|$, k is a prefix of f and $f = k\,h$, from which we deduce $k\,h\,g = k\,l$ and, by (i), $h\,g = l$. (iii) If $|f| < |k|$, f is a prefix of k and $k = f\,h$; hence we deduce $f\,g = f\,h\,l$ and, by (i), $g = h\,l$. We could also exchange f and k, and g and l, to reduce the problem to (ii). ■

Equidivisibility

More generally, we say that a monoid M is *equidivisible* if it satisfies Lemma 4.2 (without discriminating by length); that is if, for all m, n, p, and q in M such that

$$m\,n = p\,q \quad ,$$

there exists an r in M such that

$$\text{either} \quad p = m\,r \quad \text{and} \quad r\,q = n \quad , \qquad \text{or} \quad m = p\,r \quad \text{and} \quad r\,n = q \quad .$$

We have just seen that a free monoid is obviously equidivisible, but the converse is not true: an equidivisible monoid is not necessarily free.

cf. Def. III.1.3, p. 383

Furthermore, we shall say that a M is *graded* if there exists a function $\varphi\colon M \to \mathbb{N}$ (called a *gradation*) such that $\varphi(m) = 0$ only if $m = 1_M$ and if for all m and n in M, we have

$$\varphi(m\,n) = \varphi(m) + \varphi(n) \quad .$$

A free monoid is graded (the length of words is a gradation) but the converse is not true. The two notions of equidivisibility and gradation are linked by the following proposition.

Proposition 4.3 [**Levi**] *A graded and equidivisible monoid is free.* ■

Exercises & Problems

• 4.1 (a) Let $\gamma\colon \{a,b\}^* \to \mathbb{Z}$ be the morphism defined by $a\gamma = +1$ and $b\gamma = -1$ (where $\mathbb{Z}$ is a monoid for addition). Show that $(0)\gamma^{-1} = Z_1$.

cf. Exam. 3.6, p. 22

(b) Let $\alpha\colon \{a,b\}^* \to \mathbb{Z}$ be the morphism defined by $a\alpha = +2$ and $b\alpha = -3$. Characterise $\mathsf{Im}\,\alpha$ and $(0)\alpha^{-1}$.

4.2 (a) Show that every congruence of a^* is generated by a single pair (a^p, a^{p+r}).

(b) Describe all the quotients of $a^* = \mathbb{N}$.

• 4.3 Show that the mirror image map is an *anti-isomorphism* from A^* to itself; that is:

Property 4.2 $\forall f, g \in A^* \qquad (f\,g)^{\mathsf{t}} = g^{\mathsf{t}}\,f^{\mathsf{t}}$.

• 4.4 Consider morphisms from one free monoid to another. Show that:

(a) every morphism is the composition of an injective morphism and a projection;

(b) every continuous morphism is the composition of an injective morphism and a continuous alphabetic morphism.

4.5 Let π_A and π_B be the projections of $(A \cup B)^*$ on to A^* and B^* respectively. Verify that $\pi_A \times \pi_B \colon (A \cup B)^* \to A^* \times B^*$ is surjective if and only if A and B are disjoint.

• 4.6 Let f and g be words in A^* such that $f\,g = g\,f$. Show that there exists a word h and integers m and n, such that $f = h^m$ and $g = h^n$.

4.7 **Conjugate words.** Let f and g be words in A^* for which there exists a word h such that $f\,h = h\,g$. Show that there exist words u and v, and integers m and n such that $f = (u\,v)^n$, $g = (v\,u)^n$ and $h = u\,(v\,u)^m$.

• 4.8 Let f_1, f_2, f_3, f_4 and g_1, g_2, g_3, g_4 be two quadruples of words in A^*. Show that the three equalities $f_1 f_4 = g_1 g_4$, $f_1 f_2 f_4 = g_1 g_2 g_4$ and $f_1 f_3 f_4 = g_1 g_3 g_4$ imply the fourth one: $f_1 f_2 f_3 f_4 = g_1 g_2 g_3 g_4$.

4.9 (a) Verify that a group is equidivisible.

(b) Verify that an equidivisible monoid has neither divisors of the identity, nor any other idempotent than the identity.

(c) Give an example of an infinite monoid which is not equidivisible. Why did we specify 'infinite'?

• 4.10 **A characterisation of free monoids.** (a) Let W be a submonoid of A^*, a free monoid. Show that $B = (W \setminus 1_{A^*}) \setminus (W \setminus 1_{A^*})^2$ is the minimum generating set of W.

(b) Show that Lemma 4.1 is characteristic of free monoids.

• 4.11 Prove Proposition 4.3.

5 Semirings

Rings and fields, modules and vector spaces: these are fundamental structures on which a large part of mathematics is built. When we really want to formalise the objects that we are studying, to organise and prove their properties, we can only use 'less rich' structures, which are to the aforementioned as monoids are to groups. This section just gives a few definitions and examples. It is complemented by a short primer on non-commutative linear algebra in Chapter III. *Sec. III.7, p. 488*

Semirings

A semiring $\mathbb{K}$ is a set equipped with two binary operations, *addition* and *multiplication*, which satisfy the following axioms:

• SA1 $\mathbb{K}$ is a *commutative* monoid for addition, written $+$, whose neutral element, called the *zero* of $\mathbb{K}$, is written $0_{\mathbb{K}}$ (or 0);

• SA2 $\mathbb{K}$ is a monoid (not necessarily commutative) for multiplication, written by a dot, or more often by simple juxtaposition, whose neutral element, called the *identity* of $\mathbb{K}$, is written $1_{\mathbb{K}}$ (or 1);

• SA3 multiplication distributes left and right over addition; that is,

$$\forall i, j, k \in \mathbb{K} \qquad i \cdot (j + k) = (i \cdot j) + (i \cdot k) \quad \text{and} \quad (i + j) \cdot k = (i \cdot k) + (j \cdot k) \ ;$$

• SA4 the neutral element for addition is a zero for multiplication (which justifies the terminology):

$$\forall k \in \mathbb{K} \qquad k \cdot 0_{\mathbb{K}} = 0_{\mathbb{K}} \cdot k = 0_{\mathbb{K}} \ .$$

If $1_{\mathbb{K}} = 0_{\mathbb{K}}$, then $\mathbb{K}$ is reduced to this single element. In what follows $\mathbb{K}$ denotes a semiring and we assume $1_{\mathbb{K}} \neq 0_{\mathbb{K}}$.

A semiring is *commutative* when its multiplication is a commutative operation.

Examples 5.1 (i) The 'Boolean' semiring $\mathbb{B} = \{0, 1\}$ is completely defined by the axioms and the equation $1 + 1 = 1$.
(ii) The set $\mathbb{N}$ of non-negative integers, equipped with the usual operations of addition and multiplication, is a semiring. Likewise, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{Q}_+$, $\mathbb{R}$ and $\mathbb{R}_+$ denote the semirings of integers, of rationals, of non-negative rationals, of reals, and of non-negative reals, equipped with the usual operations.
(iii) The set $\mathcal{N} = \mathbb{N} \cup \{+\infty\}$ is a semiring when we extend addition and multiplication by the rules

$$n + +\infty = +\infty + n = +\infty + +\infty = +\infty \ ,$$
$$n \cdot +\infty = +\infty \cdot n = +\infty \qquad \text{for all } n \text{ in } \mathbb{N}_* \ ,$$
$$0 \cdot +\infty = +\infty \cdot 0 = 0 \ , \quad +\infty \cdot +\infty = +\infty \ .$$

(iv) The *same* set $\mathbb{N} \cup \{+\infty\}$ can be equipped with two *other* operations, the min operation as addition, and the $+$ operation as multiplication; this is *another semiring*[22] which we write $\mathcal{M} = \langle \mathbb{N} \cup \{+\infty\}, \mathsf{min}, + \rangle$. All these semirings are commutative. □

Example 5.2 The union (of graphs) and composition (of relations) make $\mathfrak{R}_E$, the set of relations from a set E to itself, a semiring. It is not commutative (when E has more than one element). □

If $\mathbb{K}$ is a semiring, $\mathbb{K} \times \mathbb{K}$ and, more generally, $\mathbb{K}^n$ for every integer n, is a semiring.

Morphisms

A *morphism* of semirings $\varphi \colon \mathbb{K} \to \mathbb{L}$ is a map from $\mathbb{K}$ to $\mathbb{L}$ which is a morphism of monoids both for the additive and the multiplicative structure of $\mathbb{K}$ and $\mathbb{L}$:

$$\forall k, l \in \mathbb{K} \qquad \begin{cases} (k + l)\varphi = k\varphi + l\varphi & \text{and} \qquad 0_{\mathbb{K}}\varphi = 0_{\mathbb{L}} \ , \\ (kl)\varphi = [k\varphi][l\varphi] & \text{and} \qquad 1_{\mathbb{K}}\varphi = 1_{\mathbb{L}} \ . \end{cases}$$

A subset $\mathbb{J}$ of $\mathbb{K}$ is a *sub-semiring* if the inclusion is a morphism of semirings; that is, if $\mathbb{J}$ is a submonoid of $\mathbb{K}$ both for addition and multiplication. In particular, $1_{\mathbb{J}} = 1_{\mathbb{K}}$ and $0_{\mathbb{J}} = 0_{\mathbb{K}}$.

A semiring $\mathbb{K}$ is *positive* if the sum of two non-zero elements of $\mathbb{K}$ is non-zero, in other words if the map $\psi \colon \mathbb{K} \to \mathbb{B}$ such that $k\psi = 1_{\mathbb{B}}$ for all $k \neq 0_{\mathbb{K}}$ (and

[22] A semiring often called *tropical* in the context of automata theory, at least by the Parisian and Brazilian schools (see [236]).

$0_{\mathbb{K}}\psi = 0_{\mathbb{B}}$) is a morphism of semirings; $\mathbb{N}$, $\mathcal{N}$, $\mathcal{M}$, $\mathbb{Q}_+$, $\mathbb{R}_+$ and $\mathfrak{R}_E$ are positive, while $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ are not.

Richer structures

A *ring* is a semiring which is a (commutative) group for addition. We know that $\mathbb{Z}$, $\mathbb{Q}$ and $\mathbb{R}$ are rings.

A ring all of whose non-zero elements are invertible (for multiplication), that is, one which without 0 is a *group* for multiplication, is called a *division ring*, or *skew field*. Such a ring, if it is *commutative*, is a *field*.[23] We know that $\mathbb{Q}$ and $\mathbb{R}$ are fields. We shall have the occasion, in Chapter IV, to construct and use some remarkable skew fields.

Characteristic functions

For any subset F of a set E considered as a fixed 'universe', and generally understood, the *characteristic function* of F, often denoted by $\underline{F}$, is the map from E into $\mathbb{B}$, or into any semiring $\mathbb{K}$, that sends the elements of F onto $1_{\mathbb{K}}$ and the elements of E which do not belong to F onto $0_{\mathbb{K}}$.

Exercises

• 5.1 **Questions.** (a) What is the other semiring with two elements?

(b) Why is $\mathbb{N}_*$ not a semiring?

(c) Why can we not make $\mathcal{Z} = \mathbb{Z} \cup +\infty$ a semiring, as we did for $\mathcal{N}$?

5.2 **An selection of semirings.** We write $\langle E, \oplus, \otimes \rangle$ for the structure comprised of the set E equipped with the operation $\oplus$ for addition and the operation $\otimes$ for multiplication. Verify that the following structures are indeed semirings and find their identity and zero.

(a) $\mathcal{M} = \langle \mathbb{N} \cup \{+\infty\}, \mathsf{min}, + \rangle$ (Exam. 5.1 (iv)); and also the three other 'tropical' structures: $\langle \mathbb{Z} \cup \{-\infty\}, \mathsf{max}, + \rangle$, $\langle \mathbb{Q} \cup \{-\infty, +\infty\}, \mathsf{max}, \mathsf{min} \rangle$, $\langle \mathbb{R} \cup \{-\infty, +\infty\}, \mathsf{min}, \mathsf{max} \rangle$.

(b) Let E be a set. Answer the same questions for $\langle \mathfrak{P}(E), \cup, \cap \rangle$, for $\langle \mathfrak{P}(E), \cap, \cup \rangle$, and for $\mathfrak{R}_E$ (Example 5.2).

(c) Let M be a monoid. Answer the same questions for $\langle \mathfrak{P}(M), \cup, \cdot \rangle$, where multiplication is defined as in Exercise 2.3.

(d) Show that $\mathcal{M}$ is a sub-semiring of $\langle \mathfrak{P}(a^*), \cup, \cdot \rangle$.

• 5.3 **Centre.** The *centre* of a semiring $\mathbb{K}$, written $\mathbb{K}_{\mathsf{c}}$, is the set of elements of $\mathbb{K}$ which commutes with all the elements of $\mathbb{K}$. Show that $\mathbb{K}_{\mathsf{c}}$ is a sub-semiring of $\mathbb{K}$.

[23] I cannot resist the pleasure of quoting P. M. Cohn [65] on this subject: 'Here is a case (rare but not unknown even in non-technical English) where the range of a noun is restricted by *omitting* a qualifying adjective: a field is a particular kind of skew field, just as tea is a particular kind of leaf tea.'

6 Matrices

A matrix is not, properly speaking, a *structure* in the algebraic sense of the term, as a monoid or semiring is; it is rather a mathematical object, or notational convention, which we recall here because we shall use it frequently, especially from Chapter III onwards. The non-mathematical reader should not be nervous or discouraged; there is nothing difficult about it.[24]

It will be convenient to consider that vectors and matrices are arrays indexed not by integers or pairs of integers but but elements or pairs of elements taken from a *set* or the product of two sets.

Let P and Q be two sets. A *matrix* of dimension $P \times Q$ with entries in a set E is a set of elements of E indexed by $P \times Q$. A *square matrix* of dimension Q is a set indexed by $Q \times Q$. Let p be in P and q in Q; we write $m_{p,q}$ for the entry with index (p,q) of the matrix m of dimension $P \times Q$. If P has a single element, we also say that m is a *row vector*; if Q has a single element, that m is a *column vector*.

The notion of a matrix only really makes sense if the set E of entries is equipped with a structure, which is then inherited by the set of matrices of a given dimension. The addition of matrices of dimension $P \times Q$ (and with entries in a commutative monoid) and the multiplication of a matrix of dimension $P \times Q$ by a matrix of dimension $Q \times R$ (where Q is finite and the matrix's entries are in a semiring) are defined as usual:

$$\begin{aligned} &\forall p,q \in P \times Q \qquad (m+n)_{p,q} = m_{p,q} + n_{p,q} \ , \\ &\forall p,r \in P \times R \qquad (m \cdot n)_{p,r} = \sum_{q \in Q} (m_{p,q} \cdot n_{q,r}) \ . \end{aligned}$$

Thus, the square matrices of dimension Q with entries in a semiring $\mathbb{K}$ form a semiring, written $\mathbb{K}^{Q \times Q}$.

Relations and boolean matrices

For example, if $\mathbb{K} = \mathbb{B}$, the Boolean semiring, the square matrices (of dimension Q) with entries in $\mathbb{B}$ are called *boolean matrices* and form a semiring (which we have indeed already met). Let α be a relation from a finite set Q to itself, identified with its graph $\widehat{\alpha}$, a subset of $Q \times Q$. We naturally associate α with a boolean matrix, say $\mathsf{m}(\alpha)$:

$$\mathsf{m}(\alpha)_{p,q} = \begin{cases} 1 & \text{if} \quad (p,q) \in \widehat{\alpha} \\ 0 & \text{otherwise} \end{cases}$$

and, conversely, every boolean matrix of dimension Q defines a relation from Q to itself, that is, an element of $\mathfrak{R}_Q$. We easily verify:

[24] *Note added in translation:* I have been warned by readers of the French edition that the non-mathematical reader would never have reached this point. This is rather unfair since, as said in the introduction, this chapter is only to be read when needed.

Property 6.1 *The natural bijection between* $\mathfrak{R}_Q$ *and* $\mathbb{B}^{Q\times Q}$ *is an isomorphism of semirings.* ■

Matrices will be a basic tool for computations and constructions used in this book, starting in Chapter III, and the reader is invited to become familiar with their manipulation, if necessary (Exer. 6.1 to 6.6).

Block decomposition

Since $\mathbb{K}^{Q\times Q}$ is a semiring, we can take its elements as the entries of another square matrix, of dimension R, say; we thus obtain a matrix in the semiring $(\mathbb{K}^{Q\times Q})^{R\times R}$. This matrix can also be seen as an element of $\mathbb{K}^{(Q\times R)\times(Q\times R)}$. Conversely, every matrix in the semiring $\mathbb{K}^{(Q\times R)\times(Q\times R)}$ can be considered as a square matrix of dimension R with entries in the semiring of square matrices of dimension Q with entries in $\mathbb{K}$.

It is this going from an element m of $\mathbb{K}^{(Q\times R)\times(Q\times R)}$ to the corresponding element of $(\mathbb{K}^{Q\times Q})^{R\times R}$ that we call the *block decomposition* of the matrix m and which has the property of being 'transparent' to the operations of the semiring. We can furthermore generalise the process to a decomposition where the blocks are not necessarily all square matrices. *cf. Exer. 6.9*

Exercises

6.1 Compute the product $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$.

6.2 Compute the products of the following matrices ($\mathbb{K} = \mathbb{Z}$):

(a) $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ (b) $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$ (c) $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$

(d) $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$ (e) $\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -2 & 1 \end{pmatrix}$

6.3 **Diagonal matrix.** A matrix m in $\mathbb{K}^{Q\times Q}$ is called *diagonal* if $m_{p,q} = 0$ when $p \neq q$. Compute m^r, for all integers r. Verify that the set of diagonal matrices is a semiring isomorphic to the direct product of $\|Q\|$ copies of $\mathbb{K}$. Comment.

6.4 **Monomial matrix.** A matrix is called *row monomial* (resp. *column monomial*) if each of its rows (resp. columns) contains at most one non-null element. Verify that the product of two row-monomial (resp. column-monomial) matrices is a row-monomial (resp. column-monomial) matrix.

6.5 **Triangular matrix.** Suppose that $Q = [n]$ is the *ordered* set of integers from 1 to n.

(a) A matrix m in $\mathbb{K}^{Q\times Q}$ is called *upper triangular* if $m_{p,q} = 0$ when $p > q$. Verify that the set of upper triangular matrices is a sub-semiring of $\mathbb{K}^{Q\times Q}$. Compute the diagonal of m^r, for all integers r.

(b) A matrix m in $\mathbb{K}^{Q\times Q}$ is called *strict upper triangular* if $m_{p,q} = 0$ when $p \geqslant q$. Verify that the set of strict upper triangular matrices is a sub-semiring of $\mathbb{K}^{Q\times Q}$. Compute $m^{\|Q\|}$.

6.6 **Transpose of a matrix.** The *transpose* of a matrix m of dimension $P \times Q$ is the matrix of dimension $Q \times P$, written m^{t}, and defined by:

$$\forall p \in P\,,\ \forall q \in Q \qquad (m^{\mathsf{t}})_{q,p} = m_{p,q} \ .$$

Verify that if m and n are two matrices (one multipliable by the other), we have

$$(m \cdot n)^{\mathsf{t}} = n^{\mathsf{t}} \cdot m^{\mathsf{t}} \ .$$

6.7 **Trace of a matrix.** The *trace* of a matrix m in $\mathbb{K}^{Q \times Q}$, written $\mathsf{Tr}(m)$, is the element of $\mathbb{K}$ defined by

$$\mathsf{Tr}(m) = \sum_{q \in Q} m_{q,q} \ .$$

We obviously have $\mathsf{Tr}(m + m') = \mathsf{Tr}(m) + \mathsf{Tr}(m')$. Verify that if $\mathbb{K}$ is *commutative*, we have

$$\forall m \in \mathbb{K}^{Q \times R}\,,\ \forall m' \in \mathbb{K}^{R \times Q} \qquad \mathsf{Tr}(m\,m') = \mathsf{Tr}(m'\,m) \ . \tag{6.1}$$

6.8 Verify Property 6.1.

6.9 Let $m = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ a matrix decomposed into four blocks, where a is a square matrix of dimension Q and d is a square matrix of dimension R and hence b is a matrix of dimension $Q \times R$ and c is a matrix of dimension $R \times Q$. Let $m' = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$ a square of the same dimension $Q \cup R$ decomposed into blocks in the same way. Verify that we can form the product $m \cdot m'$ from the product of the blocks.

Generalise to the case of square matrices of dimension $P = \bigcup_{i \in I} Q_i$.

• 6.10 **Faithful representations of the free monoid.**

(a) Let $A = \{a, b\}$. Show that the morphism $\alpha \colon A^* \to \mathbb{N}^{2 \times 2}$ defined by

$$a\alpha = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad b\alpha = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \ ,$$

is *injective* (that is, α is a *faithful representation* of A^* by (2×2) matrices over $\mathbb{N}$).

(b) Prove the same property for the morphism $\beta \colon A^* \to \mathbb{N}^{2 \times 2}$ defined by

$$a\beta = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \qquad \text{and} \qquad b\beta = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \ .$$

(c) Generalise to a morphism $\zeta \colon Z^* \to \mathbb{N}^{2 \times 2}$ in the case where Z contains k letters.

• 6.11 Let M be a monoid. Let $\sigma \colon M \to (\mathfrak{P}(M) \times \mathfrak{P}(M))^{2 \times 2}$ be the map defined by

$$\forall m \in M \qquad (m)\sigma = \begin{pmatrix} (m, 1) & \{(u, v) \mid u\,v = m\} \\ 0 & (1, m) \end{pmatrix} \ .$$

Verify that σ is a morphism when M is equidivisible.

7 Lexicon of graph theory

The automata that we shall study are *graphs*, or rather, we shall define them as such. We shall obviously give all the necessary definitions of graph theory needed for our purposes, and not use any results that we have not proved.[25]

It seems nonetheless useful to recall some basic definitions of graph theory, particularly those for which the automata-theoretic terminology differs. We will then be able to indicate the graph-theoretic counterpart of some automata-theoretic results.

A *(directed) graph* $\mathcal{G}$ is defined by giving two sets V and E: V is the set of *vertices* of $\mathcal{G}$, and E the set of its *edges*. Each edge in $\mathcal{G}$ has a *source* and a *destination* which are vertices of $\mathcal{G}$. This is the same as saying that as well as V and E, $\mathcal{G}$ is defined by two functions ι and τ from E to V which relate each edge respectively to its source and its destination. Graphs are called graphs because we draw them: we represent, as in Figure 7.1, vertices by circles and edges by arrows which join the source to the destination.

Figure 7.1: A multigraph and the corresponding simple graph

If for each pair of vertices (x, y) there is at most one edge in $\mathcal{G}$ which goes from x to y – in which case $\mathcal{G}$ is called a *simple graph* or 1-*graph*, as opposed to a *multigraph* in the contrary case – we can identify E with a subset of $V \times V$, a *relation* from V to itself. If $\mathcal{G}$ is a multigraph, we can naturally associate a simple graph with it by replacing all the edges which have the same source and destination with just one. A simple graph $\mathcal{G} = \langle V, E \rangle$ is *finite* if the set V is finite and it is then natural to represent $\mathcal{G}$ by a boolean matrix of size V, called the *incidence matrix* of $\mathcal{G}$: the entry (s, t) of the matrix is 1 if there exists (at least) one edge with source s and destination t, and 0 otherwise. For example, the incidence matrix of the simple graph in Figure 7.1 is

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} .$$

A *path* in $\mathcal{G} = \langle V, E \rangle$, of *length* n, is a sequence of n edges $(v_1, v_2, \ldots, v_n)$ such that for each i less than n, the destination of v_i is the source of v_{i+1}. It is consistent with this definition to consider each vertex of $\mathcal{G}$ to be both the source and destination of a path of length 0.

A *circuit* is a path the destination of whose last edge is the source of the first. The relation 'there exists a path from s to t' is the *transitive closure* of the relation defined by $\mathcal{G}$ on V. We also call the corresponding graph the transitive closure of $\mathcal{G}$.

cf. Sec. 1 and Prop. 1.2

[25] There is therefore no reason to have exercises in this section. There will be time for those relevant to automata later; those not bearing on automata are useless to us.

Given a graph $\mathcal{G} = \langle V, E \rangle$, we consider the relation Γ on V defined by $x \Gamma y$ if and only if there exists a circuit which contains both x and y, or, which amounts to the same thing, if there exists a path which goes from x to y *and* a path which goes from y to x. The relation Γ is clearly both symmetric and transitive, and its reflexive extension is an equivalence relation called *(the relation of) strong connectedness*. The classes of V modulo this relation are the *strongly connected components* of $\mathcal{G}$. In the graph in Figure 7.1, the strongly connected components are reduced to singletons; that is, the relation of strong connectedness is reduced to the identity.

In the last section, we recalled that not only do relations on a set Q and boolean matrices of dimension Q correspond bijectively but also that under this identification the composition of relations corresponds to matrix multiplication: the bijection is structural. Relations are composed, matrices are multiplied, and graphs are *traversed*, that is to say, we compute their paths; and we observe that the correspondence between graphs and boolean matrices, via incidence matrices, is also structural – a fundamental lemma to which we shall return several times.

cf. Propty 6.1

Lemma 7.1 *Let $\mathcal{G} = \langle V, E \rangle$ be a finite simple graph and G its incidence matrix. Then G^k is the incidence matrix of paths of length k (that is, $(G^k)_{s,t} = 1$ if and only if there exists a path of length k which goes from s to t); and the reflexive and transitive closure of $\mathcal{G}$ is given by the incidence matrix:*

$$T = \sum_{k=0}^{k=n-1} G^k \quad , \tag{7.1}$$

where $n = \|V\|$.

Proof. By induction on k. The assertion is true for $k = 1$ (and also for $k = 0$ by convention). The definition of the $(k+1)$th power of G:

$$\forall s, t \in V \qquad (G^{k+1})_{s,t} = \sum_{u \in V} (G^k)_{s,u} \, G_{u,t} \quad ,$$

expresses that every path of length $k+1$ is the concatenation of a path of length k with an edge. Elementary reasoning ensures that if there exists a path between s and t then there exists at least one of length less than n, hence (7.1). ■

cf. § I.1.3.2, p. 69

8 Complexity and decidability

One could write a whole book on this subject. What am I saying, a book? whole shelves of books! We aim therefore only to introduce a naive point of view with which we shall tackle these two eminently subtle questions. We must nevertheless attack these questions, as our subject, the theory of automata, is intrinsically linked to them. Its major problems are formulated, its most beautiful results take their meaning, only within a framework put in place by these concepts, fortunately in their

most elementary form, defined by the two theories of complexity and decidability. To underscore this, I would like to begin by quoting S. Eilenberg:[26]

> *One very characteristic feature of the subject should be mentioned here. All arguments and proofs are constructive. A statement asserting that something exists is of no interest unless it is accompanied by an algorithm for producing this "something." Thus each proof is also an algorithm. Usually, but not always, the simplest proof also gives the most efficient algorithm.*

I will add to this on two levels. I am inclined to think that it is the proof that corresponds to the efficient algorithm which lays bare the essence of the property which one wants to prove, without artefact or detour. Also, if the proofs of two distinct properties translate into the same algorithm, then there exists a sort of 'parent structure' and a result on this structure of which our two initial properties are merely instances or corollaries.

These assertions are not inviolable universal properties, and they can doubtless be contradicted by examples taken from this very work; but they express a philosophy, a habit of thought, and are a precious guide for unearthing and giving shape to the fundamental notions.

In any case we need to explain, at least briefly, what the 'efficiency' of an algorithm is, and, on the other hand, what a situation looks like when there is no algorithm.

Constructive proofs and complexity

Looking more closely at Lemma 7.1 above, we see that its statement conceals two results, of different kinds. The first is that the transitive closure of a graph, which assumes *a priori* that the infinite set of all the paths in a graph is examined, can be computed by an algorithm, hence by a finite number of operations. The second is that Equation (7.1) is itself an algorithm, the description of a way of computing the matrix T, and we can try to calculate the number of elementary operations that this takes.

The computation of each of the n^2 entries of the product of two matrices of size n requires $2\,n$ elementary operations; the entire product thus requires $2\,n^3$, and that of the matrix T a number given by a polynomial in n of degree 4. In this case we say that the algorithm, described by (7.1), 'is $O(n^4)$'.

As explained in the first chapter of all books on algorithms, as well as all elementary analysis texts, this notation, called 'big-oh notation', represents a precise idea. If f and g are functions from $\mathbb{N}$ to $\mathbb{N}$ or from $\mathbb{R}$ to $\mathbb{R}$, then g is called '$O(f)$' if there exist constants c and K such that $g(x) \leqslant c\,f(x)$ for all x greater than K. We thus describe equivalence classes on the set of functions that enable us to quantify the complexity of a procedure without worrying about the details of its implementation.

[26] From the preface to *Automata, Languages and Machines* [80].

Note that we have in this way defined time complexity, the evaluation of the number of elementary operations performed, as opposed to space complexity, the evaluation of the quantity of memory necessary to realise these operations in a machine, to which we shall not refer. Note also that this complexity corresponds to a worst-case evaluation. We start from the size of an object, measured by one or more parameters. (For the graph $\mathcal{G}$, this was the number n of vertices; we could have considered the size of $\mathcal{G}$ to be measured by the pair (n, m) where m is the number of edges.) We then compute an upper bound on the number of operations performed by the algorithm when it is executed on any object of this size.

This mode of evaluation, which could seem rather clumsy, turns out to be effective, both for practical and theoretical purposes. In our example, what counts is that besides (7.1) we can find ways of computing T whose complexity is $O(n^3)$ or even $O(mn)$ and which lead to a better description of the operations which are really being performed.

Undecidability of the Post Correspondence Problem

In the same way that the most interesting propositions are generally effective, i.e. that a result $A = B$ is in fact an algorithm that starts with the datum A, or with a one of type A, transforms it into the datum B, or into one of type B, to know that such an algorithm *cannot exist* is in itself a result, albeit negative, an admission of failure, but nevertheless a contribution to our knowledge of A and B.

The notion of an *algorithmically undecidable property* or *recursively undecidable property*,[27] due to Alan Turing, is certainly one of the most profound concepts discovered by mathematicians in the twentieth century.

The proof that a given property (P) is undecidable always consists of the *reduction* of this property to another, (Q), already known to be undecidable. We show that (Q) can be transformed *effectively* into the property (P). If this last has an algorithmic solution, we could deduce one for (Q), which is impossible. On the contrary, property (P) is *decidable* if it has an algorithmic solution.

The undecidable problem *par excellence* is the halting problem for Turing machines, but we can use as our starting point any other that is already known to be undecidable. The one we shall use is known as the Post Correspondence Problem, abbreviated to (PCP):

Let B be an alphabet containing at least two letters. Given an integer k and sets $\{u_1, \ldots, u_k\}$ and $\{v_1, \ldots, v_k\}$ of k words from B^*, does there exist a sequence of indices $i_1, \ldots, i_p$ in $[k]$ such that $u_{i_1} u_{i_2} \cdots u_{i_p} = v_{i_1} v_{i_2} \cdots v_{i_p}$?

[27] Which should not be confused with that of a *logically undecidable property*, due to Kurt Gödel, and also a fundamental concept. An algorithmically undecidable property is either true or false, but no algorithm can be constructed which gives the answer. A logically undecidable property can be considered to be either true or false without being to deduce from either hypothesis any contradiction with the rest of mathematics. We only consider *recursively undecidable* properties, which we call simply *undecidable*.

Theorem 8.1 [**Post**] *(PCP) is undecidable.* ■

This proposition holds for the problem in all generality. If we want to know its status as a function of the number k of words which enables us to find a solution, the situation is more complex. Call (PCP_k) the above problem in which the integer k is fixed. It is known that (PCP_2) is decidable and that (PCP_k) is undecidable for $k \geqslant 7$. The status of (PCP_k) is unknown for k between 3 and 6 inclusive. *cf. bibliographic notes*

The reason for our choice is not only that the Post problem can be stated in terms of words, which are one of the fundamental objects with which we shall work, but also that it takes a particularly simple form in terms of *morphisms between free monoids*. Let $\theta\colon A^* \to B^*$ be such a morphism; in this context we shall use a new notation for what is almost the graph of θ:

$$U_\theta = \{(f, f\theta) \mid f \in A^+\} = \widehat{\theta} \cap A^+ \times A^* \ .$$

We consider an instance of (PCP_k): an alphabet B (with at least two letters) and two sets $\{u_1, \ldots, u_k\}$ and $\{v_1, \ldots, v_k\}$ of k words in B^*. Write

$$\mathbf{k} = \{0, 1, \ldots, k-1\}$$

for the canonical alphabet with k letters and define two morphisms $\theta\colon \mathbf{k}^* \to B^*$ and $\mu\colon \mathbf{k}^* \to B^*$ by $i\theta = u_{i+1}$ and $i\mu = v_{i+1}$ for all i in $\mathbf{k}$. Saying that there exists a sequence of indices $i_1, i_2, \ldots, i_p$ such that $u_{i_1} u_{i_2} \cdots u_{i_p} = v_{i_1} v_{i_2} \cdots v_{i_p}$ is exactly the same as saying that there exists a word $f = i_1 i_2, \cdots i_p$ in $\mathbf{k}^+$ such that $f\theta = f\mu$ and hence is equivalent to the inequality

$$U_\theta \cap U_\mu \neq \emptyset \ . \tag{8.1}$$

Conversely, if $\theta\colon A^* \to B^*$ and $\mu\colon A^* \to B^*$ are morphisms, then Equation (8.1) is an instance of the Post problem. In other words, we can reformulate Theorem 8.1 in the following manner.

Theorem 8.2 [**Post**] *Let θ and μ be two morphisms from A^* to B^*. It is undecidable whether there exists a word f in A^+ such that $f\theta = f\mu$.* ■

We shall use the following terminology and definitions in the rest of the book. Let E be a set, and $\mathcal{F}$ a family of subsets of E, *described in an effective manner*; we further suppose that E belongs to $\mathcal{F}$.

The *intersection* is called decidable (or undecidable) in $\mathcal{F}$ if it is decidable (or undecidable) whether the intersection of two effectively given elements of $\mathcal{F}$ is *empty* or not.

The *universality problem*, (or, for brevity, *universality*), is decidable (or undecidable) in $\mathcal{F}$ if it is decidable (or undecidable) whether an effectively given element of $\mathcal{F}$ is *equal to all of* E or not.

The *equivalence problem*, (or *equivalence*), is decidable (or undecidable) in $\mathcal{F}$ if it is decidable (or undecidable) whether two effectively given elements of $\mathcal{F}$ are *equal* or not.

The *inclusion problem*, (or *inclusion*), is decidable (or undecidable) in $\mathcal{F}$ if it is decidable (or undecidable) whether two effectively given elements of $\mathcal{F}$ are *included one in the other* or not.

If inclusion is decidable, so is equivalence and hence universality. Contrariwise, if universality is undecidable, so are equivalence and inclusion. However, there exist cases in which inclusion is undecidable but equivalence is decidable.

cf. Prop. IV.5.20, p. 589

If $\mathcal{F}$ and $\mathcal{G}$ are effectively described families of subsets of a set E such that $\mathcal{G} \subseteq \mathcal{F}$, we say that $\mathcal{G}$ is *decidable* (or *undecidable* in $\mathcal{F}$ if it is decidable (or undecidable) whether an effectively given element of $\mathcal{F}$ belongs to $\mathcal{G}$.

Finally, if $\mathcal{F}$ is a Boolean algebra of subsets of E, we say that it is an *effective Boolean algebra* if, (effectively) given elements X and Y of $\mathcal{F}$, we can effectively compute a description of $X \cup Y$ and of $\complement X$, and if we can decide whether X is empty or not.

SOLUTIONS TO THE EXERCISES

1. Relations

1.3 *Let $\alpha\colon E \to F$ be a function, and let $P, Q \subseteq F$ and $A \subseteq E$ such that $A\alpha = Q$. Show that $P \bigcap Q = (P\alpha^{-1} \cap A)\alpha$.*

Solution: Since $(P\alpha^{-1})\alpha = P$ and $A\alpha = Q$, $P \bigcap Q \supseteq (A \cap P\alpha^{-1})\alpha$. Conversely

$$f \in P \cap Q \quad \Longrightarrow \quad f \in Q \quad \Longrightarrow \quad \exists e \in A \quad e\alpha = f \quad \Longrightarrow$$
$$e \in P\alpha^{-1} \quad \Longrightarrow \quad f \in (A \cap P\alpha^{-1})\alpha \ .$$

1.5 *Let $\alpha\colon E \to F$ be a function. Characterise the relation $\alpha\alpha^{-1}$ in such a way that it becomes obvious that it is an equivalence relation on E.*

Solution: We have $(e', e'') \in \widehat{\alpha\alpha^{-1}}$ if and only if $e'\alpha = e''\alpha$ and equality (of the images) is obviously reflexive, symmetric and transitive.

1.6 *Let $\alpha\colon E \to F$ be a relation. Verify that α is a function if and only if $\alpha^{-1}\alpha$ is the restriction of the identity on $\mathsf{Im}\,\alpha$.*

Solution: There exists $(f, g) \in \widehat{\alpha^{-1}\alpha}$, with $f \neq g$, if and only if there exists $e \in E$ such that $e \in f\alpha^{-1}$ and $g \in e\alpha$ and, since $f \in e\alpha$, if and only if $\{f, g\} \subseteq e\alpha$; that is, if and only if α is not a function.

1.9 ***Generated Boolean algebras.*** *Let $\mathcal{F}$ be a set of subsets of a set E: $\mathcal{F} \subseteq \mathfrak{P}(E)$.*

(a) *Characterise the atoms of the Boolean algebra generated by a set $\mathcal{F}$.*

(b) *If $\mathcal{F}$ is finite and contains n subsets of E, what is the largest possible number of elements (that is, of subsets of E) of the Boolean algebra generated by $\mathcal{F}$?*

Solution: (a) The family $\mathcal{F}$ allows us to define an equivalence on E: two elements e and f of E are equivalent (or *indistinguishable*) modulo $\mathcal{F}$ if they belong to the same elements of $\mathcal{F}$:

$$e \equiv f \bmod \mathcal{F} \quad \Longleftrightarrow \quad (\forall F \in \mathcal{F} \qquad e \in F \Leftrightarrow f \in F) \ ,$$

from which we deduce that the classes for this equivalence are

$$[e]_{\mathcal{F}} = \left(\bigcap_{\substack{F \in \mathcal{F} \\ e \in F}} F \right) \setminus \left(\bigcup_{\substack{G \in \mathcal{F} \\ e \notin G}} G \right) . \tag{SE.1.1}$$

with the (natural) convention that

$$\bigcap_{\emptyset} F = E \ , \qquad \text{which is the counterpart of} \qquad \bigcup_{\emptyset} F = \emptyset \ .$$

The classes $[e]_{\mathcal{F}}$, e in E, thus form a partition of E and are the atoms of the Boolean algebra $\mathcal{C}$ which they generate: $\mathcal{C} \subseteq \mathcal{B}$. Each F in $\mathcal{F}$ either contains or is disjoint from each of the $[e]_{\mathcal{F}}$: it is the union of the $[e]_{\mathcal{F}}$ which it contains, and belongs to $\mathcal{C}$. Hence $\mathcal{B} \subseteq \mathcal{C}$.

(b) By the foregoing, each atom of $\mathcal{B}$ is completely determined by the set of F's in which it is contained. There are thus at most 2^n atoms in $\mathcal{B}$. Each element of $\mathcal{B}$ is a union of atoms: there are at most 2^{2^n} elements in $\mathcal{B}$. This bound is tight for families of sets 'in general position'.

2. Monoids

2.1 *What is the other monoid with two elements?*

Solution: $\mathbb{Z}/2\mathbb{Z} = \{0, 1\}$ (additive!): 0 is the neutral element and $1 + 1 = 0$.

$\boxed{\mathbf{2.2}}$ *How many elements do $\mathfrak{R}_E$ and $\mathfrak{T}_E$ have, if $\|E\| = n$?*

Solution: To define a relation from E to itself we must choose for each element of E a subset of E, hence $\|\mathfrak{R}_E\| = (2^n)^n = 2^{(n^2)}$. To define a map from E to itself, we must choose for each element of E an element of E, hence $\|\mathfrak{T}_E\| = n^n$. There are n^{n+1} functional relations from E to itself.

$\boxed{\mathbf{2.3}}$ *Verify that $\mathfrak{P}(M)$ equipped with the operation $P \cdot Q = \{p \cdot q \mid p \in P, q \in Q\}$ is a monoid. What is its neutral element? Does it have a zero?*

Solution: Associativity follows directly from that of M; $1_{\mathfrak{P}(M)} = \{1_M\}$, $0_{\mathfrak{P}(M)} = \emptyset$.

$\boxed{\mathbf{2.4}}$ *Show that an equivalence relation ε on M is a congruence if and only if*

$$\forall p, q, r \in M \qquad p \equiv q \bmod \varepsilon \quad \Longrightarrow \quad \{\, p\,r \equiv q\,r \bmod \varepsilon \quad \text{and} \quad r\,p \equiv r\,q \bmod \varepsilon \,\}\ .$$

Solution: The condition is obviously necessary. It is also sufficient, since if we have

$$p \equiv q \bmod \varepsilon \quad \text{and} \quad s \equiv t \bmod \varepsilon\ , \qquad \text{we deduce} \qquad p\,s \equiv q\,s \equiv q\,t \bmod \varepsilon\ .$$

$\boxed{\mathbf{2.8}}$ ***Product.*** (a) *Verify that the direct product $M \times N$ is 'naturally' a monoid. What is its neutral element? Under what conditions does it have a zero?*

(b) *Show that for every monoid T and every pair of morphisms $\varphi\colon T \to M$ and $\psi\colon T \to N$ there exists a unique morphism, written $\varphi \times \psi$: $\varphi \times \psi\colon T \to M \times N$ which makes the adjacent diagram commute (where we write π_M and π_N for the projections from $M \times N$ to M and N respectively).*

Characterise $\varphi \times \psi$.

(c) *Show that conversely every monoid with this property with respect to M and N is isomorphic to $M \times N$.*

$$\begin{tikzcd} & M \\ T \arrow[ur, "\varphi"] \arrow[r, "\varphi\times\psi"] \arrow[dr, "\psi"'] & M \times N \arrow[u, "\pi_M"'] \arrow[d, "\pi_N"] \\ & N \end{tikzcd}$$

Solution: (a) The operation on $M \times N$ defined by $(m, n)\,(m', n') = (m\,m', n\,n')$ is associative. We have $1_{M\times N} = (1_M, 1_N)$; $0_{M\times N}$ exists if M and N have a zero: $0_{M\times N} = (0_M, 0_N)$.

(b) For all t in T, $t\,\varphi \times \psi = (t\varphi, t\psi)$.

(c) In the diagram, take $T = M \times N$, $\varphi = \pi_M$ and $\psi = \pi_N$: the isomorphism follows.

$\boxed{\mathbf{2.9}}$ *In general, which submonoids of $M \times M$ are isomorphic to M?*

Solution: They are: $1_M \times M$, $M \times 1_M$, *and* the diagonal $D = \{(m, m) \mid m \in M\}$.

$\boxed{\mathbf{2.11}}$ ***Normal subgroups.*** *A subgroup H of a group G is called* normal *if for every g in G we have: $g\,H = H\,g$. Verify that the congruences of G are in bijection with its normal subgroups.*

Solution: If α is a congruence of G, the class of 1_G modulo α is a normal subgroup H. Indeed, H contains 1_G, is closed under product, since $1_G \cdot 1_G = 1_G$, and under inversion, since $(1_G)^{-1} = 1_G$. The following sequence of equivalences shows that H is normal: for all g and k in G, we have

$$k \in H\,g \quad \Leftrightarrow \quad k\,g^{-1} \in H \quad \Leftrightarrow \quad k\,g^{-1} \equiv 1_G \bmod \alpha \quad \Leftrightarrow \quad k \equiv g \bmod \alpha \quad \Leftrightarrow$$
$$g^{-1}\,k \equiv 1_G \bmod \alpha \quad \Leftrightarrow \quad g^{-1}\,k \in H \quad \Leftrightarrow \quad k \in g\,H\ .$$

Conversely, if H is a normal subgroup of G, we define a relation α by

$$g \equiv k \bmod \alpha \quad \Longleftrightarrow \quad g\,H = k\,H\ .$$

By its definition, α is an equivalence; it is a congruence since

$$\left.\begin{array}{r} g\,H = k\,H \\ g'\,H = k'\,H \end{array}\right\} \quad \Longrightarrow \quad g\,H\,g'\,H = g\,g'\,H\,H = g\,g'\,H = k\,H\,k'\,H = k\,k'\,H\,H = k\,k'\,H\ .$$

3. Words and languages

3.1 *How many letters has the alphabet in which English words are written?*

Solution: Because it relates to language, a subtle domain, this banal-seeming question is most delicate. Of course, English uses the Roman alphabet, which consists of twenty-six letters. However, to *write words* dictionaries use upper-case as well as lower letters, and accented letters such as æ and é (to write loan-words from other languages), - (as in 'great-grandfather'), ' (as in 'isn't') and even the space (as in 'half tide' – if we consider it to be *one* word), or at least fifty-seven characters. Depending on the dictionary, it may have more.

3.2 *If the alphabet A has k letters, how many words of length n are there in A^*? How many of length n which contain l occurrences of a given letter in A?*

Solution: k^n ; $C_n^l (k-1)^{n-l}$.

3.3 *How many distinct factors of length* 3, 4 *and* 5 *are there in the following three words of length* 20*?*

(a) $f = a\,b\,a\,a\,b\,a\,b\,a\,a\,b\,a\,a\,b\,a\,b\,a\,a\,b\,a\,b$; (b) $g = a\,a\,a\,b\,a\,b\,b\,b\,b\,a\,a\,b\,b\,a\,b\,a\,a\,a\,a\,b$;

(c) $h = a\,b\,a\,a\,a\,b\,a\,a\,a\,b\,a\,a\,a\,b\,a\,a\,a\,b\,a\,a$.

Solution: (a) 4; 5; 6. (b) 8; 16; 16. (c) 4; 4; 4.

3.7 *What is* $\mathsf{Pre}(Z_1)$ *?*

Solution: $\mathsf{Pre}(Z_1) = A^*$.

3.8 *Describe* $\mathsf{Pal} \cap Z_1 \cap a^*b^*a^*b^*a^*$.

Solution:
$$\mathsf{Pal} \cap Z_1 \cap a^*b^*a^*b^*a^* = \{a^p b^q a^r b^q a^p \mid p,q,r \in \mathbb{N} \quad 2q = 2p + r\} \cup \{a^p b^q a^p \mid p,q \in \mathbb{N} \quad q = 2p\} .$$

3.9 *Let* $B_1 = \{a\,b, a\,b\,a^2 b, \ldots, a\,b\,a^2 b \cdots b\,a^{n-1} b\,a^n b, \ldots\}$. *Describe* $A^*b \setminus B_1$.

Solution: $A^*b \setminus B_1 = \{a^{k_1} b\, a^{k_2} b \cdots a^{k_{n-1}} b\, a^{k_n} b \mid n \in \mathbb{N}_*\,,\ \exists i \in [n] \quad i \neq k_i\}$.

3.10 ***Lexicographic order.*** *Let A be a* totally ordered *alphabet.*

(a) *Verify that $\preccurlyeq$ is a total order, called the* lexicographic order, *but is not a well ordering.*

(b) *Explain why the words in an English dictionary* are not *in lexicographic order. Describe the order.*

Solution: (a) If $f \neq g$ one of the following four conditions is satisfied: (i) $f = g\,h$ or (ii) $g = f\,h$ with $h \in A^+$; or $f = u\,a_i v$, $g = u\,a_j w$ with (iii) $i < j$ or (iv) $j < i$. Lexicographic order is a total order because $g \preccurlyeq f$ in cases (i) and (iv), and $f \preccurlyeq g$ in cases (ii) and (iii).

Lexicographic order is not a well ordering because in particular the set a^*b , with $a < b$, does not have a smallest element:
$$b \succ a\,b \succ a\,a\,b \succ \cdots \succ a^n b \succ \cdots$$

(b) It is the punctuation characters: -, ' and space that are problematic, and the accented letters as well. Consider for example the following sequence of entries from the dictionary:[28]

godfather **god-fearing** **godforsaken** .

It is not ordered lexicographically: if **f** $\prec$ **-**, **god-fearing** should follow **godforsaken**; if the reverse, **god-fearing** should precede **godfather**.

[28] In this case, the *Robert & Collins Dictionary.*

The dictionary order is obtained by removing all diacritics: hyphens, inverted commas and spaces are replaced by the empty word, and accented letters are replaced by non-accented letters; 'æ' is replaced by 'ae', and 'œ' by 'oe'. The words thus obtained are then ordered lexicographically, and the deleted marks re-added. When the removal of accents makes two words the same, the accented word follows the unaccented word.

4. Free monoids

4.1 (a) *Let* $\gamma\colon \{a,b\}^* \to \mathbb{Z}$ *be the morphism defined by* $a\gamma = +1$ *and* $b\gamma = -1$ *(where* $\mathbb{Z}$ *is a monoid for addition). Show that* $(0)\gamma^{-1} = Z_1$.

(b) *Let* $\alpha\colon \{a,b\}^* \to \mathbb{Z}$ *be the morphism defined by* $a\alpha = +2$ *and* $b\alpha = -3$. *Characterise* $\mathsf{Im}\,\alpha$ *and* $(0)\alpha^{-1}$.

Solution: (a) By induction on the length of f in $\{a,b\}^*$, we obtain $f\gamma = |f|_a - |f|_b$ by which $(0)\gamma^{-1} = \{f \mid |f|_a = |f|_b\} = Z_1$. (b) $\mathsf{Im}\,\alpha = \mathbb{Z}$; $0\alpha^{-1} = \{f \mid 2\,|f|_a - 3\,|f|_b = 0\}$.

4.3 *Prove Property 4.2: for all* f, g *in* A^*, $(f\,g)^{\mathsf{t}} = g^{\mathsf{t}} f^{\mathsf{t}}$.

Solution: By induction on the length of g. The property is true, for all f, for $|g| = 0$ and $|g| = 1$; then $(f\,g\,a)^{\mathsf{t}} = a\,(f\,g)^{\mathsf{t}} = a\,g^{\mathsf{t}} f^{\mathsf{t}} = (g\,a)^{\mathsf{t}} f^{\mathsf{t}}$.

4.4 *Consider morphisms from one free monoid to another. Show that:*

(a) *every morphism is the composition of an injective morphism and a projection;*

(b) *every continuous morphism is the composition of an injective morphism and a strictly alphabetic morphism.*

Solution: Let $\alpha\colon A^* \to B^*$.

(a) As a bijection between A and another alphabet A' extends to a bijective morphism, we can assume without loss of generality that A and B are disjoint. Set $C = A \cup B$ and write π_A and π_B for the projections of C onto A and B respectively. We define $\theta\colon A^* \to C^*$ by

$$\forall a \in A \qquad a\theta = a\,(a\alpha) \ .$$

On one hand, θ is injective since $(f\theta)\pi_A = f$ for all f in A^*; on the other, $\alpha = \theta\pi_B$.

(b) Take $D = (A \times B) \cup B$ and define $\varphi\colon D^* \to B^*$ by $(a,b)\varphi = b$ and $b\varphi = b$, and $\psi\colon D^* \to A^*$ by $(a,b)\psi = a$ and $b\psi = 1_{A^*}$ for all a in A and all b in B. The morphism φ is strictly alphabetic. Then, define $\sigma\colon A^* \to D^*$ by

$$\forall a \in A \qquad a\sigma = (a, b_1)\,b_2 \cdots b_k \qquad \text{if} \quad a\alpha = b_1 b_2 \cdots b_k \ .$$

Since α is continuous, σ is too, and above all, for all f in A^*, $(f\sigma)\psi = f$ and hence σ is injective. Also, $\alpha = \sigma\varphi$.

4.6 *Let* f *and* g *be words in* A^* *such that* $f\,g = g\,f$. *Show that there exists a word* h *and integers* m *and* n *such that* $f = h^m$ *and* $g = h^n$.

Solution: If $|f| = |g|$, we have $f = g$; take $h = f = g$, and the property is proved. Otherwise, and up to exchanging f and g, we can assume $|f| < |g|$. We prove the property by induction on $|f| + |g|$. If $|f| + |g| = 0$, there is nothing to prove. Otherwise, $f\,g = g\,f$ implies, by Lemma 4.2, that there exists a word k such that $g = k\,f$ and $g = f\,k$. The induction hypothesis implies that there exists h such that $f = h^m$ and $k = h^r$; we then have $g = h^{m+r}$.

4.8 *Let* f_1, f_2, f_3, f_4 *and* g_1, g_2, g_3, g_4 *be two quadruples of words in* A^*. *Show that the three equalities:*

$$f_1 f_4 = g_1 g_4 \qquad (*) \ , \qquad f_1 f_2 f_4 = g_1 g_2 g_4 \qquad (**) \ , \qquad f_1 f_3 f_4 = g_1 g_3 g_4 \qquad (***) \ ,$$

imply the fourth equality: $\qquad f_1 f_2 f_3 f_4 = g_1 g_2 g_3 g_4$.

Solution: [29] Up to exchanging the f's and g's, we can assume that $|f_1| \geqslant |g_1|$ and $(*)$ implies that there exists a word w in B^* such that

$$f_1 = g_1 w \qquad \text{and} \qquad w\, f_4 = g_4 \ .$$

By substituting these two equalities into $(**)$ then into $(***)$, we obtain

$$g_1 w f_2 f_4 = g_1 g_2 w f_4 \ , \qquad \text{thus} \quad w f_2 = g_2 w \qquad \text{and likewise} \qquad w f_3 = g_3 w \ .$$

We then compute, step by step,

$$f_1 f_2 f_3 f_4 = g_1 \underbrace{w f_2} f_3 f_4 = g_1 g_2 \underbrace{w f_3} f_4 = g_1 g_2 g_3 \underbrace{w f_4} = g_1 g_2 g_3 g_4 \ .$$

$\boxed{\textbf{4.10}}$ ***A characterisation*[30] *of free monoids.*** (a) *Let W be a submonoid of a free monoid A^*. Show that $B = (W \setminus 1_{A^*}) \setminus (W \setminus 1_{A^*})^2$ is the minimum generating set of W.*

(b) *Show that if a monoid L is such that for all monoids M and S and for every morphism $\lambda\colon L \to M$ and every* surjective *morphism $\sigma\colon S \to M$ there exists a morphism $\theta\colon L \to S$ such that $\lambda = \theta \circ \sigma$, then L is a free monoid (that is, Lemma 4.1 is* characteristic *of free monoids).*

Solution: (a) Since $B \subseteq W$ and W is a submonoid, $\langle B \rangle \subseteq W$. To show the reverse inclusion, suppose $W \setminus \langle B \rangle$ is non-empty and let f be a word from A^* in $W \setminus \langle B \rangle$ of minimal length. By the definition of B, f is the product of two elements of $W \setminus 1_{A^*}$: $f = g_1\, g_2$, and by the hypothesis of f's minimality, g_1 and g_2 are in $\langle B \rangle$, hence so is f, which is a contradiction.

Let G be a generating set of W and suppose that $B \setminus G$ is not empty: let h be one of its elements of minimum length. Since h is in B, it belongs to $W = \langle G \rangle$, hence $h = g_1\, g_2 \cdots g_n$ is a product of elements of G. If this product is non-trivial, that is, if there exist at least two words g_i different from 1_{A^*}, then h is not in B; if it is, h is in G; in either case, we have a contradiction.

(b) Let $\mathrm{V}(L)$ be the free monoid generated by L and $\alpha\colon \mathrm{V}(L) \to L$ the canonical morphism:

$$(l_1, l_2, \ldots, l_n)\alpha = l_1 \cdot l_2 \cdot \cdots \cdot l_n \ ,$$

where $\cdot$ denotes multiplication in L. By considering the identity morphism $\iota\colon L \to L$, the hypothesis implies that there exists a morphism $\theta\colon L \to \mathrm{V}(L)$ such that $\iota = \theta\alpha$. The morphism θ is injective since ι is.

$$\begin{tikzcd} & L \arrow[dl, "\theta"'] \arrow[d, "\lambda"] \\ S \arrow[r, "\sigma"'] & M \end{tikzcd} \qquad \begin{tikzcd} & L \arrow[dl, "\theta"'] \arrow[d, "\iota"] \\ \mathrm{V}(L) \arrow[r, "\alpha"'] & L \end{tikzcd} \qquad \begin{tikzcd} & L \arrow[dl, shift right, "\lambda"'] \arrow[d, dashed, "\theta"] \arrow[dr, "\iota"] & \\ \mathrm{V}(B) \arrow[ur, shift right, "\sigma"'] \arrow[r, "\beta"'] & \langle B \rangle \arrow[r, "\alpha"'] & L \end{tikzcd}$$

Let $W = L\theta$ and B be its minimum generating set: $W = \langle B \rangle$; θ is an isomorphism between L and $\langle B \rangle$. Let $\beta\colon \mathrm{V}(B) \to \langle B \rangle$ be the canonical morphism. By Lemma 4.1 there exists $\lambda\colon \mathrm{V}(B) \to L$ such that $\beta = \lambda\theta$. By assumption on L, there exists $\sigma\colon L \to \mathrm{V}(B)$ such that $\theta = \sigma\beta$. The morphism σ is injective since θ is. Thus, $\beta = \lambda\sigma\beta$. By construction, β is the identity on B, hence so is $\lambda\sigma$. The unique extension of $\lambda\sigma$ as a morphism from $\mathrm{V}(B)$ to itself is the identity, thus surjective, and σ is surjective: L and $\mathrm{V}(B)$ are isomorphic.

$\boxed{\textbf{4.11}}$ *Prove that* a graded and equidivisible monoid is free *(Proposition 4.3).*

Solution: See [144, p. 103] (the original reference is [149]).

---

[29] Taken from [225]; *cf.* the solution of Exer. V.1.3, p. 651.

[30] Adapted from [103, Th. 1].

5. Semirings

5.1 ***Questions.*** (a) *What is the other semiring with two elements?*

(b) *Why is* $\mathbb{N}_*$ *not a semiring?*

(c) *Why can we not make* $\mathcal{Z} = \mathbb{Z} \cup +\infty$ *a semiring, as we did for* $\mathcal{N}$*?*

Solution: (a) $\mathbb{Z}/2\mathbb{Z} = \{0, 1\}$ equipped with usual operations: $0_{\mathbb{Z}} = 0$ and $1_{\mathbb{Z}} = 1$.

(b) Because it is not a monoid for addition, but only a semigroup; that is, it has no zero.

(c) Because it is impossible to give a value to $x = +\infty + [(-1)+\infty]$. Distributivity implies $x = [1 + (-1)] \cdot +\infty = 0$ but this would contradicts associativity as $x = +\infty + +\infty = +\infty$.

5.3 ***Centre.*** *Show that* $\mathbb{K}_{\mathsf{c}}$*, the centre of* $\mathbb{K}$*, is a sub-semiring of* $\mathbb{K}$*.*

Solution: Clearly, $1_{\mathbb{K}}$ and $0_{\mathbb{K}}$ are in $\mathbb{K}_{\mathsf{c}}$. For all c and d in $\mathbb{K}_{\mathsf{c}}$ and all k in $\mathbb{K}$

$$k\,(c+d) = k\,c + k\,d = c\,k + d\,k = (c+d)\,k \qquad \text{and} \qquad k\,(c\,d) = c\,k\,d = (c\,d)\,k \ .$$

6. Matrices

6.10 ***Faithful representations of the free monoid.***

(a) *Let* $A = \{a, b\}$*. Show that the morphism* $\alpha\colon A^* \to \mathbb{N}^{2\times 2}$ *defined by*

$$a\alpha = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \textit{and} \qquad b\alpha = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \ ,$$

is injective *(that is,* α *is a* faithful representation *of* A^* *by* (2×2) *matrices over* $\mathbb{N}$*).*

(b) *Prove the same property for the morphism*[31] $\beta\colon A^* \to \mathbb{N}^{2\times 2}$ *defined by*

$$a\beta = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \qquad \textit{and} \qquad b\beta = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \ .$$

(c) *Generalise to a morphism* $\zeta\colon Z^* \to \mathbb{N}^{2\times 2}$ *in the case where* Z *contains* k *letters.*

Solution: (a) Let f be in A^* and set[32]

$$f\alpha = \begin{pmatrix} m_1 & m_2 \\ m_3 & m_4 \end{pmatrix} \ .$$

There then follows

$$(f\,a)\alpha = \begin{pmatrix} m_1 & m_1 + m_2 \\ m_3 & m_3 + m_4 \end{pmatrix} \qquad \text{and} \qquad (f\,b)\alpha = \begin{pmatrix} m_1 + m_2 & m_2 \\ m_3 + m_4 & m_4 \end{pmatrix} \ .$$

We thereby observe that for all f in A^* the entries of $f\alpha$ are non-negative integers and that

$$\begin{array}{llllll} \text{if} & f = f'a & \text{then} & m_1 \leqslant m_2 & \text{and} & m_3 \leqslant m_4 \ , \\ \text{if} & f = f'b & \text{then} & m_1 \geqslant m_2 & \text{and} & m_3 \geqslant m_4 \end{array}$$

and at least one of the inequalities is strict.[33] Set

$$u = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \qquad \text{and} \qquad v = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \ .$$

---

[31] *Cf.* [80, Exer. V.2.2]

[32] An elegant solution taken from [182].

[33] *Note added in translation:* This implies that a matrix m in $(A^*)\alpha$ is either the image of words all ending with an a or all ending with a b.

There follows

$$a\alpha\, u = u(a\alpha) = b\alpha\, v = v(b\alpha) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad .$$

Now suppose that there exist f and g in A^* such that $f\alpha = g\alpha$ and choose them such that the sum of their lengths is minimal. We have $f = f'a$ and $g = g'a$ or $f = f'b$ and $g = g'b$. In the first case, $f'\alpha = f\alpha\, u = g\alpha\, u = g'\alpha$, in the second, $f'\alpha = f\alpha\, v = g\alpha\, v = g'\alpha$; this contradicts the minimality of the choice of f and g.

(b) Write $\overline{f}$ for the *integer* represented in base 2 by *the word* f in A^*, by interpreting a by the digit 0 and b by the digit 1. We then verify, by induction on the length of f, that *cf. Exam. I.1.2, p. 52*

$$f\beta = \begin{pmatrix} 1 & \overline{f} \\ 0 & 2^{|f|} \end{pmatrix} \quad ,$$

hence injectivity.

(c) If Z has k elements, we identify Z with $\mathbf{k} = \{0, 1, \ldots, k-1\}$. We define ζ by

$$(i)\zeta = \begin{pmatrix} 1 & i \\ 0 & k \end{pmatrix} \quad , \qquad \text{for } 0 \leqslant i \leqslant k-1 \, ,$$

and we verify, again by induction on the length of f, that

$$f\zeta = \begin{pmatrix} 1 & \overline{f} \\ 0 & k^{|f|} \end{pmatrix} \quad ,$$

if we write $\overline{f}$ the integer represented, in base k, by f in $\mathbf{k}^*$, hence injectivity.

6.11 *Let M be a monoid. Let $\sigma\colon M \to (\mathfrak{P}(M) \times \mathfrak{P}(M))^{2\times 2}$ be the map*[34] *defined by*

$$\forall m \in M \qquad (m)\sigma = \begin{pmatrix} (m,1) & \{(u,v) \mid u\,v = m\} \\ 0 & (1,m) \end{pmatrix} \quad .$$

Verify that σ is a morphism when M is equidivisible.

Solution: This comes down to verifying that, for all k, l, m and n in M,

$$k\,l = m\,n \quad \Longrightarrow \quad k\sigma\, l\sigma = m\sigma\, n\sigma \quad ,$$

or, since the other entries pose no problem, that

$$(k,1)\left[\sum_{r\,s=l}(r,s)\right] + \left[\sum_{p\,q=k}(p,q)\right](1,l) = (m,1)\left[\sum_{y\,z=n}(y,z)\right] + \left[\sum_{u\,v=m}(u,v)\right](1,n) \quad ;$$

that is,

$$\sum_{r\,s=l}(k\,r,s) + \sum_{p\,q=k}(p,q\,l) = \sum_{y\,z=n}(m\,y,z) + \sum_{u\,v=m}(u,v\,n) \quad . \qquad \text{(SE.6.1)}$$

Since M is equidivisible, and up to swapping (k,l) and (m,n), there exists v such that $k\,v = m$ and $v\,n = l$. Consider a pair $(k\,r,s)$ from the left hand side of (SE.6.1); we have $v\,n = r\,s$ and there are two possible cases. Either there exists w such that $v\,w = r$ and $w\,s = n$; then $(k\,r,s) = (m\,w,s)$ is a pair from the right hand side. Or, there exists w such that $r\,w = v$ and $w\,n = s$; then $(k\,r,s) = (k\,r,w\,n)$ is a pair from the right hand side since $k\,r\,w = m$.

[34] *Note added in translation:* This map, which seems to come out of the blue, is indeed the construction that underlies the *Schützenberger product* and which we shall touch on twice at Chapter II (*cf.* p. 356).

NOTES & REFERENCES

Since this chapter is not meant to be read, it does not really have a bibliography; we restrict ourselves to a few general indications.

There are countless works on the elements of algebra. Two great classics are:

S. Lang, *Algebra*. Addison-Wesley, 1965. (Revised third edition: Springer, 2002). [145]

R. Godement, *Algèbre*. Hermann, 1970. [105]

On the combinatorics of words, a less hackneyed subject, one can consult

M. Lothaire, *Combinatorics on Words*. Addison-Wesley, 1983. (Reprint: Cambridge University Press, 1997). [158]

or the corresponding chapter of the *Handbook of Formal Languages*:

C. Choffrut and J. Karhumäki, Combinatorics of Words. In *Handbook of Formal Languages*, vol. 1. Elsevier, 1997. [59]

There are also many introductions to graph theory. Two classics:

R. Diestel, *Graph Theory*. Springer, 1997. [76]

M. Sakarovitch,[35] *Optimisation combinatoire*, vol. 1 Graphes et programmation linéaire. Hermann, 1984. [215]

Notions relating to complexity – in the naive form in which we have tackled it – are explored in all works on algorithms or on the mathematics of computer science. True complexity theory, which deals with the classification of problems and which starts with the distinction between the classes P and NP, still unresolved, has not even been mentioned. Two examples:

A. V. Aho, J. E. Hopcroft and J. D. Ullman, *The Design and Analysis of Computer Algorithms*. Addison Wesley, 1974. [4]

D. Beauquier, J. Berstel, and Ph. Chrétienne, *Éléments d'algorithmique*. Masson, 1992. [23]

On the idea of decidability, two samples taken off the shelf are:

M. Davis, *The Undecidable*. Raven Press, 1965. [71]

P. Dehornoy, *Complexité et décidabilité*. Springer, 1993. [72]

[35] Si ce n'est moi, c'est donc mon frère.

Part I

The three stages of rationality

This page intentionally left blank

Chapter I

The simplest possible machine...

Automata on a free monoid

In the beginning, we define *automata* as *labelled graphs*. This point of view enables a simple presentation of the basic properties of the *languages recognised* by *finite* automata – called *recognisable languages* – and leads naturally to the successive generalisations which will be the subject of subsequent chapters.

We will then consider the family of recognisable languages as the result of a direct construction on the *algebra of languages*, that is, the set of subsets of a free monoid equipped with three operations, called *rational* operations. This is the substance of Kleene's Theorem, and the source of some of the properties of this family of languages which from now on we shall also call *rational*.

Next, we will return to automata but with a functional point of view which is more suitable for the modelling of a calculating machine. This leads directly to the notion of a *deterministic automaton* and on to that of a *minimal automaton*.

The fourth section is an introduction to the theory of rational expressions, which provides an axiomatic perspective on the construction of rational languages. This theory is only sketched, but its definitions allow us to compare different modes of calculation and to introduce the idea of the *derivation* of an expression, which is another means of linking rational languages and automata.

By then we will also have presented, particularly for readers new to the subject, a 'naive' theory of finite automata, which is yet sufficiently rich to include, in their elementary form, the central results of the theory: Kleene's Theorem; the decidability of the equivalence of finite automata and of expressions; the existence, uniqueness and computability of a minimal automaton; iteration lemmas in various forms, including the one due to Ehrenfeucht, Parikh and Rozenberg which gives a combinatorial sufficient condition for a language to be recognisable. This chapter alternates between automata and rational languages, driven by the need to prove the principal results as simply and as early as possible. But this does not prevent us from giving several proofs of the same result, particularly for the two directions of Kleene's Theorem: each illuminates a particular aspect of the property, corresponds to a different algorithmic realisation, and will lead to different generalisations.

In the first supplementary section, we return to the problem of the transformation of a rational expression into an automaton, which is a central problem for practical applications of finite automata. In the second, we tackle the problem of *star height*, which can be seen as a refinement of Kleene's Theorem, a difficult subject in the study

of rational languages, which is still problematic.[1] The subsequent section allows us to review some different models of machines, equivalent to finite automata, that have been proposed at various times and in different contexts. Thence, we will (re)make the link to our intuitive view of a physical device with a finite memory that reacts to reading a sequence of symbols. In the last section the reader will find, in the form of exercises, some properties of languages recognised by finite automata, properties which for the most part will be better understood or explained in the next two chapters.

Contents of the chapter

1. What is an 'automaton'?
2. Rational languages
3. The functional point of view
4. Rational expressions

Deeper & broader

5. From expressions to automata
6. Star height
7. A field of automata
8. A crop of properties

Solutions to the exercises

Notes & references

[1] *Note added in translation:* The recent work of D. Kirsten has brought major improvement and clarification to the solution of this problem; see below.

1 What is an 'automaton'?

An automaton is an automaton is an automaton...

Gertrude Stein

An automaton is a *structure*, but that tells us nothing – because everything is structure, in other words a set equipped with operations – except that we should remember this when defining a map from one automaton to another. Furthermore, an object can be viewed as a structure in several ways. Since they describe the *same* object, they are equivalent, so can be translated into each other. Each illuminates certain aspects and leaves others obscure. To begin, we have to pick one!

1.1 First definitions – first examples

An *automaton* is a *directed graph* which is *labelled* with letters of the alphabet, and in which two subsets of vertices are distinguished. The rest of the subsection elaborates this basic definition.

1.1.1 States, transitions, *etc.*

Definition 1.1 An *automaton* $\mathcal{A}$ is specified by giving the following elements:

- a non-empty set Q, called the set of *states* of $\mathcal{A}$;
- a set A, also non-empty, called the *(input) alphabet* of $\mathcal{A}$;
- two subsets I and T of Q; I is the set of *initial states* and T is the set of *final* or *terminal states* of $\mathcal{A}$;
- a subset E of $Q \times A \times Q$, called the set of *transitions* of $\mathcal{A}$.

We write, therefore, fairly naturally, $\mathcal{A} = \langle\, Q, A, E, I, T \,\rangle$ and we say that $\mathcal{A}$ is an *automaton over (the alphabet) A*. □

An automaton $\mathcal{A}$ is a labelled graph; if we ignore the transition labels and the initial and final states of $\mathcal{A}$, we obtain the graph *underlying* $\mathcal{A}$, which is, in general, a multigraph.

An automaton which we will use no fewer than 40 times in what follows.

Example 1.1 The automaton $\mathcal{A}_1$. A first automaton: the input alphabet is $A = \{a, b\}$, as will be the case in most of our examples, the set of states Q_1 is a set of 3 elements $\{p, q, r\}$; p is the unique initial state, r is the unique final state, and the set of transitions E_1 is described by

$$E_1 = \{(p, a, p), (p, b, p), (p, a, q), (q, b, r), (r, a, r), (r, b, r)\} \ .$$

We write therefore: $\mathcal{A}_1 = \langle\, \{p, q, r\}, \{a, b\}, E_1, \{p\}, \{r\} \,\rangle$. □

This definition does not show us much! As graphs, automata have a natural graphical representation. With the conventions of Figure 1.1 below, they take the form already given in the prologue: the states are represented by circles (inside which the name of the state may be written), the transitions by arrows, with their labels written next to them, the initial states are marked with short arrows arriving at them, and the final states by short arrows leaving them.[2] The graph underlying $\mathcal{A}_1$ is the multigraph of Figure 0.7.1.

cf. p. 33

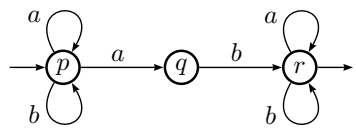


Figure 1.1: Graph representation of the automaton $\mathcal{A}_1$

Conversely, a figure such as Figure 1.1 can be easily transcribed as an automaton in the form of Definition 1.1.

Example 1.2 The 'divider by 3'. Figure P.2 (a) from the prologue, reproduced here as Figure 1.2(a), represents the automaton:

$$\mathcal{P}_2 = \langle\, \{\mathbf{0}, \mathbf{1}, \mathbf{2}\}, \{0, 1\}, F_2, \{\mathbf{0}\}, \{\mathbf{0}\} \,\rangle \ ,$$

where $\quad F_2 = \{(\mathbf{0}, 0, \mathbf{0}), (\mathbf{0}, 1, \mathbf{1}), (\mathbf{1}, 0, \mathbf{2}), (\mathbf{1}, 1, \mathbf{0}), (\mathbf{2}, 0, \mathbf{1}), (\mathbf{2}, 1, \mathbf{2})\}.$

Note the typographical trick used to distinguish between the states and the letters of the input alphabet of $\mathcal{P}_2$. To conform to Definition 1.1, we had to choose a final state which does not appear in Figure P.2 . To make reading easier, we will use the alphabet $A = \{a, b\}$ instead of $\{0, 1\}$ from now on; a word f from A^* can be considered as a number written in binary, written $\overline{f}$, interpreting a as the digit 0 and b by the digit 1. (For example: $\overline{ba} = 2$, $\overline{aaa} = 0$, $\overline{abaaba} = 18$.) □

Example 1.3 Another automaton, $\mathcal{Z}_1 = \langle\, \mathbb{Z}, \{a, b\}, X_1, \{0\}, \{0\} \,\rangle$, with

$$X_1 = \{(z, a, z+1), (z, b, z-1) \mid z \in \mathbb{Z}\} \ ,$$

is shown in Figure 1.3. □

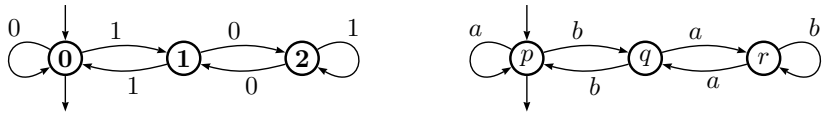
In the following, in accord with general usage, we will identify a singleton set with the element it contains, and omit the braces.

---

[2] It used to be conventional to represent final states by a double circle (*cf.* [121, 27, 183] to cite only a few spread over the last thirty years). The rival notation of short leaving arrows appears in many recent works (*cf.* [80, 123]). I have adopted the new notation for at least two reasons: it better expresses the duality between initial and final state, and it gives a natural way of associating a label with a final state, which we will wish to do several times in the chapters to come.

For this last reason moreover, a state which is both initial and final is not marked with a double-headed arrow (↔) as in [80, 123] for example, but with two distinct arrows, one entering and one leaving (see figures 1.2 and 1.3).

In cases where one is interested in the recognition of *infinite words* – and hence not in this work – the initial and final states are no longer symmetric, and the double circle notation is therefore preferable (see [22] for example).



(a) Prologue version (b) Chapter I version

Figure 1.2: $\mathcal{P}_2$, the 'divider by 3'

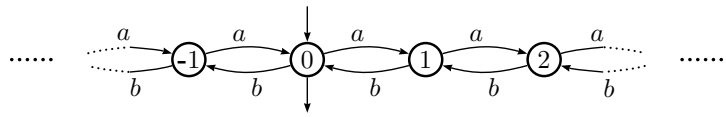


Figure 1.3: The automaton $\mathcal{Z}_1$

Remark 1.1 We could consider the input alphabet as not being part of the automaton – as suggested by the phrase 'an automaton over the alphabet A' – and write $\mathcal{A} = \langle Q, E, I, T \rangle$. We could also write $\mathcal{A} = \langle I, E, T \rangle$ and recover A and Q from projections of the elements of E (*cf.* below). In Section 3, we will introduce another way of naming and describing the *same* five elements.

The notation may therefore vary from one place to another, but *in every case* an automaton is defined by the five sets described above. □

An automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ *is finite* if the sets Q, A and E that appear in its definition are finite. It is enough to assume that E is finite, since, as mentioned above, we can recover Q and A by projection, or at least their useful parts. Conversely, if the alphabet A is considered as given and is finite, the automaton $\mathcal{A}$ is finite if and only if Q is finite, and it is this that we shall take as our definition, at least for the present chapter.

Definition 1.2 Let A be a finite alphabet. An automaton over A is *finite* if and only if its set of states is finite. □

In the preceding examples, all over the alphabet $\{a, b\}$, the automata $\mathcal{A}_1$ and $\mathcal{P}_2$ are finite and $\mathcal{Z}_1$ is infinite.

Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton; then a subset R of Q defines an automaton $\mathcal{B} = \langle R, A, F, J, U \rangle$, where:

$$F = E \cap (R \times A \times R) \ , \qquad J = I \cap R \qquad \text{and} \qquad U = T \cap R \ .$$

The automaton $\mathcal{B}$ is a *sub-automaton* of $\mathcal{A}$. Similarly, a subset H of E defines an automaton $\mathcal{C} = \langle Q, A, H, I, T \rangle$ which we will also call a sub-automaton[3] of $\mathcal{A}$.

Another way to describe and notate an automaton comes from the notion of the *incidence matrix* of a graph, which looks ahead to the point of view which we will *see p. 402*

[3]Graph theorists are more precise and speak of a *subgraph* in the first case, of a *partial graph* in the second, and of a *partial subgraph* when they are combined. We will not need to make this distinction.

adopt in Chapter III, and is only mentioned here in preparation (and to clarify the definition of a transposed automaton).

cf. § *1.1.3, p. 58*

Definition 1.3 Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be a finite automaton. The *transition matrix* of $\mathcal{A}$ is the square matrix of dimension Q, which we write E, as it is just another way to describe the set of transitions. The entry $E_{p,q}$ is the set of labels of transitions that go from p to q in $\mathcal{A}$. It is therefore reasonable to write $\mathcal{A}$ as a triple:

$$\mathcal{A} = \langle I, E, T \rangle \ ,$$

where E is the transition matrix and where the two subsets I and T are identified with two boolean vectors of dimension Q. For reasons which will become clear in Chapter III, I is a *row vector* and T a *column vector*. □

Example 1.1 (continued) We can therefore write:

$$\mathcal{A}_1 = \left\langle \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} \{a,b\} & a & 0 \\ 0 & 0 & b \\ 0 & 0 & \{a,b\} \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\rangle \ .$$ □

For the rest of this section, $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is an automaton over A.

1.1.2 Computations, recognised language *etc.*

If $e = (p, a, q)$ is a transition of $\mathcal{A}$, that is, if e is in E, we say that a is the *label* of e and we will write $p \xrightarrow{a} q$, or $p \xrightarrow[\mathcal{A}]{a} q$ where it might be ambiguous which automaton we are considering. We also say that p is the *source* and q the *destination* of the transition e.

Definition 1.4 A *computation* c in $\mathcal{A}$ is a sequence of transitions where the source of each is the destination of the previous one, which can be written thus:[4]

$$c \ := \ p_0 \xrightarrow{a_1} p_1 \xrightarrow{a_2} p_2 \cdots \xrightarrow{a_n} p_n \ .$$

The state p_0 is the *source* of the computation c, and p_n its *destination*. The *length* of the computation c is n, the number of transitions which make up c. The *label* of c is the concatenation (product) of the labels of the transitions of c; in the above case, the label of c is $a_1 a_2 \cdots a_n$ and is written thus:

$$c \ := \ p_0 \xrightarrow{a_1 a_2 \cdots a_n} p_n \qquad \text{or} \qquad p_0 \xrightarrow[\mathcal{A}]{a_1 a_2 \cdots a_n} p_n \ .$$

A computation in $\mathcal{A}$ is *successful* if its source is an initial state and its destination is a final state. A word in A^* is called *accepted* or *recognised* by $\mathcal{A}$ if it is the label of a *successful computation* in $\mathcal{A}$. □

[4] We write $c := p_0 \xrightarrow{a_1} p_1 \cdots$ rather than $c = p_0 \xrightarrow{a_1} p_1 \cdots$ to avoid any ambiguity in the case where the description of a computation begins with an equality between two states (*cf.* Section 1.4).

Thus, in translating from 'graph language' to 'automaton language', *state* replaces *vertex*, *transition* replaces *edge*, and *computation* replaces *path*.

Note that a computation has a (single) label, by definition, but that the same word can be the label of several distinct computations, and even of several distinct successful computations.

Example 1.1 (continued) The word $abab$ is the label of the four following computations in $\mathcal{A}_1$:

$$c_1 := p \xrightarrow{a} p \xrightarrow{b} p \xrightarrow{a} p \xrightarrow{b} p \ , \qquad c_2 := r \xrightarrow{a} r \xrightarrow{b} r \xrightarrow{a} r \xrightarrow{b} r \ ,$$

$$c_3 := p \xrightarrow{a} p \xrightarrow{b} p \xrightarrow{a} q \xrightarrow{b} r \qquad \text{and} \qquad c_4 := p \xrightarrow{a} q \xrightarrow{b} r \xrightarrow{a} r \xrightarrow{b} r \ .$$

The last two are successful computations. □

The computations of length 1 in $\mathcal{A}$ are exactly the transitions of $\mathcal{A}$. By convention, we consider each state to be both the source and destination of a (unique) computation of length 0 whose label is the empty word 1_{A^*}, the unique word in A^* of length 0. Following from this convention, we have:

Property 1.1 *The empty word 1_{A^*} is accepted by an automaton $\mathcal{A}$ if and only if there is a state of $\mathcal{A}$ which is both initial and final (that is, if $I \cap T \neq \emptyset$).* ■

Definition 1.5 (i) The language *accepted*, or *recognised* by $\mathcal{A}$, also called the *behaviour of* $\mathcal{A}$, written $L(\mathcal{A})$ or $|\mathcal{A}|$, is the set of words accepted (or recognised) by $\mathcal{A}$:

$$L(\mathcal{A}) = \{f \in A^* \mid \exists p \in I \, , \ \exists q \in T \quad p \xrightarrow[\mathcal{A}]{f} q\} \ .$$

(ii) Two automata are *equivalent* if they recognise the same language. □

We have now introduced three apparently synonymous terms and two notations for the language 'associated' with an automaton. We did this to familiarise ourselves with the terminology that we will need later when we generalise automata to structures other than free monoids. For the rest of *this* chapter, we will continue to use the most standard terminology:[5] $L(\mathcal{A})$ is the language *recognised* by $\mathcal{A}$.

Example 1.1 (continued) A successful computation in $\mathcal{A}_1$ must contain the sequence of transitions $(p, a, q)(q, b, r)$, and every word recognised by $\mathcal{A}_1$ contains a factor ab. Conversely, a word that contains a factor ab can be written $f = f' \, a \, b \, f''$, and is the label of a successful computation:

$$p \xrightarrow{f'} p \xrightarrow{a} q \xrightarrow{b} r \xrightarrow{f''} r \ .$$

Thus, $L(\mathcal{A}_1)$ is the set of words in $\{a, b\}^*$ that contain a factor ab . □

[5] Even if it is in fact infelicitous. See Rem. 1.2.

Example 1.2 (continued) If everything we discussed in the prologue makes sense, $L(\mathcal{P}_2)$ is the set of binary representations of numbers divisible by 3. Notice that we obtain all the representations, such as $aaabba$ (that is, 000110), and not just the normalised representations, without leading a's (that is, without leading zeros). □

cf. Exer. 2.8, p. 89

Example 1.3 (continued) We verify that the language recognised by the automaton $\mathcal{Z}_1$ is the set of words in $\{a,b\}^*$ that contain as many a's as b's: that is, the language that we wrote[6] Z_1: $Z_1 = L(\mathcal{Z}_1) = \{f \in \{a,b\}^* \mid |f|_a = |f|_b\}$. □

cf. Exam. 0.3.6, p. 22

Let $\mathcal{B} = \langle R, A, F, J, U \rangle$ be a sub-automaton of $\mathcal{A} = \langle Q, A, E, I, T \rangle$; every computation in $\mathcal{B}$ is a computation in $\mathcal{A}$ since $F \subseteq E$, and every successful computation in $\mathcal{B}$ is a successful computation of $\mathcal{A}$ since $J \subseteq I$ and $U \subseteq T$, from which we conclude:

Property 1.2 *If $\mathcal{B}$ is a sub-automaton of $\mathcal{A}$, $L(\mathcal{B})$ is a subset of $L(\mathcal{A})$.* ■

Definition 1.6 Let A be a finite alphabet. A language L of A^* is called *recognisable* if it is recognised by a *finite* automaton over A: that is, if there exists a *finite automaton* $\mathcal{A}$ over the alphabet A such that $L = L(\mathcal{A})$.

The family of recognisable languages of A^* is written $\operatorname{Rec} A^*$. □

In a way, this definition is the basis of the theory. Finite automata are not merely an object of study in themselves: they also give a procedure for defining or describing in finite terms possibly infinite subsets of A^* – for example, the set of words that contain a factor ab or the set of binary representations of numbers divisible by 3 – and, by this roundabout means, to be able to deduce the properties of the latter from those of the former.

Remark 1.2 The term 'recognisable language' and the notation $\operatorname{Rec} A^*$ for the languages recognised by a finite automaton follow common usage, but are not ideal insofar as they fail to correspond exactly with those we will be using in the next chapter and thereafter. They are so common that changing them is inconceivable, and they introduce no difficulty so long as we assume that the alphabet A is *finite*. We therefore require that A be finite for the rest of this chapter. We will state it explicitly, as in the definition above, when it is necessary to make statements compatible with future definitions.[7] The case of infinite alphabets will be touched on in Chap. II. □

cf. Sec. II.2.2, p. 238

cf. Exer. II.2.12, p. 243

Remark 1.3 If it is languages that matter rather than automata, then the set of states of an automaton is immaterial and can be replaced by any set in bijection with it. In particular, when we show an automaton in graphical form we will often omit to name the states. □

Let us now give some simple supplementary samples of recognisable languages.

---

[6] And that we will continue to write thus, at least until Section II.6.

[7] In particular for readers who do not read this book in order.

Property 1.3 *Every word* $f = a_1 a_2 \cdots a_n$ *in* A^* *of length* n *is recognised by an automaton* $\mathcal{A}_f$ *with* $n+1$ *states.*

Proof. The set of states of $\mathcal{A}_f$ is the set of prefixes of f; the initial state is 1_{A^*}, the final state is f, and for each proper prefix g of f, (g, a, ga) is a transition of $\mathcal{A}_f$ if and only if ga is a prefix of f (*cf.* Figure 1.4). ■

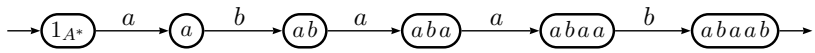
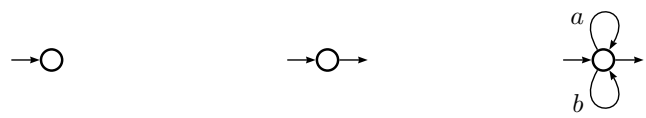


Figure 1.4: $\mathcal{A}_{abaab}$

This construction may equally be used for the automaton which recognises the *empty word* 1_{A^*}. The automata that recognise the empty set and the whole of A^* are just as simple (*cf.* Figure 1.5).

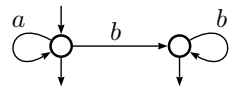


(a) (b) (c)

Figure 1.5: Some automata: (a) for $\emptyset$, (b) for 1_{A^*} and (c) for A^*

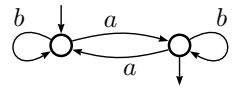
Example 1.4 The language K_1 of words in A^* such that all the a's precede all the b's is recognised by a finite automaton (see adjacent figure):

$$K_1 = \{a^n b^m \mid n, m \in \mathbb{N}\} \ . \qquad \square$$



Example 1.5 The language L_2 of words in A^* containing an odd number of a's is recognised by a finite automaton (see adjacent figure):

$$L_2 = \{f \in A^* \mid |f|_a \equiv 1 \bmod 2\} \ . \qquad \square$$



Remark 1.4 A simple cardinality argument allows us to reassure ourselves that Definition 1.6 is non-trivial: in other words, that there exist languages which are not recognisable. For a given A there exists, for each positive integer n, a finite number of automata over A having n states; we can therefore *enumerate* all the finite automata over A and hence the recognisable languages of A^*: $\mathrm{Rec}\, A^*$ is a countably infinite set. Furthermore, $\mathfrak{P}(A^*)$ has a cardinality strictly greater than that of A^*, which is countably infinite; $\mathfrak{P}(A^*)$ is therefore uncountable and therefore contains non-recognisable languages. *cf. Exer. 1.2*

This argument is of the same type as that which says that, since the set of real numbers is uncountable (by diagonalisation), and since the rational numbers are countable, there must be irrational real numbers. In the same way as that argument does not prove that a particular real number such as $\sqrt{2}$ is irrational, our counting

argument does not tell us which particular sets of words are non-recognisable. Later, we will see (§1.2.3 and §1.3.3) some methods for obtaining that type of result. □

cf. p. 62 and p. 70

1.1.3 Transposition and left–right duality

The *transpose* of $\mathcal{A}$ is the automaton obtained by reversing all the arrows in $\mathcal{A}$ – that is, by exchanging the source and destination of each transition, *and* the initial and final states; it is written $\mathcal{A}^{\mathsf{t}}$:

$$\mathcal{A}^{\mathsf{t}} = \langle\, Q, A, E^{\mathsf{t}}, T, I \,\rangle \quad \text{with} \quad E^{\mathsf{t}} = \{(p,a,q) \in Q \times A \times Q \mid (q,a,p) \in E\}\,.$$

cf. Def. 1.3, p. 54

The terminology and notation are justified by reference to matrix notation. In fact, if we write:

$$\mathcal{A} = \langle\, I, E, T \,\rangle \;, \qquad \text{we obtain} \qquad \mathcal{A}^{\mathsf{t}} = \langle\, T^{\mathsf{t}}, E^{\mathsf{t}}, I^{\mathsf{t}} \,\rangle \;,$$

cf. Exer. 0.6.6, p. 32

where the superscript ^{t} denotes *matrix transposition.*

cf. Sec. 0.4, p. 24

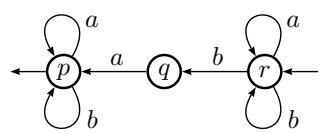
Recall that if $f = a_1 a_2 \cdots a_n$ is a word in A^*, the *mirror image* of f (or *transpose* of f) is the word $f^{\mathsf{t}} = a_n a_{n-1} \cdots a_1$, and that the mirror image operation extends additively to subsets of A^*:

$$\forall L \subseteq A^* \qquad L^{\mathsf{t}} = \{f^{\mathsf{t}} \mid f \in L\}\,.$$

The label of a computation in $\mathcal{A}^{\mathsf{t}}$ is obviously the mirror image of the corresponding computation in $\mathcal{A}$; a successful computation in $\mathcal{A}^{\mathsf{t}}$ corresponds to a successful computation in $\mathcal{A}$; and therefore:

$$L(\mathcal{A}^{\mathsf{t}}) = [L(\mathcal{A})]^{\mathsf{t}}\,.$$

Example 1.1 (continued) The transpose of $\mathcal{A}_1$, shown in the adjacent figure, is



$$\mathcal{A}_1^{\mathsf{t}} = \langle\, \{p,q,r\}, \{a,b\}, E_1^{\mathsf{t}}, \{r\}, \{p\} \,\rangle \qquad \text{with}$$

$$E_1^{\mathsf{t}} = \{(p,a,p), (p,b,p), (q,a,p), (r,b,q), (r,a,r), (r,b,r)\}\,.$$

The language $L(\mathcal{A}_1^{\mathsf{t}})$ is the set of words in A^* that contain a factor ba. □

Since the transpose of a finite automaton is clearly finite we have, anticipating the results of the next subsection:

Proposition 1.1 *The mirror image of a recognisable language is another recognisable language:* Rec A^* *is closed under the mirror image operation.* ■

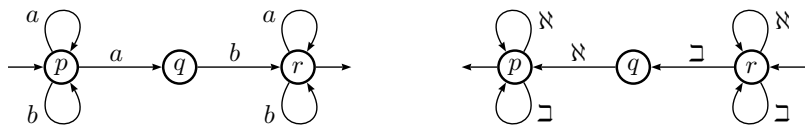
Remark 1.5 There is more to this elementary construction than its rather trivial ability to reverse words. Bearing in mind that automata are models for *computing machines*, transposition corresponds to a change in the machine's mode of operation, from *reading left-to-right* to *reading right-to-left.*

cf. Sec. 7, p. 171

Here are two words:

$$abaab \qquad \text{and} \qquad \text{בראבא}$$

and here are two automata:



Nothing in these representations shows that the words should be read, that the labels of the automata's transitions should follow one another, forwards or backwards. (Nothing, except for the information given implicitly by the alphabetic symbols themselves.) We have encountered this difficulty already in the prologue, where we had to specify in which direction numbers were to be read: this is a case where the orientation is not fixed since, for example, addition works from right to left but division from left to right.

When we say that 'the label of c is the *concatenation* of the labels of the transitions of c' with no further explanation, we have implicitly assumed that concatenation is performed on the right, whereas we could just as well have decided to do it on the left. In the first case we will say, by analogy with machines that read from or write to a tape, that the automaton 'reads' or 'writes' the word *from left to right*, and in the second case that it reads (or writes) *from right to left*.

cf. Def. 1.4, p. 54

cf. Sec. 7.1, p. 171

To be explicit, we may therefore speak of *left automata* in the first case and of *right automata* in the second. Since an automaton $\mathcal{A}$ is just a labelled graph, it can be seen as a left or right automaton, so we would need some notation, such as $\overset{\rightsquigarrow}{\mathcal{A}}$ and $\overset{\leftsquigarrow}{\mathcal{A}}$, to indicate which type we mean. We would have therefore:

cf. also § V.1.2.2 and Rem. V.1.6, p. 654

$$L(\overset{\leftsquigarrow}{\mathcal{A}}) = \left[L(\overset{\rightsquigarrow}{\mathcal{A}})\right]^{\mathsf{t}} \qquad \text{and} \qquad L(\overset{\rightsquigarrow}{\mathcal{A}}) = L(\overset{\leftsquigarrow}{\mathcal{A}^{\mathsf{t}}}) \ .$$

In fact, with the notable exception of *transducers* in Chapter V, we will only consider left automata from now on, which we will simply call automata, and we will not need the notations above. But we will remember that transposition expresses the duality between the two directions of reading, between source and destination, between initial and final states, and, for many properties, we will say that $\mathcal{A}$ is 'co-P' to mean that $\mathcal{A}^{\mathsf{t}}$ has the property P. □

Exercises

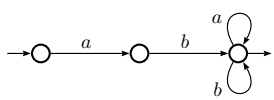
- 1.1 Give the matrix form of the automaton $\mathcal{P}_2$.

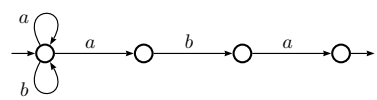
- 1.2 How many automata over $\{a, b\}$ have three states (approximately)?

- 1.3 Justify the description of $L(\mathcal{Z}_1)$ given above (*cf.* Example 1.3, p. 56).

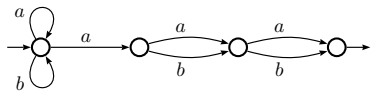
- 1.4 Describe the languages recognised by the following automata (in everyday terms, as we did in Examples 1.1 and 1.3):

cf. p. 56

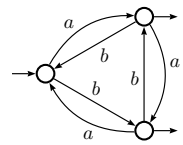
(a) 

(b) 

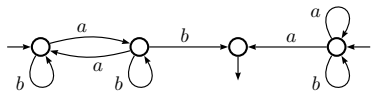
(c)



(d)



(e)



• 1.5 Construct finite automata that accept that following languages ($A = \{a, b\}$):

(a) $L_3 = \{f \in A^* \mid |f|_a \geqslant 3\}$; (b) $L_4 = \{f \in A^* \mid |f|_a \not\equiv |f|_b \bmod 2\}$;

(c) L_5 = the set of words in A^* such that each b is followed by an a.

• 1.6 Imitating the construction given in the prologue, give an automaton that computes the remainder after division by 5 of numbers written in binary,

(a) reading the numbers from right to left; (b) reading the numbers from left to right.

1.7 Do the same for the division by 5 of numbers written in ternary.

1.8 Do the same for the division by 4 of numbers written in binary.

1.2 Basic constructions, basic properties

From the very definition of automata we can conclude directly, or almost directly, the basic properties of recognisable languages. Indeed, from one or more automata we can easily construct other automata. What is the effect of these constructions on the languages recognised? Conversely, how do operations on languages translate into operations on automata?

cf. Def. 1.6, p. 56

These questions echo the remark following the definition of recognisable languages: automata are *one way* of describing these languages, and this type of problem will appear throughout this work. Here, to start with, the questions and the answers are simple.

1.2.1 Union

Let $\mathcal{A}' = \langle Q', A, E', I', T' \rangle$ and $\mathcal{A}'' = \langle Q'', A, E'', I'', T'' \rangle$ be two automata over A. Let us adopt a convention: *when we use this phrasing, we will always assume that the sets of states Q' and Q'' are* disjoint. (This assumption is without loss of generality, since the behaviour of an automaton is invariant under renaming of its states.)

cf. Rem. 1.3, p. 56

The *union* of $\mathcal{A}'$ and $\mathcal{A}''$ is by definition the automaton $\mathcal{B}$ defined by:

$$\mathcal{B} = \mathcal{A}' \cup \mathcal{A}'' = \langle Q' \cup Q'', A, E' \cup E'', I' \cup I'', T' \cup T'' \rangle \ .$$

A transition is in $\mathcal{B}$ if and only if it is a transition of either $\mathcal{A}'$ or of $\mathcal{A}''$; it follows that a (successful) computation of $\mathcal{B}$ must be a (successful) computation in $\mathcal{A}'$ or $\mathcal{A}''$. Thus:

$$L(\mathcal{A}' \cup \mathcal{A}'') = L(\mathcal{A}') \cup L(\mathcal{A}'') \ . \tag{1.1}$$

As $\mathcal{A}' \cup \mathcal{A}''$ is finite if and only if $\mathcal{A}'$ and $\mathcal{A}''$ are, it follows that:

Proposition 1.2 *The union of two recognisable languages is a recognisable language: that is,* $\mathrm{Rec}\, A^*$ *is closed under finite union.* ■

Example 1.1 (continued) The set of words in A^* that contain (at least once) a factor ab or a factor ba is the union of the languages recognised by $\mathcal{A}_1$ and by $\mathcal{A}_1^{\mathsf{t}}$: it is recognised by the automaton in Figure 1.6. □

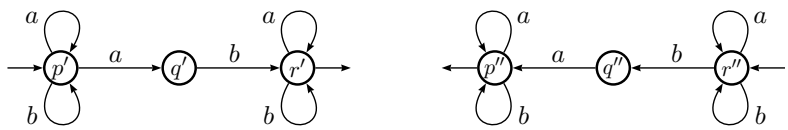


Figure 1.6: The union of $\mathcal{A}_1$ and $\mathcal{A}_1^{\mathsf{t}}$

From Proposition 1.2 and Property 1.3, we immediately obtain:

Corollary 1.3 *Every finite subset of* A^* *is recognisable.* ■

1.2.2 Cartesian product

The *cartesian product* or *direct product* – or, for brevity, the *product* – of two automata corresponds to the *intersection* of the languages recognised.[8] Let $\mathcal{A}' = \langle\, Q', A, E', I', T' \,\rangle$ and $\mathcal{A}'' = \langle\, Q'', A, E'', I'', T'' \,\rangle$ be two automata over A. The *product* of $\mathcal{A}'$ and $\mathcal{A}''$ is the automaton $\mathcal{C}$ defined as

$$\mathcal{C} = \mathcal{A}' \times \mathcal{A}'' = \langle\, Q' \times Q'', A, E, I' \times I'', T' \times T'' \,\rangle \ ,$$

where the set E of transitions is described by

$$E = \{((p', p''), a, (q', q'')) \mid (p', a, q') \in E' \quad \text{and} \quad (p'', a, q'') \in E''\} \ .$$

Thus, each state of $\mathcal{C}$ has two components: the first is a state of $\mathcal{A}'$, and the second a state of $\mathcal{A}''$. A transition is in $\mathcal{C}$ if and only if the projection on the first component of the states is a transition of $\mathcal{A}'$ and the projection on the second component of the states is a transition of $\mathcal{A}''$, and these two projections have the same label.

It follows that a (successful) computation of $\mathcal{C}$ projects both on the first component to a (successful) computation of $\mathcal{A}'$ and on the second component to a (successful) computation of $\mathcal{A}''$. Hence:

$$L(\mathcal{A}' \times \mathcal{A}'') = L(\mathcal{A}') \cap L(\mathcal{A}'') \ . \tag{1.2}$$

As $\mathcal{A}' \times \mathcal{A}''$ is finite if $\mathcal{A}'$ and $\mathcal{A}''$ are, we have:

Proposition 1.4 *The intersection of two recognisable languages is a recognisable language, that is,* $\mathrm{Rec}\, A^*$ *is closed under intersection.* ■

[8] And not the product of the languages, which is formed by another construction that we will see later (*cf.* Sec. 2, p. 82).

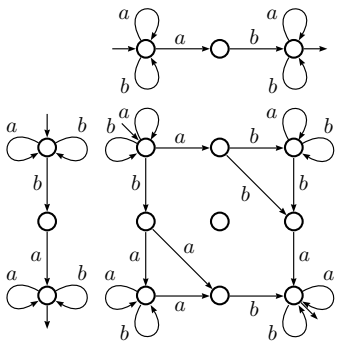


Fig. 1.7: Product of $\mathcal{A}_1$ and $\mathcal{A}_1^t$

Example 1.1 (continued) An automaton over $A = \{a, b\}$ that recognises words containing a factor $a\,b$ *and* a factor $b\,a$ can be constructed by taking the product of $\mathcal{A}_1$ and $\mathcal{A}_1^t$: *cf.* Figure 1.7. □

The construction, and Equation (1.2), hold for all automata, not just for finite ones.

Example 1.3 (continued) The product of $\mathcal{Z}_1$ with the automaton from Example 1.4 gives the automaton $\mathcal{Z}_1'$ shown in Figure 1.8, and recognises the language:

$$P_1 = Z_1 \cap K_1 = \{a^n b^n \mid n \in \mathbb{N}\} \ . \qquad \square$$

Obviously, an automaton $\mathcal{A}$ and its cartesian square $\mathcal{A} \times \mathcal{A}$ are equivalent.

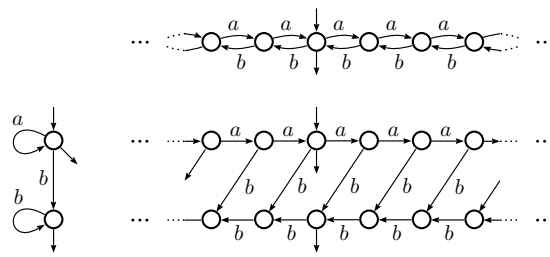


Figure 1.8: An automaton $\mathcal{Z}_1'$ that recognises P_1

Having proved Propositions 1.2 and 1.4, which establish the closure of $\operatorname{Rec} A^*$ under union and intersection, the question naturally arises whether $\operatorname{Rec} A^*$ is also closed under complement: that is, whether $\operatorname{Rec} A^*$ is a Boolean algebra. The answer is affirmative, but its proof requires a rather more elaborate construction, which will be seen in Section 3.

cf. p. 105 et seq.

The reason for this difficulty is easy to see. A word is recognised by an automaton if there exists at least one successful computation whose label it is, hence the complement of a language accepted by an automaton is the set of words that are not the label of any successful computation; the problem to solve is the replacement of an existential quantifier by a universal quantifier.

1.2.3 Quotient (of a language)

We will now define an operation on words that will play a fundamental role in the study of finite automata: the (right or left) *quotient*. Let us begin naively. Just as we can define in $\mathbb{N}$ the *subtraction* of x from y, the inverse operation of addition, with the condition that $y \geqslant x$, we can define in A^* an inverse operation of product between two words f and g with the condition that f is a *prefix* of g, an operation that will not therefore be total: for all f and g in A^*, the *(left) quotient* of g by f,

written $f^{-1}g$, is defined by:

$$f^{-1}g = \begin{cases} h & \text{if} \quad g = f\,h \ , \\ \text{undefined} & \text{if} \quad f \text{ is not a prefix of } g \ . \end{cases}$$

For example, $(aba)^{-1}abaababa = ababa$ and $(abb)^{-1}abaababa$ is undefined. Note that for all f in A^* we have $f^{-1}f = 1_{A^*}$.

Despite appearances, the expression

$$f^{-1}g$$

must not be seen as the product of a new element f^{-1} plucked out of thin air by the word g. It is much more accurate and useful to consider that taking the (left) quotient by f amounts to associating with f a partial function from A^* to itself, and that $f^{-1}g$ is the result of applying this operator[9] f^{-1} to the word g. The (left) quotient extends *additively* to languages:

$$f^{-1}L = \bigcup_{g \in L} \{f^{-1}g\} = \{h \mid f\,h \in L\} \ ,$$

which leads us rather to consider that $f^{-1}g = \emptyset$ when f is not a prefix of g. Symmetrically, we can define the *right quotient*. We will study the properties of the quotient further, in particular in Section 3. *cf.* § *3.3.1, p. 112*

Now let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton and, for each f in A^*, let I_f be the set of states of $\mathcal{A}$ that are the destination of a computation with label f, whose source is in I (we will say that they are *accessible* in $\mathcal{A}$ by f):

$$I_f = \{q \in Q \mid \exists i \in I \quad i \xrightarrow[\mathcal{A}]{f} q\} \ .$$

We then write $f^{-1}\mathcal{A} = \langle Q, A, E, I_f, T \rangle$ the automaton obtained by replacing $\mathcal{A}$'s set of initial states by I_f (*cf.* Figure 1.9). It follows that:

$$L(f^{-1}\mathcal{A}) = f^{-1}[L(\mathcal{A})] \ . \tag{1.3}$$

Indeed, if $f\,g$ is in $L(\mathcal{A})$, there exists a path

$$i \xrightarrow[\mathcal{A}]{f} q \xrightarrow[\mathcal{A}]{g} t$$

where i is in I and t in T; hence q is in I_f and g in $L(f^{-1}\mathcal{A})$. Conversely, if g is in $L(f^{-1}\mathcal{A})$, we have:

$$q \xrightarrow[f^{-1}\mathcal{A}]{g} t$$

where q is in I_f and t in T; therefore there exists i in I such that $i \xrightarrow[\mathcal{A}]{f} q$ and fg is in $L(\mathcal{A})$.

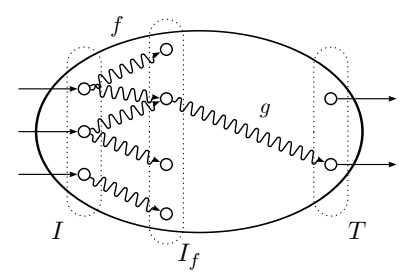


Fig. 1.9: Construction of the automaton $f^{-1}\mathcal{A}$

If $\mathcal{A}$ is finite, $f^{-1}\mathcal{A}$ is too, and we have therefore proved:

[9] Written on the left rather than the right in order to distinguish it easily from the right quotient.

Proposition 1.5 *The quotient of a recognisable language by a word is a recognisable language: that is,* $\mathrm{Rec}\, A^*$ *is closed under quotient by words.* ■

In fact, we have just proved a much stronger property. If $\mathcal{A}$ is finite, there is only a finite number of distinct subsets of Q, so it follows from (1.3) that:

Property 1.4 *If L is a recognisable language of A^*, there is only a finite number of languages of the form $f^{-1}L$, where f ranges over A^*.* ■

cf. Prop. 3.10, p. 114

We will see in Section 3 that this remarkable property is *characteristic* of recognisable languages and that the languages $f^{-1}L$ play a fundamental role in the description of L. But for now, we must be patient. This property already gives us a necessary condition for a language to be recognisable, and therefore a means of proving that certain languages are *not* recognisable: a first answer to the problem raised by Remark 1.4.

cf. p. 57

Example 1.6 The language $P_1 = \{a^n b^n \mid n \in \mathbb{N}\}$ is not recognisable since b^k is in $(a^r)^{-1}P_1$ if and only if $k = r$, and hence P_1 has an infinite number of distinct quotients by the words of $\{a, b\}^*$. □

Again, because there is only a finite number of distinct subsets of Q, Equation (1.3) implies a stronger property than Property 1.4. For every subset K of A^*, the quotient of L by K is defined as:

$$K^{-1}L = \bigcup_{f \in K} f^{-1}L \ .$$

Hence, if L is recognisable by the automaton $\mathcal{A}$:

$$K^{-1}L(\mathcal{A}) = \bigcup_{f \in K} L(f^{-1}\mathcal{A}) \ .$$

Since the union above, which looks infinite, is in fact composed of a finite number of distinct automata, we obtain:

Proposition 1.6 *If L is a recognisable language of A^*, the language $K^{-1}L$ is recognisable* for every subset K *of A^*.* ■

Remark 1.6 This last property is remarkable because it is one of the rare counter-examples to the general assertion that statements about finite automata and recognisable languages correspond to effective procedures for the construction of the relevant automata: we know that $K^{-1}L$ is recognisable *even though we cannot calculate* K, therefore even if we cannot calculate the automaton that recognises $K^{-1}L$. □

Exercises

- 1.9 Finish the proof of: (a) Equation (1.1); (b) Equation (1.2).
- 1.10 Construct an automaton that recognises the set of words that contain an odd number of a's and an even number of b's.

- 1.11 Let L be a language of A^*. Which languages are defined by $[A^*]^{-1}L$, $L\,[A^*]^{-1}$ and $[A^*]^{-1}L\,[A^*]^{-1}$?

- 1.12 Let $K_1 = \{a^n b^m \mid n, m \in \mathbb{N}\}$ be the language from Example 1.4.
Describe the languages $[a^*]^{-1}K_1$, $K_1\,[b^*]^{-1}$ and $[a^*]^{-1}K_1\,[b^*]^{-1}$. Do the same for $P_1 = Z_1 \cap K_1 = \{a^n b^n \mid n \in \mathbb{N}\}$.

- 1.13 Prove: $\forall f, g, h \in A^* \qquad (f\,g)^{-1}h = g^{-1}(f^{-1}h)$.

- 1.14 Prove: $\forall f, g, h \in A^* \qquad (f^{-1}g)^{-1}h = g^{-1}(f\,h) \quad$ if $\quad |f| \leqslant |g|$.

- 1.15 Show that a non-empty language L of A^* is prefix if and only if one of the following conditions is satisfied: *cf. Sec. 0.3, p. 22*

  (a) $f^{-1}L = 1_{A^*}$ for all f in L; (b) $L^{-1}L = 1_{A^*}$; (c) $L \cap L\,A^+ = \emptyset$.

- 1.16 **Shuffle product.** In this exercise, we introduce a notion widely used for the analysis of parallel processes, in the form of a *new product* on words. This product is called a *shuffle*.[10] The idea is that the behaviour of a set of independent processes viewed as a sequence of actions is the *shuffle*[11] of the behaviour of each process taken in isolation, hence the interest of Proposition 1.7 below.

  Definition 1.7 The shuffle of two words f and g from A^*, written $\between$, is the subset of A^* defined by:

  $$f \between g = \{f_1\,g_1\,f_2\,g_2 \cdots f_n\,g_n \mid f_i \in A^*,\ g_j \in A^* \quad f_1\,f_2 \cdots f_n = f \quad \text{and} \quad g_1\,g_2 \cdots g_n = g\}$$

  and extends additively to $\mathfrak{P}(A^*)$: $L \between K = \bigcup_{f \in L,\, g \in K} f \between g$. □

  In other words, $f \between g$ is the set of words h such that f and g are two disjoint subwords whose union covers h completely. For example: $a\,b \between b\,a = \{a\,b\,b\,a,\ b\,a\,b\,a,\ b\,a\,a\,b,\ a\,b\,a\,b\}$.

  (a) Give some (simple) examples of recognisable languages which are shuffles of two languages taken from previous examples.

  (b) Verify that shuffle makes $\mathfrak{P}(A^*)$ a commutative monoid with a zero.

  (c) Prove (taking your inspiration from the construction of the cartesian product of two automata):

  Proposition 1.7 *The shuffle of two recognisable languages is recognisable.*

- 1.17 **Infiltration product.** Infiltration is yet another product on words, very similar to shuffle, which can also be seen as a a way of describing the behaviour of parallel processes with slightly different conventions: if two similar processes each perform the same action at the same time, they *may* be identified.[12]

---

[10] The notation for this product is problematic, and varies considerably from one author to the next. Schützenberger uses the Cyrillic 'Щ' ('shch'), Eilenberg [80] a symbol inspired by it: ⧢ , which was also used in [31] and [198]. In [158, Ch. 6] and [187] I used a simple $\circ$. The symbol used here, found in the panoply of $\mathcal{AMS}$-TEX, is intended to suggest two packs of cards being shuffled together.

[11] It is notable that this product is an idea that has also been identified by mathematicians (in the study of Lie algebras; see [198] for example).

[12] This product was originally defined for the study of free groups in [54] – and was called *shuffle*! *Infiltration* is the word that I proposed in [158, Ch. 6].

The following notation is convenient for the definition of infiltration. If $h = a_1\, a_2 \cdots a_n$ is a word in A^* (the a_i are letters) and I a (possibly empty) subset of $[|h|]$, h_I denotes the word $a_{i_1}\, a_{i_2} \cdots a_{i_k}$ with $i_1 < i_2 < \cdots < i_k$ and $I = \{i_1, i_2, \ldots, i_k\}$.

Definition 1.8 The *infiltration* of two words f and g in A^*, written $\uparrow$, is the subset of A^* defined by:

$$f \uparrow g = \{h \in A^* \mid \exists I, J \subseteq [|h|] \quad I \cup J = [|h|] \ , \quad h_I = f \quad \text{and} \quad h_J = g\}$$

and extends additively to $\mathfrak{P}(A^*)$: $L \uparrow K = \bigcup_{f \in L,\, g \in K} f \uparrow g$. □

In other words, $f \uparrow g$ is the set of words h such that f and g are two subwords, not necessarily disjoint, whose union covers h completely. For example:

$$ab \uparrow ba = \{abba,\, aba,\, baba,\, baab,\, bab,\, abab\}\ .$$

(a) Give a definition of shuffle using the notation of Definition 1.8.

(b) Verify that infiltration makes $\mathfrak{P}(A^*)$ a commutative monoid with a zero.

(c) Prove:

Proposition 1.8 *The infiltration of two recognisable languages is recognisable.*

1.3 The graph perspective

As we have said, automata are fundamentally graphs, and finite automata are finite graphs. The basic properties of finite automata are those of finite graphs; and the first operation that one performs on graphs is traversal. It is thus that we will obtain some properties of automata, essential although quite elementary, by algorithms that correspond to the traversal of underlying graphs.

cf. Sec. 0.7, p. 33

The first question that one asks oneself when one is given – or rather when, in the course of a set of computations, one obtains – an automaton, is to know if that automaton really recognises something: in other words, if the language that it recognises is *non-empty*.[13] If it is not empty, a second reasonable question is to ask whether the language is *infinite*.

A third question, the inverse in some sense of those first two, is to ask whether a subset P of A^* – defined in some other way than as an automaton – is or is not recognisable: that is, whether there is a *finite* automaton which recognises it or not. We have already seen a property, the finiteness of the set of quotients, which is characteristic of recognisable languages and which therefore allows us to find the answer. However, it can be difficult to discover whether the set of quotients is finite or not. We therefore need to find a simpler way of showing the contradiction between the definition of P and the assumption that P can be recognised by a finite automaton, in the case where P is not recognisable.

cf. Propty 1.4, p. 64 and Prop. 3.10, p. 114

[13] This is the case, for example in certain automatic protocol verification systems, whose final step consists of verifying that an automaton, which represents sequences of interactions that are invalid according to the protocol, recognises no words.

These three questions refer solely to the existence of certain computations in an automaton, independently of their labels, and therefore to the existence of paths in the graph underlying the automaton. These are standard problems of graph traversal, which the assumed finiteness of the graphs in question makes trivial. We can summarise in two sentences the essence of what we will develop in this subsection: 'if there exists a path between two vertices in a graph that contains n vertices, there exists one of length less than n' and 'all paths of length greater than or equal to n in such a graph contain at least one non-trivial cycle'. No matter that the answers to these questions are conceptually obvious; once more we need methods that are algorithmically efficient.

1.3.1 Trim automata

In order to relate the properties of automata to those of the languages they recognise, we must be able to distinguish the 'useful bits' of each automaton, which are actually used by successful computations. For this, we need several definitions.

Definition 1.9 Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton.

A state q of $\mathcal{A}$ is *accessible from a state* p if there is a computation in $\mathcal{A}$ whose source is p and whose destination is q. The state q is *accessible* if it is accessible from an initial state.

A state p is *co-accessible to a state* q if there is a computation in $\mathcal{A}$ whose source is p and whose destination is q. The state p is *co-accessible* if it is co-accessible to a final state.

We will say that a state p is *useful* if it is both accessible and co-accessible. □

In other words, q is accessible from p, or p is co-accessible to q if and only if there exists a path from p to q in the graph underlying $\mathcal{A}$. We are following the convention set down in the previous subsection: p is co-accessible in $\mathcal{A}$ if and only if p is accessible in $\mathcal{A}^{\mathsf{t}}$ and *vice versa*. The useful states of $\mathcal{A}$ and $\mathcal{A}^{\mathsf{t}}$ are the same. The term 'useful' is justified by the following property.

Property 1.5 *A state p of $\mathcal{A}$ is useful if and only if there exists a successful computation in $\mathcal{A}$ which passes through p.*

Proof. If p is accessible, there exists a computation $i \xrightarrow[\mathcal{A}]{f} p$ where i is in I; if p is co-accessible, there exists a computation $p \xrightarrow[\mathcal{A}]{g} t$ where t is in T. The computation

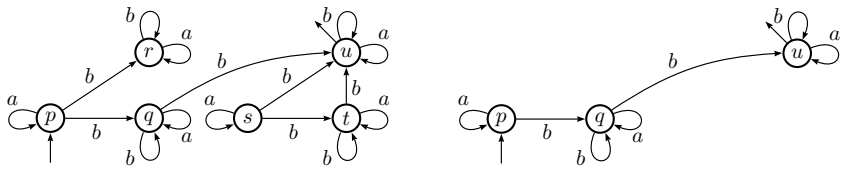
$$c \;:=\; i \xrightarrow[\mathcal{A}]{f} p \xrightarrow[\mathcal{A}]{g} t$$

is a successful calculation which passes through p. Conversely, a successful calculation which passes through p makes p *ipso facto* both accessible and co-accessible. ■

Definition 1.10 We say that an automaton $\mathcal{A}$ is *accessible* if all its states are accessible; that $\mathcal{A}$ is *co-accessible* if all its states are co-accessible; and that $\mathcal{A}$ is *trim* if it is both accessible and co-accessible: that is, if all its states are useful. □

For example, the automata $\mathcal{A}_1$ and $\mathcal{Z}_1$ above are trim. The automaton $\mathcal{Z}_2$ constructed in Figure 1.8 is neither accessible nor co-accessible. This can just as well be the case with a finite automaton: in the automaton of Figure 1.10 (a), the state r is not co-accessible and the states s and t are not accessible.

cf. Exam. III.2.5, p. 467



(a) an automaton that is neither accessible nor co-accessible

(b) the equivalent trim automaton

Figure 1.10: Trimming an automaton

Proposition 1.9 *Every automaton $\mathcal{A}$ is equivalent to a trim automaton $\mathcal{A}_\mathsf{m}$. If $\mathcal{A}$ is finite, $\mathcal{A}_\mathsf{m}$ is also and can be effectively computed from $\mathcal{A}$.*

Proof. Let $\mathcal{A} = \langle\, Q, A, E, I, T \,\rangle$ be an automaton. The set Q_m of useful states of $\mathcal{A}$ defines the sub-automaton $\mathcal{A}_\mathsf{m}$ of $\mathcal{A}$:

$$\mathcal{A}_\mathsf{m} = \langle\, Q_\mathsf{m}, A, E_\mathsf{m}, I_\mathsf{m}, T_\mathsf{m} \,\rangle \ ,$$

with $E_\mathsf{m} = E \cap [Q_\mathsf{m} \times A \times Q_\mathsf{m}]$, $I_\mathsf{m} = I \cap Q_\mathsf{m}$ and $T_\mathsf{m} = T \cap Q_\mathsf{m}$. The automaton $\mathcal{A}_\mathsf{m}$ is called the *trim (sub-)automaton* or *trim part* of $\mathcal{A}$; it is equivalent to $\mathcal{A}$.

In fact, if c is a successful computation in $\mathcal{A}$, all the states through which it passes are useful (Property 1.5), and all the transitions which form part of it are in E_m: it is a computation of $\mathcal{A}_\mathsf{m}$; conversely, every successful computation of $\mathcal{A}_\mathsf{m}$, a sub-automaton of $\mathcal{A}$, is a successful computation of $\mathcal{A}$.

cf. Propty 1.2, p. 56

Finally, $\mathcal{A}_\mathsf{m}$ is trim: every state of Q_m belongs to a successful computation of $\mathcal{A}$ by definition, hence to a successful computation of $\mathcal{A}_\mathsf{m}$. It remains to show that $\mathcal{A}_\mathsf{m}$ is effectively computable, which we will do in the next subsection. ■

The automaton in Figure 1.10 (b) was obtained from that in Figure 1.10 (a) by the construction given in the proof of Proposition 1.9: in this case, $Q_\mathsf{m} = \{p, q, u\}$. We conclude directly from Proposition 1.9 an important corollary:

cf. also Exer. 1.11, p. 65

Proposition 1.10 *The set of prefixes (respectively of suffixes, and of factors) of a recognisable language is recognisable.*

Proof. Let $L = L(\mathcal{A})$ be a recognisable language. By Proposition 1.9, we can assume that $\mathcal{A} = \langle\, Q, A, E, I, T \,\rangle$ is trim.

Let $c := i \xrightarrow[\mathcal{A}]{g} p$ be a computation whose source is an initial state. Since $\mathcal{A}$ is trim, p is co-accessible and there exists a computation $p \xrightarrow[\mathcal{A}]{h} t$ where t is a final state. Hence, $g\,h$ is in L and g is a prefix of L. Conversely, every prefix of a word in L is the label of a computation in $\mathcal{A}$ whose source is an initial state.

By this argument the set of prefixes of L, $\mathsf{Pre}(L)$, is recognised by the automaton $\langle Q, A, E, I, Q \rangle$. Similarly, the set of suffixes of L, $\mathsf{Suf}(L)$, is recognised by the automaton $\langle Q, A, E, Q, T \rangle$, and the set of factors of L, $\mathsf{Fac}(L)$, is recognised by the automaton $\langle Q, A, E, Q, Q \rangle$. ■

Remark 1.7 Proposition 1.9 *does not say* that $\mathcal{A}_{\mathsf{m}}$ is finite only if $\mathcal{A}$ is. On the contrary, we will discover that certain constructions give *a priori* infinite automata that show themselves to be finite once they are trimmed. □

cf. Exam. II.3.3, p. 268 and Exam. IV.6.1 p. 605

Remark 1.8 Abusing the terminology we will say – rightly, in the cases mentioned above – that an automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is *trim* (respectively accessible, co-accessible) even if not all the elements of Q are useful states (respectively accessible, co-accessible), but when just the sources and destinations of the elements of E have that property. □

cf. Rem. 3.2, p. 108

1.3.2 The empty and the infinite

It is a tautology to say that $L(\mathcal{A})$ is not empty if and only if there exists a successful computation, that is to say, if there exists a sequence of transitions which leads from an initial state to a final state; in other words, if there is an *accessible final state* (or a *co-accessible initial state*).

To know whether $L(\mathcal{A})$ is empty or not, we need to calculate, for each initial state i of $\mathcal{A}$, the set of states that are accessible from i: that is, to *traverse the graph* underlying $\mathcal{A}$. If $\mathcal{A}$ is finite this computation is effective, since all the sets in question are finite and the problem of emptiness is therefore decidable. Standard, well-understood graph algorithms perform these computations, that is, the traversal of the graph, in an efficient manner. In the same way that we decide whether a state is accessible we also determine whether it is co-accessible (since that is the same as deciding whether it is accessible in the transposed automaton), hence whether it is useful: the automaton $\mathcal{A}_{\mathsf{m}}$ is effectively computable from $\mathcal{A}$.

cf. Prop. 1.11 below

Remark 1.9 In light of the examples with which we have so far illustrated our definitions and results, the reader may wonder why we lay such stress on the obvious. The answer is the sheer scale of automata we encounter in practice – generated by verification programs, for example – which can have millions of states. In this situation, finding whether the language recognised by an automaton is empty or not – which is the same as knowing whether the system described by the automaton is valid or not – is neither cheap nor trivial. There is no longer any question of a graphical representation or of 'seeing' anything. The automaton is represented as a data structure, and it is important to have an algorithm that can not only answer the question but do so efficiently. □

Whether the language recognised by a finite automaton $\mathcal{A}$ (whose alphabet A is assumed to be finite) is infinite is of the same nature as the question of its emptiness. In fact, $L(\mathcal{A})$ is infinite if and only if it contains arbitrarily long words and therefore

cf. Def. 1.2, p. 53

in particular of length greater than or equal to the number $N = \|Q\|$ of states of $\mathcal{A}$. A computation which accepts such a word $f = a_1 a_2 \cdots a_l$ can be written:

$$q_0 \xrightarrow{a_1} q_1 \xrightarrow{a_2} q_2 \xrightarrow{a_3} \cdots \xrightarrow{a_{l-1}} q_{l-1} \xrightarrow{a_l} q_l \ , \tag{1.4}$$

where q_0 is in I and q_l in T. *As there are* $l+1 \gtrneqq N$ *states* q_i*, they cannot all be distinct*[14] and at least two, say q_j and q_k, must be equal to the same state p. We set:

$$u = a_1 a_2 \cdots a_j \ , \quad v = a_{j+1} a_{j+2} \cdots a_k \quad \text{and} \quad w = a_{k+1} a_{k+2} \cdots a_l \ ,$$

with the convention that $u = 1_{A^*}$ if $j = 0$, and $w = 1_{A^*}$ if $k = l$; the word v is certainly non-empty since $j \neq k$. The computation (1.4) is therefore written:

$$q_0 \xrightarrow{u} p \xrightarrow{v} p \xrightarrow{w} q_l \ . \tag{1.5}$$

From this we conclude that if $L(\mathcal{A})$ is infinite, there exists a useful state p which is accessible from itself by a path whose length is strictly positive. Conversely, if such a useful state exists in $\mathcal{A}$, there must exist in $\mathcal{A}$ a successful computation such that (1.5) holds. It follows that all the computations

$$q_0 \xrightarrow{u} \overbrace{p \xrightarrow{v} p \xrightarrow{v} p \cdots p \xrightarrow{v} p}^{n \text{ times}} \xrightarrow{w} q_l$$

are successful and that $L(\mathcal{A})$ is infinite. (We shall resume this line of reasoning in §1.3.3; here we will consider only the consequences for decidability.)

The same algorithm for traversing the graph therefore allows us to decide if a state is accessible from itself by a path of strictly positive length. The above and the performance of the algorithm which it concerns can be summarised as follows:

cf. Exer. 1.18, sol. p. 183

Proposition 1.11 *Let $\mathcal{A}$ be an automaton with k initial states and m transitions. We can, with complexity $O(k+m)$,*

(i) *decide whether $L(\mathcal{A})$ is empty, finite or infinite,*

(ii) *calculate the trim part $\mathcal{A}_{\mathsf{m}}$ of $\mathcal{A}$.* ■

1.3.3 Criteria for recognisability

cf. Exam. 1.6, p. 64

We know how to show that the language $P_1 = \{a^n b^n \mid n \in \mathbb{N}\}$ is not recognisable, using Property 1.4, based on the notion of the *quotient of a language by a word*. However, even if we had not already brought this notion out, it would have been fairly clear that P_1 could not be recognisable: how could it, with a finite number of states, remember the arbitrarily large number of a's that it reads in the first part of the word to ensure that the number of b's in the second part is the same? This is the intuition that we will formalise and translate into a necessary condition that recognisable languages must satisfy and which, in many cases, will show itself to be easier to apply than the aforementioned property. This condition will be stated in three increasingly precise forms.

[14] This argument is known as the *Pigeonhole Principle*. *Cf.* also § 3.4.2, p. 119.

Anticipating the definition and notation which will be introduced in Section 2, we will from now on write the union[15] of all the powers of a word f as $f^* = \bigcup_{n\in\mathbb{N}} f^n$. *cf. Def. 2.1, p. 83*

Lemma 1.12 [**Star Lemma (simplistic)**] *If L is a recognisable language of A^*, there exists an integer N such that, for all f in L of length greater than N, there is a factorisation of f, $f = u\,v\,w$, with v non-empty, such that, for all n in $\mathbb{N}$, the word $u\,v^n\,w$ is in L; that is, such that $u\,v^*\,w \subseteq L$.*

A word v which satisfies the conditions of the lemma will be called an *iterating factor* of L; and Lemma 1.12 may be read: 'every sufficiently long word in L contains an iterating factor'. It is no harder to prove a more precise statement that is significantly more useful: 'every sufficiently long factor of a word in L contains an iterating factor', in other words:

Lemma 1.13 [**Star Lemma**] *If L is a recognisable language of A^*, there exists an integer N such that, for all f in L and* every factorisation *$f = g_1\,h\,g_2$ such that $|h| \geqslant N$, there exists a factorisation $h = u\,v\,w$, with v non-empty, such that, for all n in $\mathbb{N}$, the word $g_1\,u\,v^n\,w\,g_2$ is in L; that is, such that*

$$g_1\,u\,v^*\,w\,g_2 \subseteq L \quad .$$

In fact, we can prove an even more precise statement.

Lemma 1.14 [**Block Star Lemma**] *If L is a recognisable language of A^*, there exists an integer N such that for every f in L and every factorisation of the form $f = u\,v_1\,v_2\cdots v_N\,w$, where the v_i are all non-empty words, there is a pair (j,k) of indices, $0 \leqslant j < k \leqslant N$, such that*

$$u\,v_1\,v_2\cdots v_j\,(v_{j+1}\cdots v_k)^*\,v_{k+1}\cdots v_N\,w \subseteq L \quad .$$

Proof. It suffices to repeat the reasoning used to characterise automata that recognise an infinite language. Let $\mathcal{A} = \langle Q, A, E, I, T\rangle$ be an automaton which recognises L. Set $N = \|Q\|$. A computation that accepts f can be written

$$i \xrightarrow{u} q_0 \xrightarrow{v_1} q_1 \xrightarrow{v_2} q_2 \xrightarrow{v_3} \cdots \xrightarrow{v_{N-1}} q_{N-1} \xrightarrow{v_N} q_N \xrightarrow{w} t \quad . \tag{1.6}$$

The $N+1$ states q_i cannot all be distinct, and at least two, say q_j and q_k, are equal to the same state p. We set

$$x = v_1\,v_2\cdots v_j \quad , \quad y = v_{j+1}\,v_{j+2}\cdots v_k \quad \text{and} \quad z = v_{k+1}\,v_{k+2}\cdots v_N \quad .$$

As above, x or z can be empty, but y is not empty since $j \neq k$. The computation (1.6) can therefore be written

$$i \xrightarrow{u} q_0 \xrightarrow{x} p \xrightarrow{y} p \xrightarrow{z} q_N \xrightarrow{w} t \quad ,$$

[15] Which is read 'f star'.

from which we conclude that, for every integer n,

$$i \xrightarrow{u} q_0 \xrightarrow{x} \overbrace{p \xrightarrow{y} p \xrightarrow{y} p \cdots p \xrightarrow{y} p}^{n \text{ times}} \xrightarrow{z} q_N \xrightarrow{w} t$$

is a successful computation of $\mathcal{A}$. ■

We can obtain Lemma 1.12 from Lemma 1.13 by setting $g_1 = g_2 = 1_{A^*}$ and $h = f$, and Lemma 1.13 from Lemma 1.14 by taking v_i from A for all i and setting[16] $g_1 = u$, $g_2 = w$, $u = x$, $v = y$, and $w = z$.We can now resume the proof that the language P_1 is not recognisable.

Example 1.6 (continued) Suppose that the language $P_1 = \{a^n b^n \mid n \in \mathbb{N}\}$ is recognisable, and let N be the integer associated with it by Lemma 1.13. Consider the word $f = a^N b^N$ and, in this word, the factor a^N. By Lemma 1.13 there exist integers r, s and t, s strictly positive, such that $r + s + t = N$ and for all n, we have $(a^r)(a^s)^n(a^t)(b^N) = (a^{r+t+n\,s})(b^N)$ is in P_1. We obtain a contradiction by taking $n = 0$: $r + t \neq N$ since s is strictly positive. □

Remark 1.10 The language P_1 is the most easily defined one that is not recognisable by a finite automaton. Very often, the proof that a language X is not recognisable consists of showing that if it were, then P_1 would be too. □

Lemmas 1.12, 1.13 and 1.14 express combinatorial properties which are necessarily satisfied by all recognisable languages. It must be added that these properties are by no means characteristic of recognisable languages: a language can very well satisfy them without being recognisable. What is more, the search for such examples will show that these properties are distinct, and each is strictly stronger than the last, since one can find non-recognisable languages that satisfy Lemma 1.12 but not Lemma 1.13, and others that satisfy Lemma 1.13 but not Lemma 1.14.

cf. Exer. 1.24

We will see in Subsection 3.4 that we can strengthen the statement of Lemma 1.14 to make it characteristic of recognisable languages.

cf. Th. 3.3, p. 119

Remark 1.11 The expressions $u\,v^*\,w$, $g_1\,u\,v^*\,w\,g_2$ and

$$u\,v_1\,v_2 \cdots v_j\,(v_{j+1} \cdots v_k)^*\,v_{k+1} \cdots v_N\,w$$

explain the generic name 'star lemma' given to these results, which are also often called *pumping lemmas*: the ability to add multiple v's to enlarge the word f while staying in the language being a sort of 'pumping'.

The use of this and similar lemmas to draw the boundaries of various families of languages is so widespread and systematic that these lemmas have become, for

[16] This relabelling is a little involved; that is the price we pay for using wording similar to that used in other works.

certain authors, emblematic of the theory of languages.[17] Without denying their utility and ubiquity, I think that on the contrary their use is an admission of ignorance. Wheeling them out to establish a property leads to easy proofs but gives no structural information other than that a certain set is finite: the very minimum in this domain. There are many examples of detours via more or less complex structures which allow us to obtain as corollaries results which we could have proved directly with the Star Lemma, but on the way we will have acquired a much better understanding of the object in question than a simple demonstration of impossibility.

It seems to me therefore that one of the goals of automata theory should be to go, whenever possible, beyond the Star Lemma. □

Exercises

• 1.18 **Complexity of accessibility.** Verify the complexity stated in Proposition 1.11.

• 1.19 **Roy–Warshall algorithm.** An algorithm for calculating the *transitive closure* of a graph $\mathcal{G} = \langle Q, G \rangle$.

We begin by totally ordering the set Q of vertices of $\mathcal{G}$, from 1 to $n = \|Q\|$: that is, to set $Q = [n]$. The main idea of the algorithm is to consider, for each integer k, $0 \leqslant k \leqslant n$, the graph $\mathcal{F}_k = \langle Q, F_k \rangle$ such that there exists an edge from i to j in $\mathcal{F}_k$ if and only if there exists a path in $\mathcal{G}$ from i to j all of whose intermediate vertices are less than or equal to k. Clearly, $\mathcal{F}_0 = \mathcal{G}$ and $\mathcal{F}_n$ is the transitive closure of $\mathcal{G}$.

Give an algorithm that calculates $\mathcal{F}_n$ with complexity $O(n^3)$.

• 1.20 **Transitive closure revisited.** Once again, let $\mathcal{G} = \langle Q, G \rangle$ be a graph. Considering the edges rather than the nodes of $\mathcal{G}$, give an algorithm that computes the transitive closure $\mathcal{F}$ of $\mathcal{G}$ with complexity $O(d\,m')$ where m' is the number of edges in $\mathcal{F}$ and d is the largest number of edges that leave any node of $\mathcal{G}$ (maximum external degree).

1.21 Show that the following languages are not recognisable:

(a) $T_1 = \{a^{n^2} \mid n \in \mathbb{N}\}$; (b) $T_2 = \{a^n \mid n \in \mathbb{N},\ n \text{ prime}\}$.

• 1.22 Show that the following languages over $A = \{a, b\}$ are not recognisable:

(a) $Z_2' = \{a^n\, b^m \mid n, m \in \mathbb{N} \quad n \leqslant m\}$; (b) $Z_3' = \{a^n\, b^m \mid n, m \in \mathbb{N} \quad n \neq m\}$;

(c) $Z_1 = \{f \in A^* \mid |f|_a = |f|_b\}$;

(d) the language of palindromes: $\mathsf{Pal} = \{f \in A^* \mid f = f^{\mathrm{t}}\}$;

(e) the language of squares: $\mathsf{Car} = \{f \mid \exists g \in A^* \quad f = g^2\}$.

• 1.23 (a) Give an example of a non-recognisable language X such that $\mathsf{Pre}(X)$ and $\mathsf{Suf}(X)$ are recognisable.

(b) Give an example of a language X such that $\mathsf{Pre}(X)$ and $\mathsf{Suf}(X)$ are non-recognisable and that $\mathsf{Fac}(X)$ is recognisable.

[17] The cover of Michael Harrison's book *Introduction to Formal Language Theory* [112] shows two lemmings (small rodents of the family *Muridæ*) using a garden pump.

1.24 **Hierarchy of star lemmas.**

(a) Show that the language $U_1 = \{a^n b^n \mid n \in \mathbb{N}\} \bigcup A^* b a A^*$ satisfies Lemma 1.12 but not Lemma 1.13.

(b) Show the same property for $U_1' = \{a^n b^m \mid n, m \in \mathbb{N} \quad n \neq m\} \bigcup A^* b a A^*$.

(c) Show that the language $U_2 = \{(a a b)^n (a b b)^n \mid n \in \mathbb{N}\} \bigcup A^* \{a a a, a a b b, b b b\} A^*$ satisfies Lemma 1.13 but not Lemma 1.14.

(d) Let $A = \{a, b, c\}$ be a three-letter alphabet and d a fourth distinct letter. Let V, W and U_3 be the languages defined as follows: $V = \{f f \mid f \in A^*\}$, $W = \{f d g \mid f, g \in A^* \quad f \neq g\}$, and $U_3 = W \cup A^* V A^* d A^* \cup A^* d A^* V A^*$. Show that the language U_3 satisfies Lemma 1.14 but is not recognisable.

(e) Show that the language $U_4 = \{a^{k_1} b a^{k_2} b \ldots a^{k_n} b \mid \exists i \quad i \neq k_i\}$ does not satisfy Lemma 1.14.[18]

• 1.25 **k-recognisable sets of numbers.** We say that a subset P of $\mathbb{N}$ is *k-recognisable* if the set of base k representations of elements of P is a recognisable language over the alphabet $\mathbf{k} = \{0, 1, \ldots, k-1\}$.

Show that the set of powers of 3, $T = \{3^n \mid n \in \mathbb{N}\}$ is 3-recognisable but not 2-recognisable.

1.4 Some supplementary definitions

To end this first section, we define three properties of automata and present a generalisation of automata over an alphabet. These will be found useful in what follows.

1.4.1 Unambiguous automata

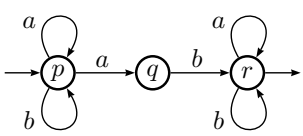
Definition 1.11 An automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is *unambiguous* when it satisfies the two following conditions:

(i) for every pair of states (p, q) and every word f in A^*, there exists *at most* one computation from p to q (that is, with source p and destination q) labelled f;

(ii) for every word f in $L(\mathcal{A})$, there exists a unique i in I and a unique t in T such that $i \xrightarrow{f} t$. □

cf. Exer. 1.28 If $\mathcal{A}$ is trim, the two conditions (i) and (ii) of this definition merge into one:

(iii) *for all f in $L(\mathcal{A})$, there exists a unique successful computation in $\mathcal{A}$ with label f.*



The automaton $\mathcal{A}_1$

For example, the automaton $\mathcal{A}_1$ is *not* unambiguous (we say that $\mathcal{A}_1$ is *ambiguous*) since the word $a b a b$ is the label of two distinct computations with source p and destination r:

$$p \xrightarrow{a} p \xrightarrow{b} p \xrightarrow{a} q \xrightarrow{b} r \qquad \text{and} \qquad p \xrightarrow{a} q \xrightarrow{b} r \xrightarrow{a} r \xrightarrow{b} r \ .$$

[18] *Note added in translation:* Both the example given by the language U_3 and a proof for (e) are due to Olivier Carton (personnal communication) who thus corrected a mistake in the French edition.

To paraphrase Definition 1.11, a (trim) automaton is unambiguous if and only if the words that it recognises are recognised *in a unique way*. We intend this uniqueness property to be important, and indeed the notion of ambiguity is a recurrent theme in this work. In Chapter III, we will make definitions that allow us to treat fully these notions and the properties which they make it possible to express. By then, we will have had the opportunity to consider some of these properties at an elementary level. *cf. pp. 408 et seq.*

Proposition 1.15 *It is decidable whether a finite automaton is unambiguous.*

Let us first make a definition. Let $\mathcal{A}\times\mathcal{A} = \langle Q\times Q, A, F, I\times I, T\times T\rangle$ be the cartesian product of $\mathcal{A} = \langle Q, A, E, I, T\rangle$ with itself; the set F of transitions is given by:

$$F = \{\big((p,r), a, (q,s)\big) \mid (p,a,q), (r,a,s) \in E\} \ .$$

The *diagonal* of $\mathcal{A}\times\mathcal{A}$ is a sub-automaton $\mathcal{D}$ of $\mathcal{A}\times\mathcal{A}$ given by the diagonal D of $Q\times Q$: $D = \{(q,q) \mid q \in Q\}$. Clearly, the states and transitions of $\mathcal{A}$ and $\mathcal{D}$ are in bijection, and therefore $\mathcal{A}$ and $\mathcal{D}$ are equivalent. Proposition 1.15 follows immediately from the following property.

Lemma 1.16 *A trim automaton $\mathcal{A}$ is unambiguous if and only if the trim part of $\mathcal{A}\times\mathcal{A}$ is the diagonal $\mathcal{D}$ of $\mathcal{A}\times\mathcal{A}$.*

Proof. By definition, $\mathcal{A}$ is ambiguous if and only if there exist two distinct successful computations c' and c'' in $\mathcal{A}$ which have the same label $f = a_1 a_2 \cdots a_n$:

$$c' \ := \ q'_0 \xrightarrow[\mathcal{A}]{a_1} q'_1 \xrightarrow[\mathcal{A}]{a_2} \cdots \xrightarrow[\mathcal{A}]{a_n} q'_n \quad \text{and} \quad c'' \ := \ q''_0 \xrightarrow[\mathcal{A}]{a_1} q''_1 \xrightarrow[\mathcal{A}]{a_2} \cdots \xrightarrow[\mathcal{A}]{a_n} q''_n :$$

that is, if and only if there exists a successful computation c in $\mathcal{A}\times\mathcal{A}$:

$$c \ := \ (q'_0, q''_0) \xrightarrow[\mathcal{A}\times\mathcal{A}]{a_1} (q'_1, q''_1) \xrightarrow[\mathcal{A}\times\mathcal{A}]{a_2} \cdots \xrightarrow[\mathcal{A}\times\mathcal{A}]{a_n} (q'_n, q''_n)$$

where, for at least one i, $0 \leqslant i \leqslant n$, we have $q'_i \neq q''_i$, and hence if and only if there exists a useful state in $\mathcal{A}\times\mathcal{A}$ that is not in $\mathcal{D}$. ■

Example 1.7 Figure 1.11 shows the construction underlying Lemma 1.16 (a) for an ambiguous automaton, and (b) for an unambiguous automaton. □

Proof of Proposition 1.15. Since the computation of the trim part of a finite automaton is effective, Lemma 1.16 proves the proposition in the case where $\mathcal{A}$ is trim, which is the interesting case.

If $\mathcal{A}$ is not trim, the lemma allows us to check conditions (i) and (ii) of Definition 1.11 for the trim part of $\mathcal{A}$. If these are satisfied, we take D to be the set of initial and final states in $\mathcal{A}\times\mathcal{A}$ and compute the trim part of this new automaton. If this is equal to $\mathcal{D}$, condition (i) is satisfied for the non-useful states of $\mathcal{A}$ and $\mathcal{A}$ is unambiguous; if not, $\mathcal{A}$ is ambiguous. ■

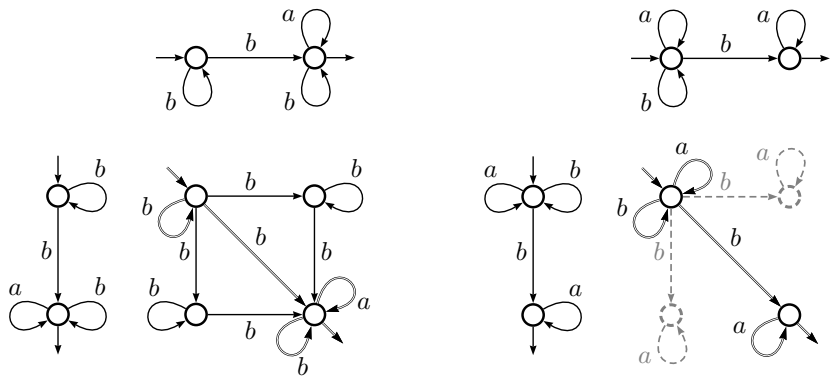
The construction of an unambiguous automaton which is equivalent to a given automaton will be performed by making the automaton deterministic; another con- *cf. Sec. 3, p. 101*

struction, which is more complex but which does not destroy the structure of the automaton, will be given in Chapter II (to be used in Chapters III and V).

cf. Sec. II.3, p. 255

Ambiguity is an 'undirected' or 'auto-dual' notion; in other words, $\mathcal{A}$ is unambiguous if and only if $\mathcal{A}^{\mathsf{t}}$ is unambiguous. We now consider two 'directed' notions (from the source to the destination of transitions, or more simply *from left to right*).

cf. § 1.1.3, p. 58



(a) the ambiguous case (b) the unambiguous case

Figure 1.11: The construction of Lemma 1.16. The double lines indicate the diagonal of the cartesian square; the dashed grey part, the subset that is accessible but not co-accessible, hence not useful.

1.4.2 Complete automata

Definition 1.12 An automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is *complete* if, for all p in Q and all a in A, there exists q in Q such that (p, a, q) is a transition in $\mathcal{A}$. □

By induction on the length of f, $\mathcal{A}$ is complete if and only if there exists at least one computation with source p and label f, for every state p and every word f in A^*.

In the examples above, $\mathcal{Z}_1$, $\mathcal{P}_2$, and many others, are complete; the automaton $\mathcal{A}_1$ is not complete, since it has no transition with source q and label a.

Since the condition depends only on the sets of transitions, it is easy to check whether a finite automaton is complete.

If $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is not complete, it is also easy to transform it into an equivalent complete automaton. Let $\mathcal{A}_{\mathsf{c}} = \langle Q_{\mathsf{c}}, A, E_{\mathsf{c}}, I, T \rangle$ with $Q_{\mathsf{c}} = Q \cup z$ where z is a new state (that is, z is not in Q). To construct E_{c}, add to E a transition (p, a, z), for each pair (p, a) in $Q \times A$ where there is no state q such that (p, a, q) is in E, and also a transition (z, a, z), for each a in A.

Example 1.1 (continued) Figure 1.12 shows the effect of the above construction on the automaton $\mathcal{A}_1$. □

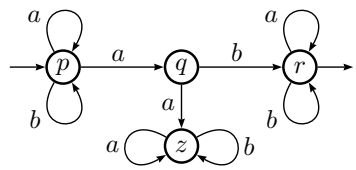


Figure 1.12: A complete automaton $\mathcal{A}_{1\mathsf{c}}$, equivalent to $\mathcal{A}_1$

By construction, $\mathcal{A}_\mathsf{c}$ is complete. On the other hand, $\mathcal{A}_\mathsf{c}$ is equivalent to $\mathcal{A}$: every (successful) computation of $\mathcal{A}$ is a (successful) computation of $\mathcal{A}_\mathsf{c}$; conversely, a computation of $\mathcal{A}_\mathsf{c}$ which is not in $\mathcal{A}$ goes through z, but every computation that goes through z terminates at z since all the transitions whose source is z have z as their destination; since z is not final such a computation is not successful. (We shall call a state like z a *sink state*, because when it is reached, the computation stays there.) Finally, if $\mathcal{A}$ is accessible (and not complete), $\mathcal{A}_\mathsf{c}$ is also accessible. We can therefore state, since each automaton is equivalent to an accessible automaton:

Property 1.6 *Every automaton is equivalent to a complete accessible automaton.* ■

Remark 1.12 The automaton $\mathcal{A}_1$ is trim but is not complete; $\mathcal{A}_{1\mathsf{c}}$ is complete but is not trim since the state z is not co-accessible. In general, given an automaton $\mathcal{A}$ it is not always possible to find an equivalent automaton which is both complete and trim. □

Remark 1.13 We can define the dual notion: an automaton $\mathcal{A}$ is called *co-complete* if $\mathcal{A}^t$ is complete: that is, if, for all q in Q and all a in A, there exists p in Q such that (p, a, q) is a transition in $\mathcal{A}$; but we will have no use for this term. □

1.4.3 Deterministic automata

The third definition introduces a fundamental property, *determinism*, which distinguishes a whole class of automata. Section 3 will be devoted mainly to their study and the proof that every automaton is equivalent to a deterministic automaton. Nevertheless we give the definition now in order to complete this first section's tour of the major properties of automata, and to be able to indicate, for the constructions we meet later, which of them preserve determinism and which do not. *cf. p. 101*

Definition 1.13 An automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is *deterministic* if the two following conditions are satisfied:

(i) for all p in Q and all a in A, there exists *at most one transition* in E with source p and label a;

(ii) there is at most one initial state: that is, I is a singleton, or empty. □

Among the preceding examples, $\mathcal{A}_1$ is not deterministic since (p, a, p) and (p, a, q) are in E_1; the automata $\mathcal{P}_2$ and $\mathcal{Z}_1$ are deterministic.

Since the conditions deal with just the sets of transitions and initial states, it is easy to test whether a finite automaton is deterministic. Notice that we could also define determinism with the notions of Lemma 1.16:

cf. Exer. 1.31; cf. also Prop. V.4.3, p. 699 and Note V.37

Property 1.7 *An accessible automaton $\mathcal{A}$ is deterministic if and only if the* accessible part *of $\mathcal{A}\times\mathcal{A}$ is the diagonal $\mathcal{D}$ of $\mathcal{A}\times\mathcal{A}$.* ■

Property 1.8 *A deterministic automaton is unambiguous.*

Proof. By induction on the length of f, we have that if $\mathcal{A}$ is deterministic there exists, for every state p and every word f in A^*, at most one computation with source p and label f. Hence we conclude that if f is the label of a computation with source p and destination q, this computation is unique. ■

The converse of this property is obviously false.

We can define the dual notion, following Definition 1.13: the automaton $\mathcal{A}$ is *co-deterministic* if $\mathcal{A}^{\mathsf{t}}$ is deterministic: that is, if:

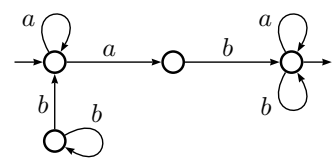
(iii) *for all q in Q and all a in A, there exists* at most one transition *in E with destination q and label a;*
(iv) *there is* a unique *terminal state: that is, T is a singleton set.*

If $\mathcal{A}$ is co-deterministic there exists, for all q in Q and all f in A^*, at most one computation with label f and destination q. We therefore have:

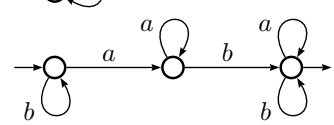
Property 1.9 *A co-deterministic automaton is unambiguous.* ■

Example 1.1 (continued) Some automata equivalent to $\mathcal{A}_1$, which is *ambiguous*:

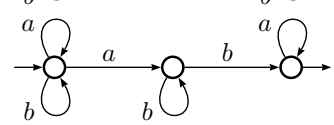
(a) An unambiguous automaton which is neither deterministic nor co-deterministic.



(b) A deterministic automaton.



(c) A co-deterministic automaton.



The union of two deterministic automata is not deterministic since it has two initial states. On the other hand, the family of deterministic automata is closed under product:

cf. Exer. 1.32

Property 1.10 *The product of two deterministic automata is a deterministic automaton.* ■

1.4.4 Automata with spontaneous transitions

The definition of a computation in an automaton is independent of the nature of the labels of the automaton's transitions. The label of a computation is the concatenation – the product – of the labels of the transitions. There is therefore nothing to stop us considering automata whose transitions have labels that are *words* in A^*, rather than restricting ourselves to cases where the labels are merely *letters*.[19] The question that then presents itself is whether these 'generalised' automata are more powerful: that is whether they recognise languages which are not recognised by classical automata. As we shall now see, they can do nothing of the kind. What we shall gain from this is not the definition of a richer family of languages, but more powerful tools to study the languages we already have.

It should already be obvious that a (finite) automaton $\mathcal{A}$ in which some transitions are labelled with non-empty words is equivalent to a (finite) automaton $\mathcal{B}$ all of whose transitions are labelled with letters, and that it is possible effectively to construct $\mathcal{B}$ given $\mathcal{A}$ (assuming $\mathcal{A}$ is finite).

cf. Exer. 1.36

A *spontaneous transition*[20] in an automaton $\mathcal{A}$ is a transition labelled with 1_{A^*}: that is, a triple $(p, 1_{A^*}, q)$. This corresponds to the intuitive idea that such a transition represents the possibility that $\mathcal{A}$ can go from state p to state q without reading anything. The set of transitions in an automaton which has spontaneous transitions is therefore the union of two sets: a set E of 'normal' transitions, and a set S of spontaneous transitions; we can therefore write

cf. the model due to Rabin & Scott, Sec. 7, p. 171

$$\mathcal{A} = \langle Q, A, E, S, I, T \rangle \ .$$

The set S, the set of triples $(p, 1_{A^*}, q)$, can also be seen as a subset of $Q \times Q$, hence as the *graph of a relation* on Q. The aforementioned equivalence takes the following form.

Proposition 1.17 *Every finite automaton with spontaneous transitions can be* effectively *transformed into an equivalent finite automaton without spontaneous transitions and with the same set of states.*

Proof. Let $\mathcal{A} = \langle Q, A, E, S, I, T \rangle$ be an automaton with spontaneous transitions. We transform $\mathcal{A}$ in two stages.

First, $\mathcal{A}$ is equivalent to another automaton with spontaneous transitions $\mathcal{A}' = \langle Q, A, E, S', I, T \rangle$ where S' has the property that if it contains $(p, 1_{A^*}, q)$ and $(q, 1_{A^*}, r)$, it also contains $(p, 1_{A^*}, r)$. In other words, S' is the graph of the transitive closure of (the relation on Q whose graph is) S and can be effectively constructed from S.

cf. Exer. 1.19, p. 73

[19] Nothing prevents us from further imagining an automaton whose transition labels are not words taken from A^* but elements from any set that has a product operation (and also, implicitly, addition). We shall do this in the following chapters.

[20] Most authors, who, as we have said, write the empty word ε, call them ε-transitions.

From $\mathcal{A}'$, we will construct an automaton $\mathcal{B} = \langle Q, A, F, J, T \rangle$ without spontaneous transitions. A computation c in $\mathcal{A}'$ is of the form:

$$c \ := \ p_0 \xrightarrow[\mathcal{A}']{x_1} p_1 \xrightarrow[\mathcal{A}']{x_2} p_2 \cdots p_{n-1} \xrightarrow[\mathcal{A}']{x_n} p_n \ ,$$

where the x_i are in $A \cup \{1_{A^*}\}$. We can assume, thanks to our assumptions about $\mathcal{A}'$, that two consecutive x_i are never equal to 1_{A^*}. The idea is to transform the computation c into an equivalent computation d in $\mathcal{B}$ replacing each sequence

$$p_{i-1} \xrightarrow[\mathcal{A}']{a} p_i \xrightarrow[\mathcal{A}']{1_{A^*}} p_{i+1} \qquad \text{with the transition} \qquad p_{i-1} \xrightarrow[\mathcal{B}]{a} p_{i+1}$$

and deleting the transition $p_0 \xrightarrow[\mathcal{A}']{x_1} p_1$ if $x_1 = 1_{A^*}$. For this, we define $\mathcal{B}$ as follows:

(i) a transition (p, a, r) is in F if and only if either it is in E, or there exists a state q such that there exist in $\mathcal{A}'$ a transition (p, a, q) and a spontaneous transition from q to r:

$$F = E \cup \{(p, a, r) \mid \exists q \in Q \quad (p, a, q) \in E \ \text{ and } \ (q, 1_{A^*}, r) \in S'\} \ ;$$

(ii) a state p is initial in $\mathcal{B}$ if it is either an initial state of $\mathcal{A}'$ or it is the destination of a spontaneous transition whose origin is an initial state of $\mathcal{A}'$:

$$J = I \cup \{q \mid \exists p \in I \quad (p, 1_{A^*}, q) \in S'\} \ .$$

The automaton $\mathcal{B}$ is equivalent to $\mathcal{A}'$ (and hence to $\mathcal{A}$). In fact, if c is a successful computation, the computation d is too since it has the same destination and, if it does not have the same source, then, by construction, p_1 is in J.

Conversely, for each successful computation of $\mathcal{B}$, one can find a corresponding successful computation of $\mathcal{A}'$ with the same label. ■

Definition 1.14 We will say that the automaton $\mathcal{B} = \langle Q, A, F, J, T \rangle$ constructed from $\mathcal{A} = \langle Q, A, E, S, I, T \rangle$ in the demonstration of Proposition 1.17 is obtained by *forward closure* of spontaneous transitions of $\mathcal{A}$, since the new transitions are obtained by considering the spontaneous transitions which *follow* the 'normal' transitions (labelled with letters).

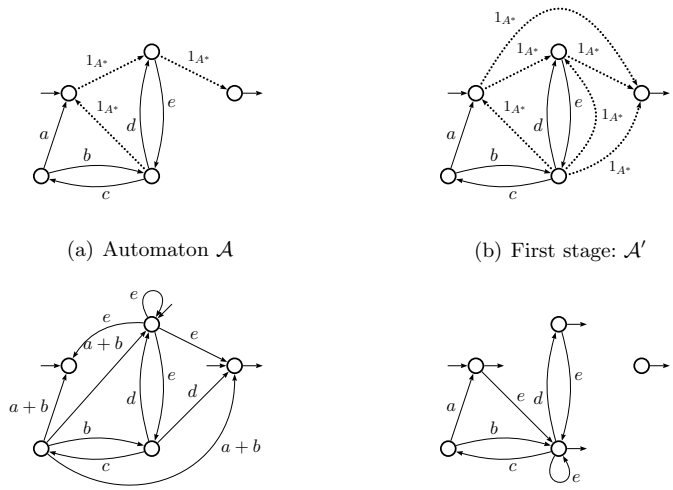
We could just as well obtain an equivalent automaton $\mathcal{C} = \langle Q, A, G, I, U \rangle$ by *backward closure* of spontaneous transitions of $\mathcal{A}$ with

$$G = E \cup \{(p, a, r) \mid \exists q \in Q \quad (p, 1_{A^*}, q) \in S' \ \text{ and } \ (q, a, r) \in E\}$$
$$\text{and} \qquad U = T \cup \{p \mid \exists q \in T \quad (p, 1_{A^*}, q) \in S'\} \ ;$$

that is to say, the new transitions of $\mathcal{C}$ are obtained by considering the spontaneous transitions which *precede* the transitions (labelled with letters) in $\mathcal{A}$. □

Figure 1.13 shows the two stages of the construction, for both the forward and backward closure. In Chapter III, we will give a more formal interpretation of these two constructions and of their relation to each other.

cf. Prop. III.2.14, p. 410



(a) Automaton $\mathcal{A}$

(b) First stage: $\mathcal{A}'$

(c) Forward closure: $\mathcal{B}$

(d) Backward closure: $\mathcal{C}$

Figure 1.13: Deletion of spontaneous transitions

Once we have established that their usage does not change the class of recognisable languages, spontaneous transitions are a versatile means of exhibiting properties of recognisable languages. For example, if $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is an automaton, the automaton obtained by adding to $\mathcal{A}$ all the spontaneous transitions $(p, 1_{A^*}, q)$, where (p, a, q) is a transition of $\mathcal{A}$, recognises the set of subwords of $L(\mathcal{A})$, from which we conclude:

Proposition 1.18 *The set of subwords of a recognisable language is recognisable.* ■

We will see many other examples of such constructions in the exercises and in the following sections.

cf. for example Sec. 5, p. 145

Exercises

• 1.26 What can be said about the union and the product

(a) of two trim automata? (b) of two unambiguous automata?

(c) of two complete automata?

1.27 Under what conditions is a complete automaton trim?

• 1.28 Verify that a trim automaton $\mathcal{A}$ is unambiguous if and only if for all f in $L(\mathcal{A})$, there exists a *unique* successful computation whose label is f.

• 1.29 What is the complexity of the procedure for deciding whether a trim automaton is unambiguous?

• 1.30 Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be a finite automaton and (p, q) and (r, s) two pairs of states (not necessarily distinct) of $\mathcal{A}$. Show directly (that is, without using the automaton $\mathcal{A} \times \mathcal{A}$ explicitly) that if there exists a word f in A^* which is the label of *two distinct computations*

in $\mathcal{A}$, one from p to q and the other from r to s, then there exists at least one such word of length less than $\|Q\|^2 + \|Q\|$.

Draw appropriate conclusions and comment.

• 1.31 Prove Property 1.7.

• 1.32 Prove Property 1.10.

• 1.33 Let $\mathcal{A}'$ and $\mathcal{A}''$ be two deterministic automata. Construct a deterministic automaton which recognises $L(\mathcal{A}') \cup L(\mathcal{A}'')$.

1.34 Verify that every automaton is equivalent to an automaton which has at most two distinct sink states.

1.35 Categorise (as deterministic, non-deterministic, ambiguous, unambiguous *etc.*) the au-
p. 59 tomata of Exercise 1.4.

1.36 Give an algorithm that transforms a finite automaton $\mathcal{A}$ where certain transitions are labelled with non-empty words from A^* into a classical finite automaton $\mathcal{B}$ over A. How many states does $\mathcal{B}$ have? Given a finite automaton $\mathcal{A}$ whose transitions are labelled with (possibly empty) words from A^*, in which order must this algorithm and that from the proof of Proposition 1.17 be applied in order to obtain a classical automaton with the fewest states possible?

• 1.37 Give a direct proof that the forward and backward closures of an automaton with spontaneous transitions are equivalent (independent from the proof that both are equivalent to the original automaton).

2 Rational languages

We will now show that the recognisable languages can also be defined with three operations – called 'rationals' – on the set of subsets of a free monoid. This is the substance of Kleene's Theorem, the central result of the theory.

In what follows, A is a non-empty alphabet.

2.1 Rational operations

We have already considered several operations on $\mathfrak{P}(A^*)$, the set of languages of A^*, such as Boolean operations and the additive extension of operations on words: transposition, quotient *etc.* We will now examine two operations derived from the monoid structure of the free monoid A^*.

2.1.1 Product of languages

The product of words extends to a product of languages, which is denoted, as in the
cf. Exer. 0.2.3 first case, by a dot, $X \cdot Y$, or by simple concatenation, XY:

$$X \cdot Y = \{f\,g \mid f \in X\,,\ g \in Y\}\ .$$

From this we obtain the definition of the *square* of a language X: $X^2 = X \cdot X$ and, by induction on n, of the nth power of X:

$$\forall n \in \mathbb{N} \qquad X^{n+1} = X^n \cdot X = X \cdot X^n \ , \tag{2.1}$$

a formula that is completed by setting

$$X^0 = 1_{A^*} \tag{2.2}$$

for all X in $\mathfrak{P}(A^*)$, *even for* $X = \emptyset$.

Examples 2.1 (i) If $X = \{a,\ ab,\ ba\}$ then

$$X^2 = \{aa,\ aab,\ aba,\ abab,\ abba,\ baa,\ baab,\ baba\} \ .$$

(ii) For every language X, $\emptyset \cdot X = X \cdot \emptyset = \emptyset$ and $1_{A^*} \cdot X = X \cdot 1_{A^*} = X$.

(iii) The set of words which have a prefix f is written $f\,A^*$, and the set of words which have a suffix g is written A^*g . The set of words that contain the factors ab or ba is written $A^*\{ab, ba\}A^*$.

(iv) The set of words of length n is A^n. □

Notation It will be convenient to write $X^{<n}$ (resp. $X^{\leqslant n}$) to denote the union of powers of X less than n (resp. less than or equal to n), for some integer n, by

$$X^{<n} = 1_{A^*} \cup X \cup X^2 \cup \cdots \cup X^{n-1} \ .$$

Thus, $a^{\leqslant 3} = \{1_{A^*}, a,\ aa,\ aaa\}$. Obviously, $X^{<0} = \emptyset$ for all X. The set of words of length less than or equal to n can therefore be written $A^{\leqslant n}$.

2.1.2 Star of a language

By combining the operations of union and product on $\mathfrak{P}(A^*)$, we will now construct a new operation which is one of the most important we shall meet in this work.

Definition 2.1 The *star* of a language X is the union of all the powers of X, and is written X^*:

$$X^* = \bigcup_{n \in \mathbb{N}} X^n \ .$$

The *proper star* of X, written X^+, is the union of all the *strictly positive* powers of X:

$$X^+ = \bigcup_{n=1}^{n=\infty} X^n \ . \qquad \square$$

The star of a word f, a language with a single element, is the set of all the powers of the word: $f^* = \bigcup_{n \in \mathbb{N}} f^n$. In particular,

$$(1_{A^*})^* = 1_{A^*} \qquad \text{and, from (2.2),} \qquad (\emptyset)^* = 1_{A^*} \ . \tag{2.3}$$

cf. p. 22 or p. 22

Example 2.2 With this notation we can rewrite the languages from Example 0.3.6 (or Examples 1.4 and 1.5) as

$$K_1 = \{a^n b^m \mid n, m \in \mathbb{N}\} = a^* b^* \ ,$$

and, since a word of even length is a product of digrams[21] and a word of odd length is the product of a letter and a word of even length,

$$L_2 = \{f \in A^* \mid |f| \equiv 1 \bmod 2\} = A\,(A^2)^* \ . \qquad \square$$

Why is the star operation important? We will first give an intuitive explanation; more formal approaches will follow.

We can take the view that there are in fact two sorts of union operation on the semiring $\mathfrak{P}(A^*)$: the union of two sets or, which comes to the same thing, the union of a finite number of sets; and the union of an arbitrary, possibly infinite, number of sets. With the latter operation we can build from its elements any subset of A^*, however complicated. The star operator falls between the two: it provides an infinite union, but restricts its application to the powers of a particular set. What is the 'power' of this operation? Or equivalently, what are the properties of the sets it can define? This is the question which we shall now answer.

cf. Note 0.10, p. 18

Remark 2.1 The notation X^* is somewhat improper because in this context X^* does not denote the free monoid generated by X but the set of products of elements of X. Nevertheless, if one takes $X = A$, the impropriety vanishes: the set of products of elements of A is indeed equal to the free monoid generated by A. $\square$

Remark 2.2 The star operation is sometimes called *iteration*. This reference to programming becomes clear if we consider the trivial program opposite where C represents a condition and f a sequence of statements. The program will execute the statements f^n if the condition C becomes false during the nth execution of the loop – or 'on the nth iteration'. Thus, the set of possible sequences of statements corresponding to this program (if it halts) is precisely f^*. $\square$

```
begin
while C do f od
end
```

An iterative program

2.1.3 Rational operations

The three *rational operations* on A^* (a mild abuse of terminology since they are in fact operations on $\mathfrak{P}(A^*)$) are:

(i) *union*: $X \cup Y$ (also written $X + Y$);
(ii) *product*: $X \cdot Y$ (also written $X\,Y$);
(iii) and *star*: X^*.

[21] Words of length 2.

By definition we have: $X^* = 1_{A^*} \cup X^+$. We verify that the product on $\mathfrak{P}(A^*)$ is (left and right) distributive over union:

$$X \cdot \left[\bigcup_{i\in I} Y_i\right] = \bigcup_{i\in I} [X \cdot Y_i] \qquad \text{and} \qquad \left[\bigcup_{i\in I} Y_i\right] \cdot X = \bigcup_{i\in I} [Y_i \cdot X] \tag{2.4}$$

even for infinite unions.[22] We deduce the following series of identities:

$$X^+ = \bigcup_{n=1}^{n=\infty} X \cdot X^{n-1} = \bigcup_{n=1}^{n=\infty} X^{n-1} \cdot X = X \cdot X^* = X^* \cdot X \quad ,$$

from which we conclude $X^* = 1_{A^*} \cup X \cdot X^* = 1_{A^*} \cup X^* \cdot X$, which is better written:

$$X^* = 1_{A^*} + X\,X^* = 1_{A^*} + X^*\,X \quad . \tag{2.5}$$

We will better appreciate the importance of these elementary identities as we continue.

Note that intersection and complement are *not* rational operations.

Exercises

• 2.1 Verify:

Property 2.1 $\forall X \subseteq A^* \qquad X^* = (X \setminus 1_{A^*})^*$.

• 2.2 **Prefix languages.** (a) Show that a non-empty language L over A^* is prefix if and only if for all disjoint languages H and K over A^*, $L\,H$ and $L\,K$ are disjoint.

cf. Sec. 0.3, p. 22 and Exer. 1.15, p. 65

(b) Show (recall that we write $f \leqslant g$ if f is a prefix of g):

Property 2.2 *Let K be a prefix language of A^*. Then:*

$$\forall f, f' \in K\,,\ \forall g, g' \in A^* \qquad f\,g \leqslant f'\,g' \quad \Longrightarrow \quad f = f' \qquad \textit{and} \qquad g \leqslant g' \quad .$$

(c) Show that a non-empty language L over A^* is prefix if and only if $K\,L$ is prefix for all non-empty prefix languages K of A^*.

• 2.3 **Prefix matrices.**[23] We call *prefix family of languages* a family of languages K_i such that:

(i) the K_i are pairwise disjoint;

(ii) the union of all the K_i is a prefix language.

We say that a matrix m with entries in $\mathfrak{P}(A^*)$ is prefix if the entries of each row of m form a prefix family. Show:

Property 2.3 *The product of two prefix matrices is prefix.*

---

[22] More precisely, union and product make $\mathfrak{P}(A^*)$ a semiring, indeed, a complete semiring (*cf.* Exer. 0.5.2, p. 29 and Sec. III.5.1, p. 470).

[23] This exercise is put here in anticipation of later developments, as a continuation of the previous exercise.

2.2 Rational languages

The choice of operations on $\mathfrak{P}(A^*)$ leads naturally to the following definitions.

Definitions 2.2
(i) A family of languages over A^* is *rationally closed* if it is closed under rational operations. That is, the family $\mathcal{F}$ is rationally closed if and only if

$$\forall X, Y \subseteq A^* \qquad X, Y \in \mathcal{F} \quad \Longrightarrow \quad X + Y\,,\ X\,Y\,,\ X^* \in \mathcal{F}\ .$$

(ii) The *rational closure* of a family $\mathcal{F}$ of languages over A^* is the *smallest* family of languages over A^* that contains $\mathcal{F}$ and which is rationally closed. It is written $\mathrm{Rat}\,\mathcal{F}$.
(iii) The *rational languages* (or *rational sets*) of A^* are the elements of *the rational closure of finite sets* of A^*. The family of rational languages over A^* is denoted $\mathrm{Rat}\,A^*$.
□

Since every finite set is either empty or a finite union of singleton sets, the family of rational languages is the rational closure of the empty set and the elements of A^*, which justifies the notation $\mathrm{Rat}\,A^*$. Since each non-empty word in A^* is a finite product of letters of A, $\mathrm{Rat}\,A^*$ is the rational closure of the empty set and the letters of A (as the empty word is the star of the empty set).[24]

Definition 2.2 obviously forces the union and product of two rational languages, and the star of a rational language, to be rational. For example, since $\{a\}$, $\{b\}$ and $\{a, b\}$ are finite subsets of $\{a, b\}^*$, the language

$$(\{a, b\}^*) \cdot \{a\} \cdot \{b\} \cdot (\{a, b\}^*)\ ,$$

which is better written

$$(a + b)^* a\, b\, (a + b)^*\ , \tag{2.6}$$

cf. Exam. 1.1, p. 55 is a rational language – the language of words that contain at least one factor $a\,b$: that is, $L(\mathcal{A}_1)$. Thus, every language obtained from the letters of A (and the empty word) by a succession of unions of products and of stars is a rational language, and this language can be written in a form like that of (2.6), which we will call a *rational expression*. At this point, and for our purposes in this section, this elementary definition suffices. The two preceding examples K_1 and L_2 become

$$K_1 = a^* b^* \qquad \text{and} \qquad L_2 = (a + b)\,[(a + b)\,(a + b)]^*$$

and are therefore rational languages. In Section 4, we will give a more precise (and *cf. Def. 4.1, p. 124* more advanced) definition of what constitutes a rational expression, which will allow us to make the notion a more useful tool.

Let us note immediately, however – to show the relevance of rational expressions – that the family of languages which can be obtained from a rational language is rationally closed, and that conversely, every rationally closed family which contains the letters of A (and the empty set) contains every language that can be obtained

[24] The notation $\mathrm{Rat}\,A$ would be no worse than $\mathrm{Rat}\,A^*$ but is not used because of the notation $\mathrm{Rat}\,M$ which we will see in the next chapter.

from a rational expression. Thus, the rational languages are exactly those obtained from a rational expression. *cf. Prop. 4.1, p. 126*

The notions of *rational languages* and *recognisable languages* are linked by the following fundamental result, known as Kleene's Theorem.

Theorem 2.1 [**Kleene**] *Let A be a finite alphabet. A language of A^* is rational if and only if it is recognised by a finite automaton over A: that is,*

$$\mathrm{Rat}\, A^* = \mathrm{Rec}\, A^* \quad .$$

In the next two subsections we will prove successively the two inclusions:

$$\mathrm{Rat}\, A^* \subseteq \mathrm{Rec}\, A^* \quad \text{and} \quad \mathrm{Rec}\, A^* \subseteq \mathrm{Rat}\, A^* \quad .$$

2.3 Rational is recognisable

Proposition 2.1 *The rational languages are recognisable: that is,*

$$\mathrm{Rat}\, A^* \subseteq \mathrm{Rec}\, A^* \quad . \tag{2.7}$$

Proof. We follow a method suggested by the definition of rational languages itself. We already know that $\mathrm{Rec}\, A^*$ is closed under union (Proposition 1.2). We will show that $\mathrm{Rec}\, A^*$ is closed under product (Proposition 2.4) and star (Proposition 2.5) and thus have proved:

Proposition 2.2 $\mathrm{Rec}\, A^*$ *is rationally closed.* ■

As we know moreover that the finite subsets of A^* are recognisable, Proposition 2.1 is proved, since $\mathrm{Rat}\, A^*$ is the smallest closed rational family that contains the finite subsets. ■ *cf. Cor. 1.3, p. 61*

To show that $\mathrm{Rec}\, A^*$ is closed under product and star, we will consider automata in special forms on which we can perform the product and star operations. With the first form, the *normal form*, the constructions are immediate but result in automata with many states. For this reason we shall repeat the same proof – which is therefore not just a proof of closure – on *standard form*, which will lead to much smaller automata. *cf. also Sec. 5, p. 145*

2.3.1 Normalised automata

Definition 2.3 An automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is *normalised* or *in normal form* if:

(i) the set of initial states is a singleton set $I = \{i\}$ which is not the destination of any transition of $\mathcal{A}$;

(ii) the set of final states is a singleton set $T = \{t\}$ which is not the source of any transition of $\mathcal{A}$;

(iii) i is distinct from t. □

A normalised automaton is therefore naturally represented by the schema in Figure 2.1. It follows directly from Definition 2.3 that the empty word cannot be recognised by a normalised automaton. It is notable that, apart from this one difference, normalised automata recognise the same languages as any other automata, which fact is expressed by the following result.

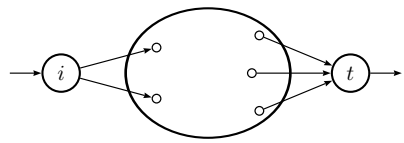


Figure 2.1: A normalised automaton

Lemma 2.3 *If L is a language recognised by a (finite) automaton over A, then $L \setminus 1_{A^*}$ is recognised by a (finite) normalised automaton.*

Proof. Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton that recognises L and $\{i, t\}$ be two states that *are not in* Q. Construct the automaton

$$\mathcal{A}_{\mathsf{nor}} = \langle Q \cup \{i, t\}, A, E \cup E', \{i\}, \{t\} \rangle \ ,$$

where E' is the set of transitions defined by

$$\begin{aligned} E' = \quad & \{(i, a, q) \mid \exists p \in I \quad (p, a, q) \in E\} \\ & \bigcup \{(p, a, t) \mid \exists q \in T \quad (p, a, q) \in E\} \\ & \bigcup \{(i, a, t) \mid \exists p \in I \, , \ \exists q \in T \quad (p, a, q) \in E\} \, . \end{aligned}$$

This construction is illustrated by Figure 2.2. We can easily verify that

$$L(\mathcal{A}_{\mathsf{nor}}) = L(\mathcal{A}) \setminus 1_{A^*} \ .$$

Indeed, let there be a successful computation in $\mathcal{A}$, of length at least 1:

$$c \ := \ p_0 \xrightarrow[\mathcal{A}]{a_1} p_1 \xrightarrow[\mathcal{A}]{a_2} p_2 \cdots p_{n-1} \xrightarrow[\mathcal{A}]{a_n} p_n \ . \tag{2.8}$$

Then it is in one-to-one correspondence with a successful computation in $\mathcal{A}_{\mathsf{nor}}$:

$$c_{\mathsf{nor}} \ := \ i \xrightarrow[\mathcal{A}_{\mathsf{nor}}]{a_1} p_1 \xrightarrow[\mathcal{A}_{\mathsf{nor}}]{a_2} p_2 \cdots p_{n-1} \xrightarrow[\mathcal{A}_{\mathsf{nor}}]{a_n} t \tag{2.9}$$

since, as c is successful, p_0 is in I and p_n is in T and, by construction, (i, a_1, p_1) and (p_{n-1}, a_n, t) are in E'.

Conversely, for each path such that (2.9), there is by construction at least one path such that (2.8). The correspondence also holds for paths of length 1 (that is the *raison d'être* of the third term in the equation that defines E') and the lemma is proved. ∎

Remark 2.3 An automaton need not be finite to be normalised. What is important is that the normalisation of a finite automaton be finite, and that is clearly the case since the normalisation merely adds two states. □ *cf. Exer. 2.5*

Remark 2.4 The construction of the normalised automaton implies that, even if $\mathcal{A}$ is deterministic or co-deterministic, $\mathcal{A}_{\mathsf{nor}}$ *is neither deterministic nor co-deterministic.* On the other hand, we can easily verify that $\mathcal{A}_{\mathsf{nor}}$ is unambiguous if and only if $\mathcal{A}$ is. □

The construction of $\mathcal{A}_{\mathsf{nor}}$ will be used in conjunction with the following definition. If L is a language (recognisable or not) of A^*, we will call the subset $\mathsf{c}\,(L) = L \cap \{1_{A^*}\}$ the *constant term*[25] of L: that is, $\mathsf{c}\,(L)$ is equal to 1_{A^*} if 1_{A^*} is in L, and to the empty set if not. It follows that

$$L = L \cap A^+ + \mathsf{c}\,(L) = (L \setminus 1_{A^*}) + \mathsf{c}\,(L) \quad . \tag{2.10}$$

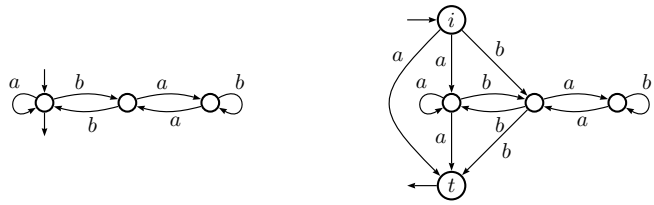
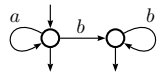
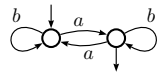


Figure 2.2: The automaton $\mathcal{P}_2$ and its normalisation

Exercises

- 2.4 Normalise the automata: (a) (b)  (c) 

- 2.5 Normalise the automaton $\mathcal{Z}_1$ (*cf.* Example 1.3, p. 52).

- 2.6 Construct a normalised automaton with three states that recognises:

  (a) a^*b ; (b) bb^*a .

- 2.7 Give a procedure for the construction of $\mathcal{A}_{\mathsf{nor}}$ from $\mathcal{A}$ using spontaneous transitions and the construction used in the proof of Lemma 2.3.

- 2.8 Construct a finite automaton that accepts normalised binary representations: that is, without leading zeros, of numbers divisible by 3.

[25] We will justify the terminology in Chapter III. It would be more accurate to define $\mathsf{c}\,(L)$ as a boolean variable which is 1 or 0 according to whether L contains 1_{A^*}, and (2.10) would be written $L = (L \setminus 1_{A^*}) + \mathsf{c}\,(L)\, 1_{A^*}$. I have simplified at the cost of a slight inexactitude.

2.3.2 Closure under product and star

Proposition 2.4 *The product of two recognisable languages is recognisable.*

Proof. If L and K are two languages over A^*, then by direct application of (2.10)

$$L\,K = (L \setminus 1_{A^*})(K \setminus 1_{A^*}) + \mathsf{c}\,(L)\;(K \setminus 1_{A^*}) + (L \setminus 1_{A^*})\;\mathsf{c}\,(K) + \mathsf{c}\,(L)\;\mathsf{c}\,(K) \quad .$$

Since recognisable languages are closed under union, we need only prove the proposition for recognisable languages without the empty word: that is, by Lemma 2.3, for languages recognised by normalised automata.

Let therefore $\mathcal{A}$ and $\mathcal{B}$ be two normalised automata which recognise L and K respectively. Then, the automaton $\mathcal{C}$ obtained by identifying the final state of $\mathcal{A}$ with the initial state of $\mathcal{B}$ (*cf.* Figure 2.3) is a (normalised) automaton which recognises the product $L\,K$. Indeed, a successful computation in $\mathcal{C}$ is *necessarily* of the form

$$i \xrightarrow[\mathcal{A}]{f} t = j \xrightarrow[\mathcal{B}]{g} u \quad , \tag{2.11}$$

and its label is $f\,g$, where f is in L since $i \xrightarrow[\mathcal{A}]{f} t$ is a successful computation of $\mathcal{A}$, and g is in K since $j \xrightarrow[\mathcal{B}]{g} u$ is a successful computation of $\mathcal{B}$. Also, $L(\mathcal{C})$ is a subset of $L\,K$. Conversely, every product $f\,g$ of words of L and K is the label of a computation of the same form as (2.11). ■

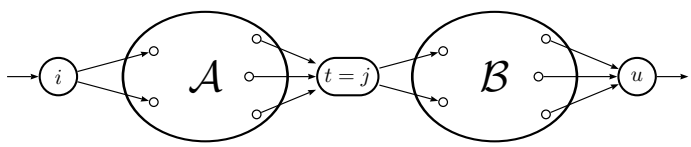
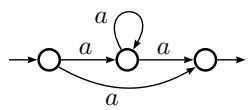


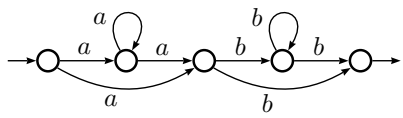
Figure 2.3: Automaton for the product of two languages

Example 2.3 We can construct, using the above procedure, an automaton which recognises the language $K_1 = a^*b^*$, the product of $L_3 = a^*$ and $K_3 = b^*$. We have seen that the normalised automaton which recognises $L_3 \setminus 1_{A^*} = a^+$ is

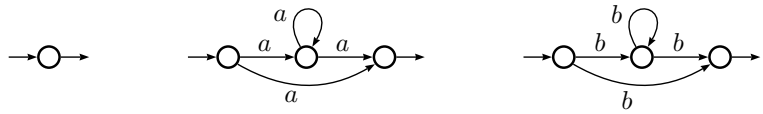
cf. Exer. 2.4 (a)



whence the automaton which recognises the product $(L_3 \setminus 1_{A^*})(K_3 \setminus 1_{A^*}) = a^+b^+$:

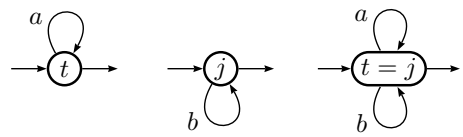


to which must be added the three automata



to recognise the whole of K_1. □

We see already in this construction the reason for the assumptions made by the definition of normalised automata: if there were a transition from t and therefore – since we can assume that $\mathcal{A}$ is trim – a circuit including t labelled f, and similarly, if there were a circuit including j labelled g, we could in some computation of $\mathcal{C}$ read g, a prefix of K, then f, a suffix of L, to obtain a word which is not in $L\,K$. An extreme case is shown in the adjacent figure, where the construction applied to the natural automata that recognise a^* and b^* gives an automaton which recognises $\{a, b\}^*$ rather than a^*b^* .



Proposition 2.5 *The star of a recognisable language is recognisable.*

Proof. Let L be a language of A^* (recognisable or not). Since we evidently have the identity *cf. Propty 2.1, p. 85*

$$L^* = (L \setminus 1_{A^*})^* \ ,$$

we need only prove the proposition for recognisable languages without the empty word, in other words, by Lemma 2.3, for languages recognised by normalised automata. Let $\mathcal{A}$ be a normalised automaton which recognises L. Then the automaton $\mathcal{D}$ obtained by identifying the initial and final states of $\mathcal{A}$ (*cf.* Figure 2.4) is a (non-normalised) automaton which recognises L^*.

In fact, a successful computation in $\mathcal{D}$ is *necessarily* of the form:

$$i \xrightarrow[\mathcal{A}]{f_1} i \xrightarrow[\mathcal{A}]{f_2} i \cdots i \xrightarrow[\mathcal{A}]{f_n} i \qquad (2.12)$$

and its label is $f_1 f_2 \cdots f_n$, where each f_k is in L since

$$i \xrightarrow[\mathcal{A}]{f_k} t = i$$

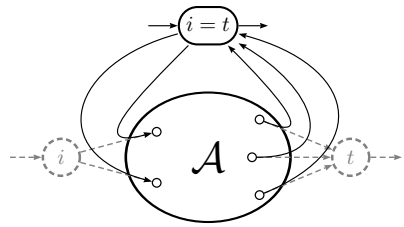


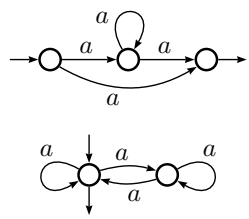
Fig. 2.4: Automaton for the star of a language

is a successful computation in $\mathcal{A}$. Hence, $L(\mathcal{D})$ is a subset of L^*. Conversely, every product $f_1 f_2 \cdots f_n$ of words in L is the label of a path such that (2.12) holds; this includes the product of zero words, which is equal to the empty word. ■

As we have already said, this demonstration completes the proof of Proposition 2.2 and hence of Proposition 2.1 (Equation (2.7)). Not only have we attained our objective but the notion of a normalised automaton is useful and will be of service again. But this proof does not just show that $\mathrm{Rec}\, A^*$ is rationally closed. It can also be seen as a way to construct, from a rational expression – that is, a sequence of unions, products and stars on letters – an automaton that recognises the language obtained by this expression. From this point of view, it is not a good proof. The two simple examples

above and below show that the automata obtained by the method induced from this proof are bigger than might reasonably be expected. *Standard form*, which is defined in §2.3.3, will make up for this failing.

Example 2.4 According to the procedure above, we construct an automaton that recognises $(a^*)^*$, the star of the language $L = a^*$, even if in this case $L^* = L$. From the normalised automaton which recognises $L \setminus 1_{a^*} = a^+$ (which we have already computed twice) – adjacent figure, top – we deduce the automaton which recognises $(a^+)^*$ – adjacent, bottom. □



Exercises

• 2.9 Construct, using the methods just introduced, an automaton that recognises:

(a) $(a^*b^*)^*$ from the normalised automaton computed in Exercise 2.4 (b) (not from that in Example 2.3!).

(b) $((a^+)^+)^*$ from the automaton computed in Example 2.4.

(c) $(a^*b + b\,b^*a)^*$ from the normalised automata computed in Exercise 2.6.

2.10 Verify the assertion of Remark 2.4: that $\mathcal{A}_{\mathsf{nor}}$ is unambiguous if and only if $\mathcal{A}$ is.

2.11 We will say that *the product $L\,K$ is unambiguous* if $f\,g = f'g'$, where $f, f' \in L$, $g, g' \in K$, implies $f = f'$ and $g = g'$. Let $\mathcal{A}$ and $\mathcal{B}$ be unambiguous automata. Show that the automaton constructed in the proof of Proposition 2.4 from $\mathcal{A}$ and $\mathcal{B}$ is unambiguous if and only if the product $L(\mathcal{A})\,L(\mathcal{B})$ is.

• 2.12 **Conjugate of a language.** Let $\mathsf{Conj}\colon A^* \to A^*$ be the relation which relates a word f to the set of its *conjugates*:

$$\mathsf{Conj}(f) = \{v\,u \mid u, v \in A^* \quad u\,v = f\} \ .$$

The relation extends by additivity: $\mathsf{Conj}(L) = \bigcup_{f \in L} \mathsf{Conj}(f)$. Show that $\mathsf{Conj}(L)$ is recognisable when L is.

2.3.3 Standard automata

Definition 2.4 An automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is *standard*, or *in standard form*, if the set of initial states is a singleton set $I = \{i\}$ which is not the destination of any transition of $\mathcal{A}$. □

A standard automaton is therefore naturally represented by the schema in Figure 2.5. Note that the definition does not forbid the initial state i from also being final – the case shown in the figure.

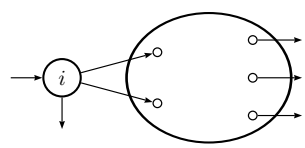


Fig. 2.5: A standard automaton

Lemma 2.6 *Every (finite) automaton is equivalent to a (finite) standard automaton.*

Proof. Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton and $\{i\}$ a state which is *not* in Q. We first construct an automaton $\mathcal{A}'$ by adding to $\mathcal{A}$ a spontaneous transition from i to each initial state j of $\mathcal{A}$, and by making the initial states of $\mathcal{A}$ non-initial. This automaton $\mathcal{A}'$ is equivalent to $\mathcal{A}$, since, if its spontaneous transitions are removed by forward closure and the result is trimmed, the result is $\mathcal{A}$. We thus remove these spontaneous transitions from $\mathcal{A}'$ by backward closure. The automaton thus obtained is standard and equivalent to $\mathcal{A}$. ■ *cf. Prop. 1.17, p. 79*

This construction is shown in Figure 2.6. The 'standardisation' of an automaton does not require that it be finite, but a standardised finite automaton is clearly so.

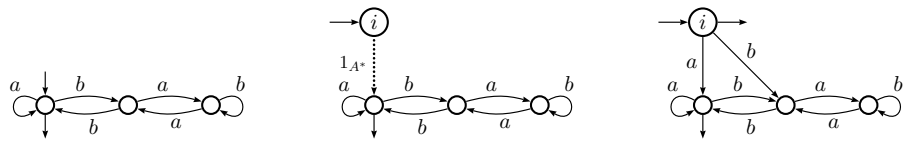


Figure 2.6: Standardisation of the automaton $\mathcal{P}_2$

Like normalised automata, standardised automata allow the construction of automata for the product and star (of recognised languages). Let $\mathcal{A}$ and $\mathcal{B}$ be two standard automata that recognise respectively L and K. We obtain a (standard) automaton $\mathcal{C}$ that recognises the language $L\,K$ in three steps: (i) we add a spontaneous transition from each final state of $\mathcal{A}$ to the unique initial state of $\mathcal{B}$; (ii) we make the final states of $\mathcal{A}$ non-final and the initial state of $\mathcal{B}$ non-initial; and (iii) we remove spontaneous transitions (by backward closure). This construction is shown in Figure 2.7.

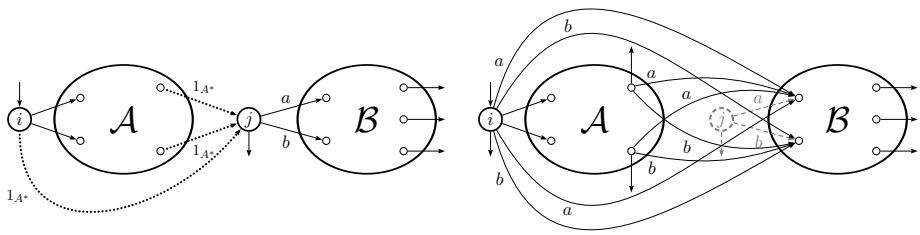


Figure 2.7: Concatenation of two standard automata
In this figure the two initial states are also final, which gives a richer construction.

Similarly, we obtain a (standard) automaton $\mathcal{D}$ which recognises the language L^* in three steps: (i) we add a spontaneous transition to each final state of $\mathcal{A}$ to the unique initial state of $\mathcal{A}$; (ii) we make the final states of $\mathcal{A}$ non-final, and make the initial state of $\mathcal{A}$ final; and (iii) we remove spontaneous transitions (by backward closure). This construction is shown in Figure 2.8.

These two constructions are another proof of Proposition 2.2 and therefore of Proposition 2.1. But, as with the constructions associated with normalised automata, they can also be seen as a method for constructing an automaton from a rational expression, to which aspect we shall return after defining rational expressions with more care, and more formally.

cf. Sec. 4 and 5

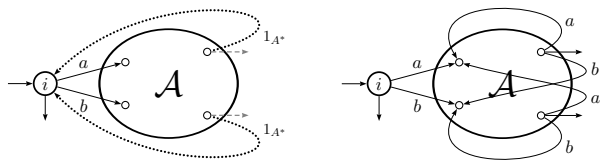
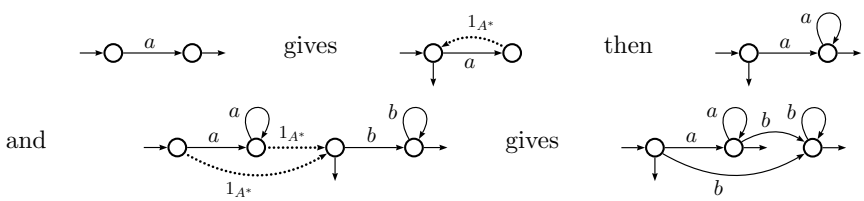


Figure 2.8: Star of a standard automaton

Example 2.5 A standard automaton which recognises the language $K_1 = a^*b^*$ is obtained by concatenating the standard automata that recognise a^* and b^*, which are themselves obtained by taking the star of the two automata that recognise a and b. In detail,



which is the automaton we want. The comparison with the automaton calculated in Example 2.3 should be obvious. □

cf. p. 90

Exercises

- 2.13 Give the standard automata which correspond to the following rational expressions: (a) $(a+b)^*$; (b) $(a^*b^*)^*$; (c) $(a^*b + b\,b^*a)^*$.

2.14 Give an example which shows that the above constructions are invalid when applied to non-standard automata.

2.4 Recognisable is rational

Th. 2.1, p. 87

To complete the proof of Kleene's Theorem it remains to show that

$$\mathrm{Rec}\, A^* \subseteq \mathrm{Rat}\, A^* \quad , \tag{2.13}$$

or more explicitly:

Proposition 2.7 *The language recognised by a finite automaton is rational.*

We will give several proofs of this fundamental statement, starting with one of the best known, called the McNaughton–Yamada algorithm, which describes a procedure for showing that the language recognised by a finite automaton can be obtained

from the letters of the alphabet by a series of unions, products and stars: that is, by a rational expression. The algorithm also makes clear the relationship between this procedure and the calculation of the transitive closure of a graph[26] by the Roy–Warshall algorithm. *cf. Exer. 1.19, p. 73*

The second method, due to Brzozowski and McCluskey, and generally known as the *state elimination method*, has the advantage of leading to the simplest computations, both for the small examples which we will compute by hand, and for implementation as a program. However, it requires that we define a more general model of automata than that of finite automata over an alphabet. A third method is based on the solution of a system of linear equations, and is in fact a more mathematical presentation of the second method.

Later in this chapter and the two that follow, we will return several more times to the result and its proof.

2.4.1 The McNaughton–Yamada algorithm, or MNY algorithm

Proof of Proposition 2.7. Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton; write $R_{p,q}$ for the set of labels of computations (in $\mathcal{A}$) whose source is p and destination q:

$$R_{p,q} = \{f \in A^* \mid p \xrightarrow[\mathcal{A}]{f} q\} \ .$$

Since $L(\mathcal{A}) = \bigcup_{p \in I, q \in T} R_{p,q}$, that is, $L(\mathcal{A})$ is a finite union of $R_{p,q}$, it is sufficient to show that each $R_{p,q}$ is rational.

The central idea of the proof is to make a *total order* on the elements of Q, from 1 to $n = \|Q\|$. We therefore define for each integer k, $0 \leqslant k \leqslant n$, and for all p and q in Q, the set $R^{(k)}_{p,q}$ of labels of computations from p to q:

(i) which are *non-trivial* (that is, of length at least 1);

(ii) and which *do not pass through intermediate states of rank greater than* k.

A computation that does not pass through any intermediate state of rank greater than 0 passes through no intermediate states, and therefore reduces to a single transition. Thus, $R^{(0)}_{p,q}$ is, for all p and q in Q, the set of labels of transitions which go from p to q: that is, $R^{(0)}_{p,q} = E_{p,q}$ if E is viewed as the matrix of transitions of $\mathcal{A}$. *cf. Def. 1.3, p. 54* Hence $R^{(0)}_{p,q}$ is a subset of A, and hence rational *since A is assumed to be finite.*

We notice next that a computation which goes from p to q without visiting intermediate states of rank greater than $k+1$ is:

(a) either a computation (from p to q) which does not visit intermediate states of rank greater than k;

(b) or the concatenation:

- of a computation from p to $k+1$ without passing through states of rank greater than k;
- followed by an arbitrary number of computations which go from $k+1$ to $k+1$ without passing through intermediate states of rank greater than k;

[26] And with Floyd's algorithm for the computation of 'shortest paths' in a valued graph.

- followed finally by a computation from $k+1$ to q without passing through intermediate states of rank greater than k.

This is shown in Figure 2.9 and corresponds to the fact that for all p and q in Q, for all $k < n$, we have:

$$R_{p,q}^{(k+1)} = R_{p,q}^{(k)} + R_{p,k+1}^{(k)} \left[R_{k+1,k+1}^{(k)}\right]^* R_{k+1,q}^{(k)} \quad , \quad (2.14)$$

which shows that each $R_{p,q}^{(k+1)}$ is rational if the $R_{p,q}^{(k)}$ are. We obviously have

$$R_{p,q} = \begin{cases} R_{p,q}^{(n)} & \text{if} \quad p \neq q \\ R_{p,q}^{(n)} + 1_{A^*} & \text{if} \quad p = q \end{cases}$$

which completes the proof of the proposition. ■

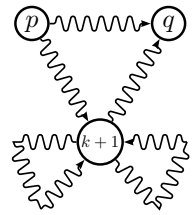
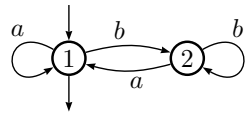


Fig. 2.9: Step $k+1$ of MNY algorithm

Example 2.6 Let us apply the MNY algorithm to the automaton $\mathcal{R}_1$ in the adjacent figure. For conciseness we shall regroup, for each k, the four $R_{p,q}^{(k)}$ into a matrix $R^{(k)}$.[27] We obtain

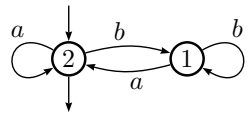


$$R^{(0)} = \begin{pmatrix} a & b \\ a & b \end{pmatrix} \quad , \qquad R^{(1)} = \begin{pmatrix} a + a(a)^*a & b + a(a)^*b \\ a + a(a)^*a & b + a(a)^*b \end{pmatrix} = \begin{pmatrix} a^+ & a^*b \\ a^+ & a^*b \end{pmatrix}$$

$$R^{(2)} = \begin{pmatrix} a^+ + a^*b(a^*b)^*a^+ & a^*b + a^*b(a^*b)^*a^*b \\ a^+ + a^*b(a^*b)^*a^+ & a^*b + a^*b(a^*b)^*a^*b \end{pmatrix} = \begin{pmatrix} (a^*b)^*a^+ & (a^*b)^+ \\ (a^*b)^*a^+ & (a^*b)^+ \end{pmatrix} \quad ,$$ [28]

$$\text{and} \quad R = \begin{pmatrix} 1_{A^*} + (a^*b)^*a^+ & (a^*b)^+ \\ (a^*b)^*a^+ & (a^*b)^* \end{pmatrix} \quad , \quad \text{or} \quad L(\mathcal{R}_1) = R_{1,1} = 1_{A^*} + (a^*b)^*a^+ \quad .$$

If we consider the other order on the states of $\mathcal{R}_1$, we have:



$$R^{(1)} = \begin{pmatrix} b^*a & b^+ \\ b^*a & b^+ \end{pmatrix} \quad , \quad R^{(2)} = \begin{pmatrix} (b^*a)^+ & (b^*a)^*b^+ \\ (b^*a)^+ & (b^*a)^*b^+ \end{pmatrix}$$

$$\text{whence} \quad R = \begin{pmatrix} (b^*a)^* & (b^*a)^*b^+ \\ (b^*a)^+ & (b^*a)^*b^* \end{pmatrix} \quad \text{and} \quad L(\mathcal{R}_1) = (b^*a)^* \quad .$$ □

Note that to make it easier to write the matrices $R^{(k)}$ we have made liberal use of the identities (2.4) and (2.5). Note also that these computations demonstrate that the same language can be obtained from different rational expressions. We shall return to these questions in Section 4.

cf. § 2.1.3, p. 85

[27] The word 'matrix' is taken here in the sense of a square table of languages indexed by $Q \times Q$, with no other structure or property.

[28] In order to keep the size of the entries of the computed matrices reasonable, we have used freely the identities (2.5) (see p. 85). This does not change the accepted languages but is more questionable when it comes to symbolic computation. We shall be more precise in Subsection 4.3.

2.4.2 The state elimination method, or BMC algorithm

The description of this algorithm uses a generalisation of the notion of automaton, leading us to anticipate some things that we shall study more systematically in the next two chapters.

A *generalised*[29] automaton is an automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ in which the transition labels are not just letters of the alphabet A but *subsets* of A^*: that is, the elements of E are triples (p, L, q) with p and q in Q and $L \subseteq A^*$. The label of a computation is, as before, the product of the labels of the transitions which form the computation. The language recognised, or accepted, by $\mathcal{A}$, still written $L(\mathcal{A})$, is the *union* of the labels of the successful computations of $\mathcal{A}$.

cf. p. 227

Note that, even if it means taking some unions, we can always assume that there is, between each pair of states of a generalised automaton, *at most one transition*, without changing the language recognised by the automaton. Note also that an automaton with spontaneous transitions is a special case of a generalised automaton.

A generalised automaton is *finite* if the set of transitions, and hence that of useful states, is finite. Starting with a generalised automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ with Q finite, we canonically construct a finite generalised automaton $\mathcal{B}$ – obviously equivalent to $\mathcal{A}$ – by setting the label of the single transition from p to q in $\mathcal{B}$ to the (finite) union of the labels of all the transitions from p to q in $\mathcal{A}$. Having made these definitions, we prove the following proposition, of which Proposition 2.7 is evidently a special case.

cf. also Th. II.1.2, p. 227

Proposition 2.8 *The language recognised by a finite generalised automaton $\mathcal{A}$ is a member of the rational closure of the family of labels of transitions of $\mathcal{A}$.*

Proof. The proof consists of the description of an algorithm which, for a finite generalised automaton $\mathcal{A}$, computes a corresponding generalised automaton $\mathcal{C}$ that can be described as trivial: it has a unique initial state i, a unique final state t (distinct from i) and a unique transition from i to t labelled by a language L (*cf.* Figure 2.10). This L is both equal to $L(\mathcal{A})$ and is obtained by a series of (finite) unions of products, pairs and stars on the transition labels of $\mathcal{A}$.

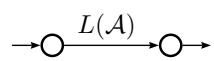


Figure 2.10: The result of the state elimination method

The first phase involves constructing a normalised generalised automaton $\mathcal{B}$ by adding to $\mathcal{A} = \langle Q, A, E, I, T \rangle$ two states i and t, distinct and not belonging to Q, a transition labelled 1_{A^*} from i to each initial state of $\mathcal{A}$, and a transition labelled by 1_{A^*} from each final state of $\mathcal{A}$ to t. The state i is the unique initial state of $\mathcal{B}$, and t its unique final state. Clearly, $\mathcal{B}$ is equivalent to $\mathcal{A}$. Since $\mathcal{A}$, and hence $\mathcal{B}$, are finite, we can assume – even if it means taking some finite unions of the transitions

cf. the transformation from Figure 2.12 to Figure 2.13 (a)

[29] *N.B.*, this is not the same as Eilenberg's definition of a generalised automaton, *op. cit.*

labels – that there is at most one transition with source p and destination q for each pair (p, q) of states in $\mathcal{B}$.

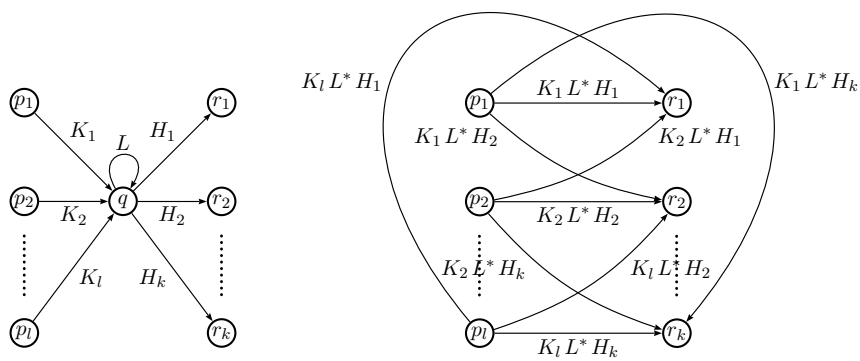
The second phase comprises as many steps as $\mathcal{A}$ has states. It consists of successively removing from $\mathcal{B}$ the elements of Q and of modifying the transitions such that at each step an automaton equivalent to that of the previous step is obtained whose labels are derived from the preceding automaton by unions, products and stars.

More precisely, let q be a particular state in Q; let $p_1, p_2, \ldots, p_l$ be those states of $\mathcal{B}$, distinct from q, that are the sources of transitions whose destination is q, and K_1, K_2, $\ldots$, K_l the labels of these transitions; let r_1, r_2, $\ldots$, r_k be those states of $\mathcal{B}$, distinct from q, that are the destinations of transitions whose source is q, and H_1, H_2, $\ldots$, H_k the labels of these transitions. Note that some r_j may coincide with some p_h. Let L be the label of the transition with source and destination q, if it exists; if not, set $L = \emptyset$ (and we therefore have $L^* = 1_{A^*}$).

cf. Equat. (2.3), p. 83

Let $\mathcal{B}'$ be the automaton obtained from $\mathcal{B}$ by first removing the state q and all the transitions that touch q described above, and second by adding, for each pair of states (p_h, r_j), $1 \leqslant h \leqslant l$ and $1 \leqslant j \leqslant k$, the transition $(p_h, K_h\, L^*\, H_j, r_j)$ (*cf.* Figure 2.11).

To each computation c in $\mathcal{B}$, there corresponds a computation c' in $\mathcal{B}'$ whose label contains that of c. Conversely, every word f in A^* belonging to the label of some computation d' in $\mathcal{B}'$ can be factorised according to the labels of the transitions of that computation, and the factorisation allows us to find a computation d in $\mathcal{B}$ whose label contains f.



(a) before the removal of state q (b) after the removal of state q

Figure 2.11: One step of the state elimination method

Thus, $\mathcal{B}$ and $\mathcal{B}'$ are equivalent. For each pair (p, r) of states in $\mathcal{B}'$, we now have at most two transitions, which we can combine into one with the union of two labels if need be. After repeating this construction n times ($n = \|Q\|$), we obtain an automaton $\mathcal{C}$ which contains none of the states of the automaton $\mathcal{A}$ with which we started, and which is therefore of the required form. ■

Example 2.6 (continued) Let us apply the method we have just described to the two-state automaton $\mathcal{R}_1$ shown in Figure 2.12. In the series of automata in Figure 2.13, we removed state q first, then state p; in the series of automata in Figure 2.14, state p was removed first, then q.[30] □

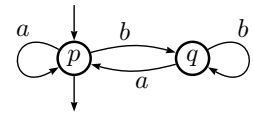


Fig. 2.12: The automaton $\mathcal{R}_1$

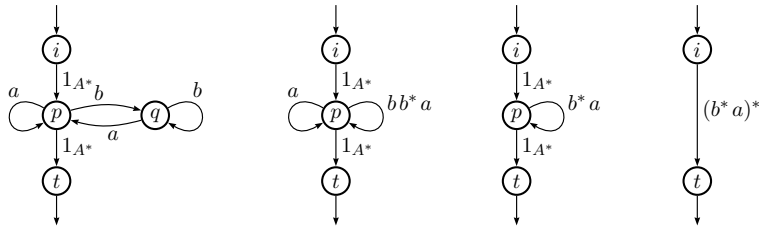


Figure 2.13: The state elimination method applied to $\mathcal{R}_1$, removing q, then p.

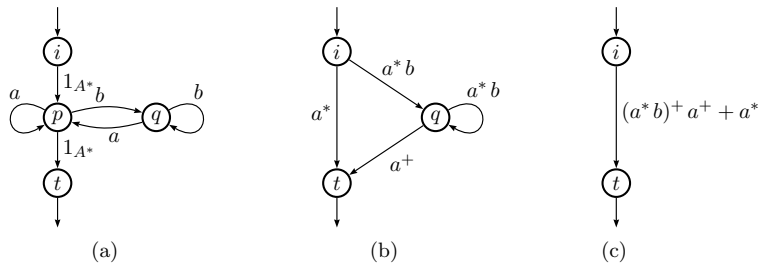


Figure 2.14: The state elimination method applied to $\mathcal{R}_1$, removing p, then q.

We also notice, as in the preceding section, that the same language can be obtained from different rational expressions. *cf. Sec. 4.3, p. 133*

2.4.3 Equation solution method

We can give a more abstract interpretation to the state elimination method by showing that it is an algorithm for solving a system of linear equations...by *elimination* of unknowns, or Gaussian *elimination*. Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be a particular finite automaton. For all p in Q, we will write L_p for the set of words that are the label of a computation from p to a final state of $\mathcal{A}$: *what we will call the* future *of p in $\mathcal{A}$ cf. Sec. II.4, p. 273*

$$L_p = \{f \mid \exists t \in T \quad p \xrightarrow[\mathcal{A}]{f} t\} \ . \tag{2.15}$$

We obviously have

$$L(\mathcal{A}) = \bigcup_{p \in I} L_p \ ,$$

[30] Here again, we have made use of the identities (2.5) for the labels of transitions. *Cf.* Subsection 4.3.

and it is therefore enough to show that each of the L_p is rational. A word f in A^* belongs to L_p if:

(a) either it is the empty word and p is in T;

(b) or it is written $f = a\,g$, where a is the label of a transition of $\mathcal{A}$ from p to q, and g is in L_q.

We can therefore write

$$\forall p \in Q \qquad L_p = \bigcup_{q \in Q} E_{p,q}\, L_q + \delta_{p,T} \ , \tag{2.16}$$

where $E_{p,q}$ is the set of labels of the transitions which go from p to q and where $\delta_{p,T}$ is 1_{A^*} if p is in T and $\emptyset$ otherwise.

We therefore obtain a system of n linear equations in n unknowns, where $n = \|Q\|$, the number of states in $\mathcal{A}$, the n unknowns are the L_p, and the coefficients are in $\mathfrak{P}(A^*)$: indeed in this case, in $\mathfrak{P}(A)$. We can solve this system by successive elimination of unknowns using the following result, known as Arden's Lemma.

Lemma 2.9 [**Arden**] *Let K and L be two subsets of A^*, where K does not contain 1_{A^*}. Then K^*L is the* unique solution *of*

$$X = K\,X + L \ . \tag{2.17}$$

Proof. It is obvious that K^*L is a solution of (2.17). Starting with $K^* = 1_{A^*} + K\,K^*$ we multiply the two sides by L (on the right) to obtain:

cf. Equat. (2.5), p. 85

$$(K^*L) = L + K\,(K^*L) \ .$$

Conversely, let P be a solution of (2.17), that is,

$$P = K\,P + L \qquad \text{from which} \qquad L \subseteq P \quad \text{and} \quad K\,P \subseteq P \ . \tag{2.18}$$

From (2.18), we conclude that $K\,L \subseteq P$, since

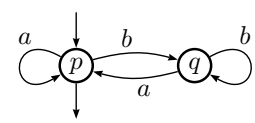
$$L \subseteq P \quad \Longrightarrow \quad K\,L \subseteq K\,P \subseteq P \ ,$$

then, by induction, that $K^n L \subseteq P$ for all n in $\mathbb{N}$, and hence also that

$$K^*L \subseteq P \ . \tag{2.19}$$

We now show the reverse inclusion $P \subseteq K^*L$ by *reductio ad absurdum*: let f be the shortest word in the set $P \setminus K^*L$, which we assume is non-empty. By hypothesis, f is not in L and since f is in P, we have f is in $K\,P$ by (2.18) and can write $f = k\,g$ with k in K and g in P. By the hypothesis of the lemma, $k \neq 1_{A^*}$, hence g, strictly shorter than f, is in K^*L, and hence also f, which is a contradiction. ■

Example 2.6 (continued) Let us write the equations corresponding to the automaton $\mathcal{R}_1$:



$$L_p = a\,L_p + b\,L_q + 1_{A^*} \tag{2.20}$$

$$L_q = a\,L_p + b\,L_q \tag{2.21}$$

From (2.21), we deduce, by Lemma 2.9, that $L_q = b^* a\, L_p$, which we substitute into (2.20):

$$L_p = a\, L_p + b^+ a\, L_p + 1_{A^*} = b^* a\, L_p + 1_{A^*} \,.$$

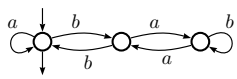
This gives $L(\mathcal{R}_1) = L_p = (b^* a)^*$. □

Remark 2.5 The symmetric equation, $X = X\,K + L$, still assuming that K does not contain 1_{A^*}, has $L\,K^*$ as its unique solution. Furthermore, if K contains 1_{A^*}, A^* is a solution of (2.17) whatever the value of L. The language $K^* L$ is therefore no longer a unique solution; by (2.19), it is the *smallest* solution of (2.17). □

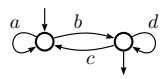
To finish this section, recall that the inclusions (2.7) and (2.13) together prove Theorem 2.1. ■

Exercises Compute, by different methods, the language recognised by the following automata:

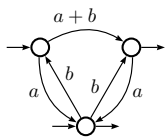
• 2.15



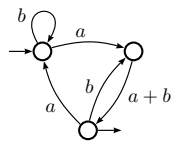
• 2.16



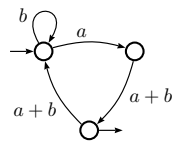
• 2.17



2.18



• 2.19



• 2.20 Give the general form of solutions of $X = K\,X + L$ when $1_{A^*} \in K$.

3 The functional perspective

We return to the study of automata – the definition of rational languages and Kleene's Theorem were almost an interlude, which I did not want to put off any longer – with a different point of view, a functional and *dynamic* point of view – we will say that a letter 'sends' a state to a set of states, as opposed to the *static* set point of view – which prevailed in Section 1, where we considered the set of transitions of an automaton. In this context, the definition of deterministic automata seems natural: a letter 'sends' a state to another state, unique and well defined, and gives a believable model of 'real machines', of physically realisable computation devices.

It is therefore essential that every (finite) automaton be equivalent to a (finite) deterministic automaton, and that we finish the proof that the family of recognisable languages – and hence that of rational languages – is a Boolean algebra, the key to showing that the equivalence of finite automata is decidable.

In fact, deterministic automata are a different sort of structure from general automata. We will understand this distinction better, and develop it, in the next chapter; in this section we will show the existence of a unique minimal deterministic automaton, canonical and computable, associated with each recognisable language.

3.1 From transitions to the transition function

Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton. The set E of transitions, a subset of $Q \times A \times Q$, can be seen as the graph of a relation δ from $Q \times A$ to Q, or, which is the same thing, as the graph of a map δ from $Q \times A$ to $\mathfrak{P}(Q)$:

$$\forall (p,a) \in Q \times A \qquad (p,a)\delta = \{q \in Q \mid (p,a,q) \in E\} \ . \tag{3.1}$$

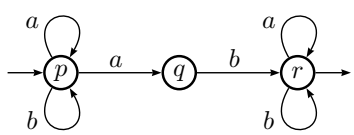
Conversely, every map δ from $Q \times A$ to $\mathfrak{P}(Q)$ defines a set of transitions of an automaton over A whose set of states is Q:

$$E = \{(p,a,q) \mid q \in (p,a)\delta\} \ .$$

We call δ the *transition function* of $\mathcal{A}$. It is convenient, if $\mathcal{A}$ is finite, to represent this function as a table whose columns are indexed by the elements of Q, and the rows by those of A, and where the value of the entry at column p and row a is $(p,a)\delta$.

The automaton $\mathcal{A}$ is completely described by this table, which is called the *transition table of* $\mathcal{A}$, if we mark those columns that correspond to initial or final states, which we do with exponents.

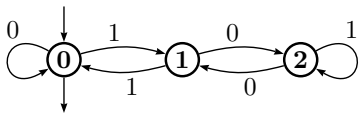
Example 1.1 (continued) With these conventions, the automaton $\mathcal{A}_1$ is described by the following transition table:



	${}^I p$	q	r^T
a	$\{p,q\}$	$\emptyset$	r
b	p	r	r

□

Example 1.2 (continued) The table in the prologue which allowed us to describe
cf. p. 6
'the machine that reads numbers from left to right and calculates the remainder of division by 3' is, up to the initial and final states, the transition table of the automaton which we called $\mathcal{P}_2$ in Section 1:



	${}^I\mathbf{0}^T$	**1**	**2**
0	**0**	**2**	**1**
1	**1**	**0**	**2**

□

The definition of an automaton by a transition function does not imply that the automaton must be finite. Only the notion of transition table is problematic in this case: it is replaced by a non-extensive definition of the transition function.

cf. p. 52 **Example 1.3** (continued) We also write $\mathcal{Z}_1 = \langle \mathbb{Z}, \{a,b\}, \zeta_1, \{0\}, \{0\} \rangle$, with, for all z in $\mathbb{Z}$, $(z,a)\zeta_1 = z+1$ and $(z,b)\zeta_1 = z-1$. □

Just as we go from transitions labelled with letters to computations labelled with words, we will extend the domain of definition of the map δ from $Q \times A$ to $\mathfrak{P}(Q) \times A^*$, in two steps. We first extend δ additively to a map from $\mathfrak{P}(Q) \times A$ to $\mathfrak{P}(Q)$:

$$\forall P \subseteq Q \,,\ \forall a \in A \qquad (P,a)\delta = \bigcup_{p \in P} (p,a)\delta \ ; \tag{3.2}$$

then, by induction on the length of words, to a map from $\mathfrak{P}(Q) \times A^*$ to $\mathfrak{P}(Q)$:

$$\forall P \subseteq Q \qquad (P, 1_{A^*})\delta = P \tag{3.3}$$

$$\forall P \subseteq Q\,,\ \forall f \in A^*\,,\ \forall a \in A \qquad (P, f\,a)\delta = ((P, f)\delta, a)\delta\ . \tag{3.4}$$

Hence we conclude: *cf. Exer. 3.4*

Property 3.1 $\forall P \subseteq Q\,,\ \forall f, g \in A^* \qquad (P, f\,g)\delta = ((P, f)\delta, g)\delta\,.$ ■

This definition is consistent with that of a computation of $\mathcal{A}$ since there exists a computation with source p, destination computation q and label f if and only if q is in the image of (p, a) under δ: that is, we have:

Property 3.2 $\forall P \subseteq Q\,,\ \forall f \in A^* \qquad (P, f)\delta = \{q \mid \exists p \in P \quad p \xrightarrow[\mathcal{A}]{f} q\}\ .$

Proof. By induction on the length of f. The property is true for $|f| = 0$, since (3.3) corresponds to the convention that each state is the source and destination of a computation of length 0 labelled with 1_{A^*}, and for $|f| = 1$ by (3.1) and (3.2). For all P in Q, all f in A^* and all a in A, we obtain:

$$\begin{aligned}(P, f\,a)\delta = ((P, f)\delta, a)\delta &= \{q \mid \exists r \in (P, f)\delta \quad r \xrightarrow[\mathcal{A}]{a} q\} \\ &= \{q \mid \exists p \in P\,,\ \exists r \in Q \quad p \xrightarrow[\mathcal{A}]{f} r \quad \text{and} \quad r \xrightarrow[\mathcal{A}]{a} q\} \\ &= \{q \mid \exists p \in P \quad p \xrightarrow[\mathcal{A}]{f\,a} q\}\ .\end{aligned}$$

The last equality expresses that every computation of length $n + 1$ is obtained by concatenating a computation of length n with a transition. ■

Then entails:[31]

Corollary 3.1 *Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton over A and δ its transition function. Then*

$$L(\mathcal{A}) = \{f \in A^* \mid (I, f)\delta \cap T \neq \emptyset\}\ .$$ ■

We see that considering a (finite) automaton as a set of transitions represented by a labelled graph is exactly equivalent to considering it as a transition function represented by its table. From now on we will use the following expressions interchangeably, without comment, according to which is best suited to our purposes:

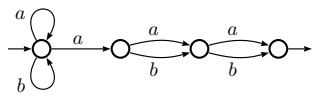
$$\mathcal{A} = \langle Q, A, E, I, T \rangle \qquad \text{or} \qquad \mathcal{A} = \langle Q, A, \delta, I, T \rangle\ .$$

[31] Yes, this is correct English, and no, I won't do it again.

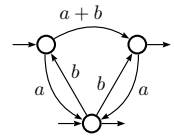
Exercises

- 3.1 Give the transition table of the following automata:

(a)



(b)



- 3.2 Give the graph representation of the automaton defined by the transition table opposite.

	${}^{I}p$	q^{T}	${}^{I}r$
a	$\{q,r\}$	$\{p,q\}$	$\{p,r\}$
b	p	$\emptyset$	q

- 3.3 Say what it means in terms of the transition function for an automaton to be complete.
- 3.4 Prove Property 3.1.

3.2 Deterministic automata

cf. § 1.4.3, p. 77 The description of an automaton by its transition function leads naturally to the (re)definition of deterministic automata and to the formulation of their properties.

3.2.1 Reformulation of the definition

cf. Def. 1.13, p. 77 **Definition 3.1** An automaton $\mathcal{A} = \langle\, Q, A, \delta, I, T \,\rangle$ is *deterministic* if two conditions are satisfied:

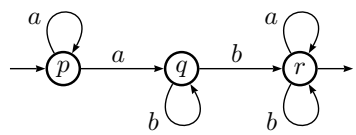
(i) the transition function $\delta\colon Q \times A \to \mathfrak{P}(Q)$ is such that, for all p in Q and all a in A, we have that $(p,a)\delta$, if it is non-empty, is a *singleton*; in other words, δ is a *function* (possibly partial) from $Q \times A$ to Q;

(ii) there is at most one initial state: that is, I is a singleton, or empty. □

As we have seen, $\mathcal{A}_1$ is not deterministic since $(p,a)\delta_1 = \{p,q\}$. The automata $\mathcal{P}_2$ and $\mathcal{Z}_1$ are deterministic. It should be clear that Definition 3.1 is, via Equation (3.1), identical to Definition 1.13.

Example 3.1 The automaton $\mathcal{A}'_1 = \langle\, \{p,q,r\}, A, \delta'_1, p, r \,\rangle$, defined by the table

	${}^{I}p$	q	r^{T}
a	q	q	r
b	p	r	r



cf. Exam. 1.1, p. 78 is deterministic, and shown above. We can easily verify that $\mathcal{A}'_1$ is equivalent to $\mathcal{A}_1$: that is, recognises the set of words that contain a factor $a\,b$. □

Equations (3.3) and (3.4) can be simplified for deterministic automata:

$$\forall p \in Q \qquad (p, 1_{A^*})\delta = p \tag{3.5}$$

$$\forall p \in Q\,,\ \forall f \in A^*\,,\ \forall a \in A \qquad (p, f\,a)\delta = ((p,f)\delta, a)\delta \ . \tag{3.6}$$

Hence we deduce the corresponding version of Property 3.1:

Property 3.3 $\forall p \in Q\,,\ \forall f, g \in A^* \qquad (p, f\,g)\delta = ((p, f)\delta, g)\delta\ .$ ■

Notation It is useful to have a simpler notation, so when it does not lead to confusion we shall write

$$p \cdot a \qquad \text{for} \qquad (p, a)\delta\ .$$

This notation evokes the fact that the letter a *acts* on the automaton $\mathcal{A}$ in state p to take it to state $q = p \cdot a$. In the same spirit we will write, for all f in A^*:

$$p \cdot f \qquad \text{for} \qquad (p, f)\delta\ .$$

In fact, this is about more than just a notation, and we will return to this theme later. Property 3.3 is therefore written *cf. § II.2.2.1, p. 238*

$$\forall p \in Q\,,\ \forall f, g \in A^* \qquad p \cdot f\,g = (p \cdot f) \cdot g\ , \tag{3.7}$$

and Corollary 3.1 becomes

$$L(\mathcal{A}) = \{f \in A^* \mid i \cdot f \in T\}\ . \tag{3.8}$$

We will also write $\mathcal{A} = \langle\, Q, A, \delta, i, T\,\rangle$ to show that there is a single initial state.

3.2.2 Determinisation

The following property, though elementary to show, is fundamental. It follows naturally from the functional form of the definition of an automaton.

Proposition 3.2 *Every automaton $\mathcal{A}$ is equivalent to a complete deterministic automaton $\mathcal{D}$. If $\mathcal{A}$ is finite with n states, $\mathcal{D}$ can be constructed with at most 2^n states.*

Proof. Let $\mathcal{A} = \langle\, Q, A, \delta, I, T\,\rangle$ be an automaton, *a priori* non-deterministic, and let

$$\mathcal{D} = \langle\, \mathfrak{P}(Q), A, \delta, \{I\}, U\,\rangle\ ,$$

be the automaton whose transition function is the additive extension of the transition function of $\mathcal{A}$ (again written δ and defined by Equation (3.2)):

$$\forall P \subseteq Q\,,\ \forall a \in A \qquad (P, a)\delta = \bigcup_{p \in P} (p, a)\delta\ , \tag{3.9}$$

and whose final states are given by

$$U = \{S \subseteq Q \mid S \cap T \neq \emptyset\}\ . \tag{3.10}$$

The automaton $\mathcal{D}$ is deterministic: its transition function δ is a function from $\mathfrak{P}(Q) \times A$ to $\mathfrak{P}(Q)$ (not to $\mathfrak{P}(\mathfrak{P}(Q))$) and in addition $\{I\}$ is the unique initial state of $\mathcal{D}$. Furthermore, it is equivalent to $\mathcal{A}$. Indeed, we have the series of equivalences:

$$\begin{aligned} \forall f \in A^* \quad f \in L(\mathcal{A}) \quad &\Longleftrightarrow \quad \exists i \in I \quad (i, f)\delta \cap T \neq \emptyset \quad \Longleftrightarrow \quad (I, f)\delta \cap T \neq \emptyset \\ &\Longleftrightarrow \quad (I, f)\delta \in U \quad \Longleftrightarrow \quad f \in L(\mathcal{D})\ . \end{aligned}$$ ■

Example 1.1 (continued) Application to the automaton $\mathcal{A}_1$. Starting from the transition table of $\mathcal{A}_1$, shown in the adjacent figure, we compute that of $\mathcal{D}_1$ using (3.9):

	${}^I p$	q	r^T
a	$\{p,q\}$	$\emptyset$	r
b	p	r	r

	$\emptyset$	${}^I\{p\}$	$\{q\}$	$\{r\}^T$	$\{p,q\}$	$\{q,r\}^T$	$\{p,r\}^T$	$\{p,q,r\}^T$
a	$\emptyset$	$\{p,q\}$	$\emptyset$	$\{r\}$	$\{p,q\}$	$\{r\}$	$\{p,q,r\}$	$\{p,q,r\}$
b	$\emptyset$	$\{p\}$	$\{r\}$	$\{r\}$	$\{p,r\}$	$\{r\}$	$\{p,r\}$	$\{p,r\}$

from which we deduce the graph representation of $\mathcal{D}_1$ (Figure 3.1). □

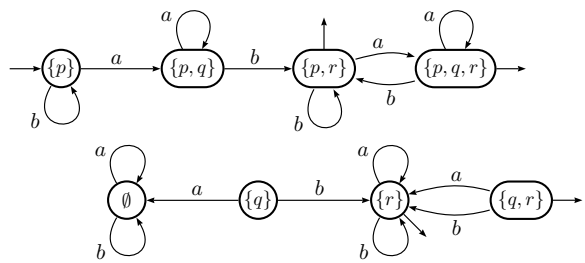


Figure 3.1: The automaton $\mathcal{D}_1$

We worked through this example in accordance with the proof of Proposition 3.2. It showed that part of the computation was 'useless' since it corresponded to an *inaccessible* subset of $\mathcal{D}_1$. This gives the following definition its interest and utility.

Definition 3.2 The *determinisation* of an automaton $\mathcal{A} = \langle Q, A, \delta, I, T \rangle$ *by subset construction* is the computation of the accessible subset of the automaton $\mathcal{D} = \langle \mathfrak{P}(Q), A, \delta, \{I\}, U \rangle$ performed step by step, starting from the initial state $\{I\}$, and considering only those states obtained by applying δ to the states already calculated. The automaton thus obtained is called the *determinisation of* $\mathcal{A}$, and written $\mathcal{A}_{\mathsf{det}}$. We will also say 'determinise $\mathcal{A}$' to mean 'compute $\mathcal{A}_{\mathsf{det}}$'. □

Example 1.1 (continued) The transition table of $\mathcal{A}_{1\mathsf{det}}$, extracted from the table of $\mathcal{D}_1$ above, is given opposite. The automaton $\mathcal{A}_{1\mathsf{det}}$ is shown in Figure 3.2. □

	${}^I\{p\}$	$\{p,q\}$	$\{p,r\}^T$	$\{p,q,r\}^T$
a	$\{p,q\}$	$\{p,q\}$	$\{p,q,r\}$	$\{p,q,r\}$
b	$\{p\}$	$\{p,r\}$	$\{p,r\}$	$\{p,r\}$

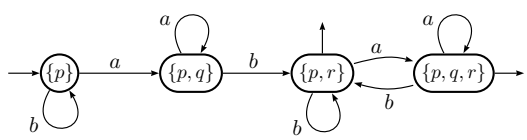


Figure 3.2: The automaton $\mathcal{A}_{1\mathsf{det}}$, determinisation of $\mathcal{A}_1$

Remark 3.1 If the automaton $\mathcal{A}$ is defined by a set of transitions rather than as a transition function $\mathcal{A} = \langle Q, A, E, I, T \rangle$, then Equation (3.9) which defines $\mathcal{A}_{\mathsf{det}}$ is instead written

$$\forall P \subseteq Q\,,\ \forall a \in A \qquad P \xrightarrow[\mathcal{A}_{\mathsf{det}}]{a} S \quad \Longleftrightarrow \quad S = \{q \mid \exists p \in P \quad p \xrightarrow[\mathcal{A}]{a} q\} \tag{3.11}$$

from which, by induction on the length of words, we get

$$\forall P \subseteq Q\,,\ \forall f \in A^* \qquad P \xrightarrow[\mathcal{A}_{\mathsf{det}}]{f} S \quad \Longleftrightarrow \quad S = \{q \mid \exists p \in P \quad p \xrightarrow[\mathcal{A}]{f} q\} \tag{3.12}$$

By definition, $\mathcal{A}_{\mathsf{det}}$ is a (deterministic) complete accessible automaton. As an example of performing the determinisation procedure, we will show that it nearly preserves trim automata: that is:

Property 3.4 *If $\mathcal{A}$ is co-accessible, then the states of $\mathcal{A}_{\mathsf{det}}$ are all co-accessible, except perhaps for the sink state $\emptyset$.*

Proof. Let P be a state of $\mathcal{A}_{\mathsf{det}}$; by definition, there exists f in A^* such that

$$P = \{p \mid \exists i \in I \quad i \xrightarrow[\mathcal{A}]{f} p\}\ .$$

If P is non-empty, there exists p in P, and p is a state of $\mathcal{A}$. By hypothesis, $\mathcal{A}$ is co-accessible, so there exists g in A^* such that $p \xrightarrow[\mathcal{A}]{g} t$ where t is in T. Thus, $P \cdot g \cap T \neq \emptyset$ and $P \cdot g = S$ is a final state of $\mathcal{A}_{\mathsf{det}}$: P is co-accessible. ∎

The transformation from a general automaton to an equivalent deterministic automaton is the key to some fundamental results we shall shortly see: Proposition 3.4, Theorem 3.1 and Corollary 3.6 on one hand, and Theorem 3.2 on the other. We therefore cannot overemphasise the importance of the determinisation procedure, which is the effective realisation of Proposition 3.2. The statement of this proposition warns of a potential combinatorial explosion; the following fact, which we will confirm in exercises, shows that the risk is real.

Fact 3.1 *For every integer n, there exist automata with n states whose determinisation has exactly 2^n states.*

It is worth mentioning here that with ideas developed in the next subsection we shall be able to make this fact more precise.

Fact 3.2 *For every integer n, there exist automata with n states such that every equivalent deterministic automaton has* at least *2^n states.*

These facts also call for the study of the cases where this explosion does not occur. Let us return for an instant to Example 2.3. There we constructed a normalised automaton which accepted a^*b^*. This non-deterministic automaton had 12 states; its determinisation is therefore liable to have 4000, compared with the 2-state deterministic automaton of Example 1.4, or that of 3 states in Example 2.5. This illustrates the fact that an effective proof, even a conceptually simple one such as that of Proposition 2.4, does not *ipso facto* give us control over the size of the objects used, and that we must concern ourselves with this as well.

cf. p. 57 and p. 94

Remark 3.2 Determinisation is an example of a situation where, as stated in Remark 1.8, we write somewhat imprecisely, as

$$\mathcal{A}_{\mathsf{det}} = \langle\, \mathfrak{P}(Q), A, \delta, \{I\}, T_{\mathsf{det}} \,\rangle \qquad \text{with} \qquad T_{\mathsf{det}} = \{S \subseteq Q \mid S \cap T \neq \emptyset\} \ ,$$

although not all the elements of $\mathfrak{P}(Q)$ are accessible and T_{det} is really a subset of R, the accessible subset of $\mathfrak{P}(Q)$.

Similarly, if $\mathcal{A} = \langle\, Q, A, E, I, T \,\rangle$ is defined as a labelled graph, we can write $\mathcal{A}_{\mathsf{det}} = \langle\, \mathfrak{P}(Q), A, E_{\mathsf{det}}, \{I\}, T_{\mathsf{det}} \,\rangle$ where

$$E_{\mathsf{det}} = \{(P, a, S) \in \mathfrak{P}(Q) \times A \times \mathfrak{P}(Q) \mid \forall s \in S\,,\ \exists p \in P \quad (p, a, s) \in E\} \ .$$

By doing this we preserve the relationship between the set of states of $\mathcal{A}$ and those of $\mathcal{A}_{\mathsf{det}}$. □

The idea of a deterministic automaton is directed: determinisation admits a dual operation. We say that an automaton $\mathcal{A}$ is *co-deterministic* if and only if its transpose is deterministic. The transpose of the determinisation of the transpose of $\mathcal{A}$ is called its *co-determinisation*:

cf. Def. 1.13, p. 77

$$\mathcal{A}_{\mathsf{cod}} = \big(\big(\mathcal{A}^{\mathsf{t}}\big)_{\mathsf{det}}\big)^{\mathsf{t}} \ .$$

Since $(\mathcal{A}^{\mathsf{t}})_{\mathsf{det}}$ and $\mathcal{A}^{\mathsf{t}}$ are equivalent, $\mathcal{A}_{\mathsf{cod}}$ is equivalent to $(\mathcal{A}^{\mathsf{t}})^{\mathsf{t}} = \mathcal{A}$. We can define, or rather formulate, the co-determinisation of $\mathcal{A}$:

$$\mathcal{A}_{\mathsf{cod}} = \langle\, \mathfrak{P}(Q), A, E_{\mathsf{cod}}, I_{\mathsf{cod}}, \{T\} \,\rangle \ ,$$

where $\quad E_{\mathsf{cod}} = \{(P, a, R) \in \mathfrak{P}(Q) \times A \times \mathfrak{P}(Q) \mid \forall p \in P\,,\ \exists r \in R \quad (p, a, r) \in E\} \ ,$

and $\quad I_{\mathsf{cod}} = \{P \subseteq Q \mid P \cap I \neq \emptyset\} \ .$

Practical computations are generally performed on the transposed automaton because the manipulation of maps is more natural than that of inverse maps.[32]

Example 1.1 (continued) The automaton $\mathcal{A}_{1\,\mathsf{cod}}$ is shown in Figure 3.3; a useful exercise is to compute its transition table directly starting from the final state r:

	${}^{I}\{p,q,r\}$	${}^{I}\{p,r\}$	$\{q,r\}$	$\{r\}^{T}$
a	$\emptyset$	$\big\{\{p,q,r\},\{p,r\},\{q,r\}\big\}$	$\emptyset$	$\{r\}$
b	$\big\{\{p,q,r\},\{p,r\}\big\}$	$\emptyset$	$\{r\}$	$\emptyset$

□

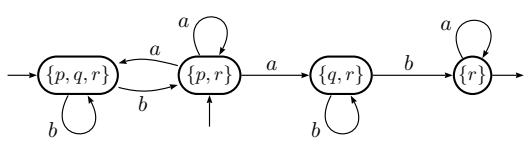


Figure 3.3: The automaton $\mathcal{A}_{1\,\mathsf{cod}}$

[32] I should point out that the notion of co-determinisation will only really be used in Chapter V.

3.2.3 The case of one-letter alphabets

In the case where the alphabet A has a single letter $A = \{a\}$, an automaton over A, which we call a *unary automaton*, necessarily has the form shown in Figure 3.4, if it is finite and deterministic. As every automaton is equivalent to a deterministic automaton, we can deduce the characterisation of recognisable unary languages.

Proposition 3.3 *A language L over $\{a\}^*$ is recognisable if and only if it can be written $L = F \cup G$ where F is a finite set and G is either empty or defined by integers r and $p \geqslant 1$ and k integers $0 \leqslant i_1 < i_2 < \cdots < i_k \leqslant p-1$ such that:*

$$G = \{a^n \mid n \geqslant r \quad \textit{and} \quad \exists j \in [k] \qquad n \equiv i_j \bmod p\} \ . \qquad \blacksquare$$

If we identify $\{a\}^*$ and (the additive monoid) $\mathbb{N}$, we can reformulate Proposition 3.3 by saying that a unary language is recognisable if and only if it is the union of a finite set and a finite union of (infinite) arithmetic progressions of the same period.

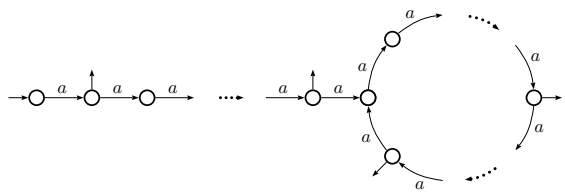


Figure 3.4: A deterministic finite unary automaton

3.2.4 Complement of recognisable languages

The fundamental property of deterministic automata is that they make the complement of the language they recognise as easy to get at as the language itself, which we express thus:

Property 3.5 *Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ be a complete deterministic automaton. The complement of $L(\mathcal{A})$ is recognised by the automaton $\langle Q, A, \delta, i, Q \setminus T \rangle$.*

Proof. Since $\mathcal{A}$ is complete, $(i, f)\delta$ exists (is non-empty) for all f in A^*, and since $\mathcal{A}$ is deterministic, $(i, f)\delta$ is a unique state of $\mathcal{A}$. Since, by definition, $L(\mathcal{A}) = \{f \in A^* \mid (i, f)\delta \in T\}$ we then have:

$$A^* \setminus L(\mathcal{A}) = \{f \in A^* \mid (i, f)\delta \notin T\} = \{f \in A^* \mid (i, f)\delta \in Q \setminus T\} \ . \qquad \blacksquare$$

From the fact that every finite automaton is equivalent to a finite complete deterministic automaton and therefore that every recognisable language is recognised by a complete deterministic automaton, we deduce the above-mentioned property.

Proposition 3.4 *The complement of a recognisable language is recognisable: that is,* $\mathrm{Rec}\, A^*$ *is closed under complement.* $\blacksquare$

Kleene's Theorem allows us to restate this proposition:

Corollary 3.5 *The complement of a rational language is rational: that is,* $\mathrm{Rat}\, A^*$ *is closed under complement.* ■

This repetition gives us nothing new, of course. However, it helps us better to understand the importance of Kleene's Theorem: in particular it is not obvious from their definition that the rational languages are closed under complement, but no other proof of this fact is known than the procedure of turning them into automata and determinising them.

cf. pp. 61, 61 and 70

We can combine Propositions 1.2, 1.4, 1.11 and 3.4, which all correspond to effective constructions on automata, into a single statement:

Theorem 3.1 $\mathrm{Rec}\, A^*$*, and hence* $\mathrm{Rat}\, A^*$*, is an effective Boolean algebra.*[33] ■

We can also point out a remarkable consequence and elaborate its proof:

Corollary 3.6 *The equivalence of finite automata is decidable.*

Proof. Thanks to Proposition 3.2, we can assume without loss of generality that the two automata $\mathcal{A}'$ and $\mathcal{A}''$ are deterministic. We then note that $L(\mathcal{A}') = L(\mathcal{A}'')$ if and only if

$$L = \left(\overline{L(\mathcal{A}')} \cap L(\mathcal{A}'')\right) \cup \left(L(\mathcal{A}') \cap \overline{L(\mathcal{A}'')}\right) = \emptyset \ ,$$

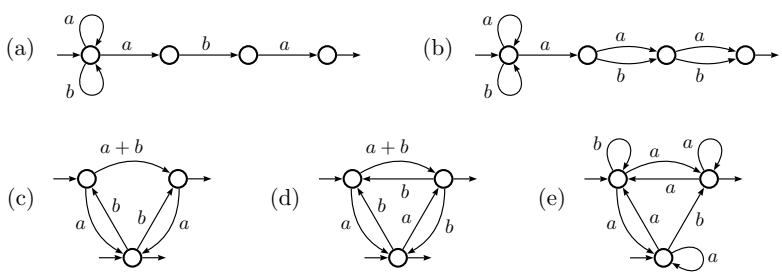
where $\overline{X}$ denotes the complement of X. By virtue of our previous work we can effectively construct from $\mathcal{A}'$ and $\mathcal{A}''$ an automaton $\mathcal{A}$ which recognises L, then decide whether the language recognised by $\mathcal{A}$ is empty or not (Proposition 1.11). ■

cf. p. 56

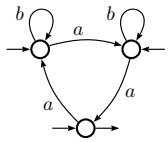
This result, and more broadly Theorem 3.1, return us to the remarks that follow Definition 1.6. Finite automata are a good way of representing recognisable languages since we know how to test whether two such objects are equal; and once again the importance of Kleene's Theorem comes from allowing us to transfer this representational ability to rational expressions.

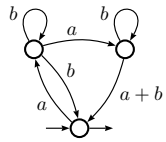
Exercises

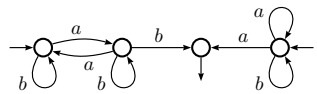
• 3.5 Determinise the following automata:



[33] A Boolean algebra $\mathcal{F}$ of subsets of a set E is called *effective* if, given (effectively) two elements X and Y of $\mathcal{F}$, (a description of) $X \cup Y$ and $\overline{X}$ can be effectively computed, and if it is decidable whether X is empty or not (*cf.* Section 0.8).

(f) 

(g) 

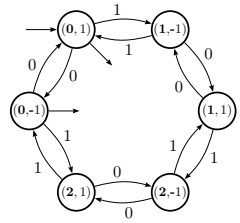
(h) 

.

• 3.6 Which of the automata in the previous exercise suggests the definition of families of automata (indexed by n) that supports Fact 3.1? *cf. p. 107*

• 3.7 The automaton shown in the adjacent figure, called $\mathcal{P}_1$, was obtained from Figure P.1 (a) in the prologue by making $(0, +1)$ and $(0, -1)$ final states; in other words, $\mathcal{P}_1$ recognises the binary representations of numbers divisible by 3, *reading from right to left*.

Find the determinisation of the transpose of $\mathcal{P}_1$.



3.8 **Longest common prefix.** Show that the longest common prefix of all the words in a recognisable language is effectively computable.

3.9 **Prefix-closed languages.** Show that it is decidable whether a recognisable language is prefix-closed.

• 3.10 **Discriminating length I.** (a) Show (using the methods of this section) that two complete deterministic automata $\mathcal{A}$ and $\mathcal{B}$, with n and m states respectively, are equivalent if and only if they accept the same words of length less than $k = n\,m$.

We will see that it is possible to do much better: cf. in particular, Exer. III.4.13, p. 462 and Rem. III.SE.4.1, p. 516.

(The *discriminating length* of two automata is the shortest length of a word accepted by one and not the other, if the two automata are not equivalent. The question above may therefore be stated: 'Show that the discriminating length of two complete deterministic automata with respectively n and m states is smaller than $n\,m$'.)

(b) Deduce an upper bound for the discriminating length of any two automata with n and m states respectively.

3.11 **Complexity of determinisation.** Calculate the complexity of determinisation by the subset construction as a function of the size of the result.

3.3 Minimisation

By studying deterministic automata we have just shown that it is decidable whether two finite automata recognise the same language. Indeed, we proved the result with great economy: having used the property of deterministic automata that they can be used to recognise both a language and its complement, we were able to rely on the general properties of finite automata with regard to the union and intersection of recognisable languages. *cf. Cor. 3.6*

We will now prove a much stronger result, of which the previous one will be just a straightforward consequence. We are going to show that a *canonical* deterministic automaton can be associated with every language – recognisable or not – that is unique (up to isomorphism), called the *minimal automaton* of the language. Then, a language L is recognisable if and only if its minimal automaton is finite, and it

is the complete deterministic automaton with the smallest number of states that recognises L. The essence of this construction is the notion of taking quotients of a language by words.

We will show next that, in the case of recognisable languages, this minimal automaton can be computed effectively from any finite deterministic automaton which defines the language. Deciding whether two recognisable languages are equal will therefore come down to testing whether their minimal automata are isomorphic.

3.3.1 The automaton of quotients of a language...

cf. § 1.2.3, p. 62

Recall that to take the *(left) quotient* by a word f of A^* means associating with f a *(partial) function* from A^* to itself, defined by

$$f^{-1}g = \begin{cases} h & \text{if} \quad g = f\,h \ , \\ \text{undefined} & \text{if} \quad f \text{ is not a prefix of } g \ , \end{cases}$$

a function which extends additively to a map from $\mathfrak{P}(A^*)$ to itself:

$$\forall X \subseteq A^* \qquad f^{-1}X = \{h \mid f\,h \in X\} \ .$$

The two following properties are obvious.

Property 3.6 $\forall f \in A^*\,,\ X \subseteq A^* \qquad 1_{A^*} \in f^{-1}X \ \Leftrightarrow\ f \in X\,.$ ■

cf. Exer. 1.13, p. 65

Property 3.7 $\forall f, g \in A^*\,,\ X \subseteq A^* \qquad (f\,g)^{-1}X = g^{-1}(f^{-1}X)\,.$

Proof. The equivalence $(f\,g)\,h \in X \Leftrightarrow f\,(g\,h) \in X$ implies the series of equivalences:

$$h \in (f\,g)^{-1}X \Leftrightarrow g\,h \in f^{-1}X \Leftrightarrow h \in g^{-1}(f^{-1}X) \ .$$ ■

It says that the quotient is an action of A^* on $\mathfrak{P}(A^*)$; cf. Chap. II.

Though obvious, this last property is no less fundamental. It is the basis of the construction we seek. Let L be any language of A^*, even though the case that really interest us is that of recognisable languages, and Q_L be the set of quotients of L:

$$Q_L = \{f^{-1}L \mid f \in A^*\} \ .$$

The set Q_L is therefore a family of subsets of A^* and we note that $L \in Q_L$ since $L = (1_{A^*})^{-1}L$. This Q_L is the set of states of an automaton

$$\mathcal{A}_L = \langle\, Q_L, A, \delta_L, \{L\}, T_L \,\rangle$$

which we call the *automaton of quotients* of L and which is defined by

$$\forall f \in A^*\,,\ \forall a \in A \qquad ((f^{-1}L), a)\delta_L = a^{-1}(f^{-1}L) = (f\,a)^{-1}L$$

and $$T_L = \{f^{-1}L \mid 1_{A^*} \in f^{-1}L\} = \{f^{-1}L \mid f \in L\} \ .$$

We conclude immediately from the definition (and from Property 3.7) that $\mathcal{A}_L$ is deterministic, that $(L, f)\delta_L = f^{-1}L$ for all f in A^* and, from the definition of T_L and from Property 3.6, it follows that:

Proposition 3.7 *For all languages L over A^*, $\mathcal{A}_L$ recognises L.* ■

Example 3.2 Let $Z_1 = L(\mathcal{Z}_1) = \{f \in \{a,b\}^* \mid |f|_a = |f|_b\}$. For all f in $\{a,b\}^*$, we have

$$f^{-1}Z_1 = \{g \in \{a,b\}^* \mid |g|_b - |g|_a = |f|_a - |f|_b\} \ .$$

cf. Exam. 1.3, p. 52, p. 56 and p. 102

Each state of $\mathcal{A}_{Z_1}$ is identified with an integer, and we obtain $\mathcal{A}_{Z_1} = \mathcal{Z}_1$. □

Example 3.3 The quotients of $P_1 = \{a^n b^n \mid n \in \mathbb{N}\}$ are described by the following equation (*cf.* Exam. 1.6, p. 64 and p. 72):

$$f^{-1}P_1 = \begin{cases} \{a^m b^{m+k} \mid m \in \mathbb{N}\} & \text{if} \quad f = a^k \ , \\ \{b^j\} & \text{if} \quad f = a^{r+j}b^r \ , \\ \emptyset & \text{otherwise} \ . \end{cases}$$

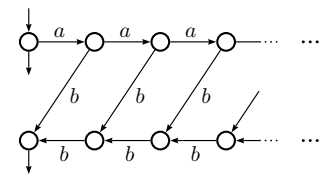


Fig. 3.5: Automaton $\mathcal{A}_{P_1}$

Hence the automaton $\mathcal{A}_{P_1}$ shown in the adjacent figure without its sink state or the transitions which lead to it. □

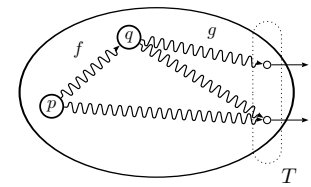
3.3.2 ...is minimal

Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ be a complete deterministic accessible automaton and $L = L(\mathcal{A})$ the language that it recognises. The application of $Q \times A^*$ in Q defined by δ is written as a dot. As in § 2.4.3, and for all p in Q, we write L_p for the set of words which send the state p to a final state:

cf. p. 99

$$L_p = \{f \in A^* \mid p \cdot f \in T\} \ .$$

Thus L_p is the language recognised by the automaton $\mathcal{A}$ in which the initial state i has been replaced by p; in particular, $L = L_i$. Because $\mathcal{A}$ is deterministic, and if, as in the figure, $p \cdot f = q$, we have the series of equivalences



$$\begin{aligned} g \in L_q \quad &\Longleftrightarrow \quad q \cdot g \in T \quad \Longleftrightarrow \quad p \cdot f\,g \in T \\ &\Longleftrightarrow \quad f\,g \in L_p \quad \Longleftrightarrow \quad g \in f^{-1}L_p \ . \end{aligned}$$

In other words, we have just proved:

Lemma 3.8 $\forall f \in A^*, \ \forall p, q \in Q \qquad p \cdot f = q \quad \Longrightarrow \quad L_q = f^{-1}L_p$. ■

Theorem 3.2 *Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ be an accessible deterministic automaton and $L = L(\mathcal{A})$ the language that recognises it. Let $\mathcal{A}_L = \langle Q_L, A, \delta_L, \{L\}, T_L \rangle$ be the automaton of quotients of L. There exists a map $\varphi\colon Q \to Q_L$ such that:*[34]

$$\forall q \in Q \ , \ \forall a \in A \qquad [(q,a)\delta]\,\varphi = (q\varphi, a)\delta_L \tag{3.13}$$

$$i\varphi = \{L\} \ , \qquad \text{and} \qquad \forall t \in T \qquad t\varphi \in T_L \tag{3.14}$$

and φ is surjective if $\mathcal{A}$ is complete.

The automaton $\mathcal{A}_L$ is called the minimal automaton *of L.*

[34] This is what we shall call an *automata morphism*, or *covering* (see Section II.3).

Proof. Let p be in Q; Lemma 3.8 tells us that for all f in A^* such that $i \cdot f = p$ we have $L_p = f^{-1}L$, which defines a map $\varphi\colon Q \to Q_L$ as $p\varphi = f^{-1}L$ (since the value of $p\varphi$ does not depend on the choice of f). We can easily verify that φ satisfies (3.13) and (3.14).

If $\mathcal{A}$ is complete, $i \cdot f$ is defined for all f and φ is surjective. ■

It is important to emphasise that what we have just written (about the automaton of quotients, or minimal automaton) holds for every language L, even if it is non-recognisable; in other words, it does not require $\mathcal{A}$ to be finite. If we suppose precisely that L is recognisable, that is to say, that $\mathcal{A}$ is finite, we conclude from the foregoing that Q_L *is finite* and can therefore make the following two statements:

Corollary 3.9 *The automaton $\mathcal{A}_L$ of quotients of a recognisable language L is the complete deterministic automaton with the smallest number of states that recognises L.* ■

Proposition 3.10 *A language L is recognisable if and only if the family of subsets $f^{-1}L$, for all f in A^*, is finite.* ■

This characterisation of recognisable languages is of the first importance, from both a theoretical and a technical point of view. It will stay with us through varied generalisations of recognisable languages obtained by considering automata over general monoids, then weighted automata, extending even to the characterisation of a certain family of functions realised by finite automata. It is a powerful tool for proving that certain sets are recognisable; we will give a first example in Subsection 3.4.

cf. p. 118

The idea of a minimal automaton allows us to restate Fact 3.2:

cf. p. 107

Fact 3.2 *For every integer n, there exists an automaton with n states which accepts languages whose minimal automaton has 2^n states.*

3.3.3 Computation of the minimal automaton

Knowing that the automaton of quotients is minimal is not enough. We must also show that this automaton is *effectively computable*, at least in the case of a recognisable language explicitly described by an automaton.

Definition 3.3 Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ be a deterministic automaton. Then, *Nerode's equivalence* is the relation ν defined on Q by

$$p \equiv q \bmod \nu \quad \Longleftrightarrow \quad L_p = L_q \ . \qquad \square$$

The definition of ν implies that it is an equivalence relation. We can simplify the notation by writing (when there is no risk of ambiguity)

$$p \sim q$$

when p and q are equivalent modulo ν.

Properties 3.8 (i) $p \sim q \quad \Longrightarrow \quad \forall f \in A^* \quad p \cdot f \sim q \cdot f$;
(ii) $p \sim q \quad \Longrightarrow \quad (p \in T \quad \Longleftrightarrow \quad q \in T)$.

Proof. Property (i) is a direct consequence of Lemma 3.8; Property (ii) follows because p is in T if and only if 1_{A^*} is in L_p. ■

We can therefore define a new deterministic automaton $\mathcal{A}/\nu$, the quotient of $\mathcal{A}$ by the equivalence ν:

$$\mathcal{A}/\nu = \langle\, Q/\nu, A, \delta_\nu, [i]_\nu, T/\nu \,\rangle$$

where $Q/\nu = \{[p]_\nu \mid p \in Q\}$ is the set of equivalence classes modulo ν and where the transition function δ_ν is defined by the action of the letters of A on the classes $[p]_\nu$:

$$[p]_\nu \cdot a = [q]_\nu \qquad \Longleftrightarrow \qquad p \cdot a = q \ .$$

This definition is valid: that is, does not depend on the choice of p and q, thanks to Property 3.8 (i); Property 3.8 (ii) ensures that T is saturated by ν.

The definition of ν implies that the map φ is a bijection between Q/ν and Q_L if $\mathcal{A}$ is complete, which allows us to state:

Proposition 3.11 *If $\mathcal{A}$ is a complete deterministic automaton which recognises L, $\mathcal{A}/\nu$ is isomorphic to $\mathcal{A}_L$.* ■

Our programme will be completed when we have shown:

Proposition 3.12 *The automaton $\mathcal{A}/\nu$ is effectively computable from the deterministic automaton $\mathcal{A} = \langle\, Q, A, \delta, i, T \,\rangle$.*

Proof. We define a series $\{\mathcal{Q}_0, \mathcal{Q}_1, \ldots\}$ of *nested partitions* of Q in the following manner:

- $\mathcal{Q}_0 = \{Q \setminus T, T\}$;
- for all k, $\mathcal{Q}_{k+1}$ is *the* refinement of $\mathcal{Q}_k$ such that, for all p and q in Q, p and q are connected in $\mathcal{Q}_k$ and disconnected in $\mathcal{Q}_{k+1}$ if there exists a letter a in A such that $p \cdot a$ and $q \cdot a$ are disconnected in $\mathcal{Q}_k$, in other words, if $p \equiv q \bmod \mathcal{Q}_k$, then:

$$p \not\equiv q \bmod \mathcal{Q}_{k+1} \quad \Longleftrightarrow \quad \exists a \in A \qquad p \cdot a \not\equiv q \cdot a \bmod \mathcal{Q}_k \ . \tag{3.15}$$

If Q is finite, then there is only a finite number of possible partitions; each successive partition $\mathcal{Q}_k$ is effectively computable. It suffices therefore to show that when the above computation terminates (that is, when the integer k is such that $\mathcal{Q}_k = \mathcal{Q}_{k+1}$) we have in fact reached the minimal automaton: that is to say that $L_p = L_q$ for every pair of states p and q that are equivalent modulo $\mathcal{Q}_k$.

Suppose that this property is false, and choose p and q equivalent modulo $\mathcal{Q}_k$ so that not only is

$$D_{p,q} = [L_p \setminus L_q] \cup [L_q \setminus L_p]$$

non-empty, but also $D_{p,q}$ contains words of minimal length for all non-empty $D_{r,s}$. Let f be one of these words of minimal length. We have $f \neq 1_{A^*}$ since $\mathcal{Q}_k$ is a

refinement of $\mathcal{Q}_0$ which distinguishes the states p for which $1_{A^*} \in L_p$ from those for which $1_{A^*} \notin L_p$. Thus $f = ag$ with a in A; we have $p \cdot a \equiv q \cdot a \bmod \mathcal{Q}_k$ by the definition of k, and g belongs to $D_{r,s}$, with $r = p \cdot a$ and $s = q \cdot a$. The word g is shorter than f, which contradicts the choice of p and q. ■

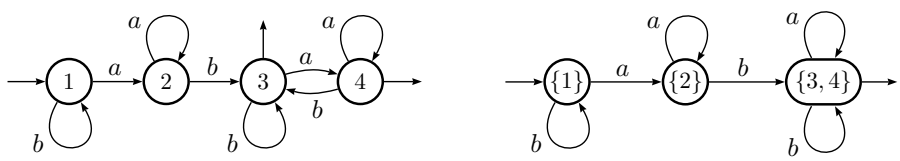
The computation described by Equation (3.15) is known as *Moore's algorithm.*

Example 1.1 (continued) Consider again the determinisation of $\mathcal{A}_1$, whose states we have renamed for clarity. We have $\mathcal{Q}_0 = \big\{\{1,2\},\{3,4\}\big\}$ and then compute:

$$\begin{aligned} 1 \cdot a = 2 \equiv 2 \cdot a = 2 \bmod \mathcal{Q}_0 \ , \qquad & 1 \cdot b = 1 \not\equiv 2 \cdot b = 3 \bmod \mathcal{Q}_0 \\ 3 \cdot a = 4 \equiv 4 \cdot a = 4 \bmod \mathcal{Q}_0 \ , \qquad & 3 \cdot b = 3 \equiv 4 \cdot b = 3 \bmod \mathcal{Q}_0 \end{aligned} \tag{3.16}$$

from which $\mathcal{Q}_1 = \big\{\{1\},\{2\},\{3,4\}\big\}$. Furthermore, Equations (3.16) already show that $\mathcal{Q}_2 = \mathcal{Q}_1$. From this we obtain the result of the computation, shown in Figure 3.6: the minimal automaton of $L(\mathcal{A}_1)$ is equal to $\mathcal{A}'_1$ (which is hardly surprising). □

cf. Exam. 3.1, p. 104



(a) The determinisation of $\mathcal{A}_1$ (b) The minimal automaton of $L(\mathcal{A}_1)$

Figure 3.6: Moore's algorithm on $\mathcal{A}_{1\mathsf{det}}$

Remark 3.3 As the minimal automaton of a language is a fundamental object, the complexity of its computation is an important question. A naive algorithm for the computation of nested partitions is $O(n^2)$ (where n is the number of states of the initial deterministic automaton). A more involved algorithm lowers the complexity to $O(n \log n)$ (see bibliographic notes). □

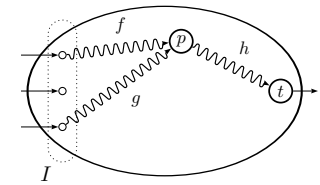
cf. p. 214

3.3.4 Another minimisation method

We can calculate the minimal automaton by a method other than the successive refinement of a partition. The new method is based on the following result.

Proposition 3.13 *The determinisation of a co-deterministic co-accessible automaton which recognises L is the minimal automaton of L.*

Proof. Let $\mathcal{A} = \langle Q, A, \delta, I, \{t\} \rangle$ be any co-deterministic co-accessible automaton which recognises L, and $\mathcal{A}_{\mathsf{det}} = \langle \mathfrak{P}(Q), A, \delta, \{I\}, U \rangle$ be its determinisation. We will show that if two words f and g in A^* are such that $f^{-1}L = g^{-1}L$ then $(I,f)\delta = (I,g)\delta$ and therefore that the states of $\mathcal{A}_{\mathsf{det}}$ are in bijection with those of the minimal automaton of L.



Let p be a member of $(I, f)\delta$; since $\mathcal{A}$ is co-accessible, there exists an h such that t is in $(p, h)\delta$ and $f\,h$ is in L.

By hypothesis, $g\,h$ is also in L and t belongs to $(I, g\,h)\delta$. As $\mathcal{A}$ is co-deterministic, there is only one computation in $\mathcal{A}$ with label h which ends at t and this computation's source is p. Hence p belongs to $(I, g)\delta$ and $(I, f)\delta \subseteq (I, g)\delta$.

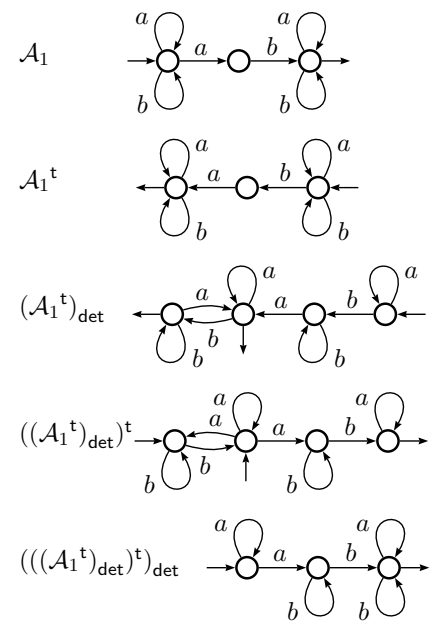
Symmetrically, $(I, g)\delta \subseteq (I, f)\delta$ and the proposition is established. ■

Proposition 3.13 is itself a procedure, due to J. Brzozowski, which makes the computation of the minimal automaton of a language into a double determinisation.

Corollary 3.14 *Let $\mathcal{A}$ be an automaton and $L = L(\mathcal{A})$ the language which it recognises. Then*

$$\mathcal{A}_L = (((\mathcal{A}^{\mathsf{t}})_{\mathsf{det}})^{\mathsf{t}})_{\mathsf{det}} \quad . \qquad ■$$

The construction described in the corollary is shown in the adjacent figure for the automaton $\mathcal{A}_1$.



We begin with an automaton $\mathcal{A}$ (deterministic or non-deterministic) which recognises L. We compute its transpose $\mathcal{A}^{\mathsf{t}}$, then the determinisation of the transpose, $(\mathcal{A}^{\mathsf{t}})_{\mathsf{det}}$.

By definition, $(\mathcal{A}^{\mathsf{t}})_{\mathsf{det}}$ is accessible, so its transpose $((\mathcal{A}^{\mathsf{t}})_{\mathsf{det}})^{\mathsf{t}}$ is co-accessible and this last is equivalent to $\mathcal{A}$. The determinisation $(((\mathcal{A}^{\mathsf{t}})_{\mathsf{det}})^{\mathsf{t}})_{\mathsf{det}}$ is the sought-for minimal automaton.

It remains to calculate the complexity of this method of minimisation and to compare it to that of the previous method.

If we start with a deterministic automaton, this method is not obviously good, because it is $O(2^n)$.

If we start with a non-deterministic automaton, determinisation does not impose an additional penalty, as we need to determinise the automaton before using Moore's algorithm.

Exercises

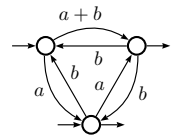
3.12 Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ be a deterministic automaton, and for every integer k let λ_k be the equivalence on Q defined by

$$p \equiv q \bmod \lambda_k \quad \Longleftrightarrow \quad [L_p] \cap A^{\leqslant k} = [L_q] \cap A^{\leqslant k} \quad .$$

Show that $\lambda_k = \mathcal{Q}_k$ (where the $\mathcal{Q}_k$ are the equivalences on Q defined in the description of Moore's algorithm).

• 3.13 Minimise the deterministic automata calculated in Exercise 3.5.

• 3.14 Which of the automata in Exercise 3.5 suggest the definition of families of automata (indexed by n) that are evidence for Fact 3.2; that is, automata with n states that recognise languages whose minimal automaton has 2^n states?



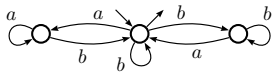
• 3.15 Find the language recognised by the automaton in the adjacent figure.

cf. Exer. 3.7, p. 111

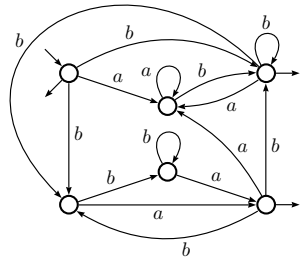
• 3.16 Find the minimal automaton of the language recognised by $\mathcal{P}_1$.

• 3.17 Determinise and minimise the following automata:

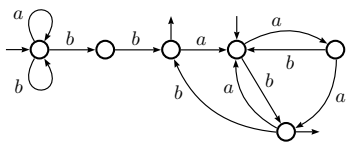
(a)



(b)



(c)



• 3.18 Give the minimal automaton of the language $\big((a\,(a+b)^2+b)^*a\,(a+b)\big)^*$.

3.19 Justify the complexity given for the second minimisation method: why is it $O(2^n)$ and not $O(2^{2^n})$ as the two successive determinisations performed by this method might suggest?

• 3.20 What is the effect of Moore's algorithm on a non-deterministic automaton?

• 3.21 **State complexity of operations on languages.** Some authors call the number of states of the minimal automaton of a recognisable language L its *state complexity*. We will write it $\mathsf{Sc}(L)$. We can measure the effect that different operations on languages have on this complexity. Show that:

(a) $\mathsf{Sc}(L \cup K) \leqslant \mathsf{Sc}(L)\,\mathsf{Sc}(K)$; (b) $\mathsf{Sc}(L \cap K) \leqslant \mathsf{Sc}(L)\,\mathsf{Sc}(K)$;

(c) $\mathsf{Sc}(L\,K) \leqslant (2\,\mathsf{Sc}(L)-1)\,2^{\mathsf{Sc}(K)-1}$; (d) $\mathsf{Sc}(L^*) \leqslant 3\;2^{\mathsf{Sc}(L)-2}$;

(e) $\mathsf{Sc}(L^{\mathsf{t}}) \leqslant 2^{\mathsf{Sc}(L)}$.

and that all these bounds are tight.

3.4 Return to the Star Lemma

cf. Lem. 1.12, 1.13 and 1.14, p. 71 et seq.

We saw, under the generic name 'Star Lemma', some properties that can be described as combinatorial and that are necessarily satisfied by every rational language. We also saw that none of these properties is characteristic of rational languages: a language may very well satisfy them without being rational.

cf. Exer. 1.24, p. 74

It appears however that if we consider *the language and its complement* together (we already know that the complement of a recognisable language is recognisable), then the Block Star Lemma is indeed characteristic of recognisable languages (and thus of rational languages). The precise statement of this property requires some preparation and definitions. The proof consists of the use, albeit simple, of a fundamental combinatorial result, known as Ramsey's Theorem.

3.4.1 Block iteration and block simplification

For a given integer h, we say that a language L over A^* satisfies the property[35] $\mathfrak{E}'_h$, or belongs to $\mathfrak{E}'_h(A^*)$, if for every word f of A^* and *every factorisation* of the form

$$f = u\,v_1\,v_2 \cdots v_h\,w \qquad \text{with} \qquad |v_i| \geqslant 1 \quad \forall i \in [h] \ , \tag{3.17}$$

there exists a pair of indices (j,k), $0 \leqslant j < k \leqslant h$, such that the word obtained by iterating in f the factor $v_{j+1} \cdots v_k$ any number of times (possibly zero) is in L if and only if f is; that is, *cf. Lem. 1.14, p. 71*

$$\forall n \in \mathbb{N} \qquad f \in L \quad \Longleftrightarrow \quad u\,v_1\,v_2 \cdots v_j\,(v_{j+1} \cdots v_k)^n\,v_{k+1} \cdots v_h\,w \in L \ .$$

As in § 1.3.2 and § 1.3.3, we adopt the convention that $v_1 v_2 \cdots v_j = 1_{A^*}$ if $j = 0$ and that $v_{k+1} v_{k+2} \cdots v_h = 1_{A^*}$ if $k = h$. The property $\mathfrak{E}'_h$ corresponds to that stated by the Block Star Lemma, but for a language and its complement at the same time. *cf. pp. 70 and 71*

We will say that a language L over A^* satisfies $\mathfrak{E}_h$, or belongs to $\mathfrak{E}_h(A^*)$, if for every word f in A^* and *every factorisation* of the form (3.17), there exists a pair of indices (j,k), $0 \leqslant j < k \leqslant h$, such that the word obtained by deleting the factor $v_{j+1} \cdots v_k$ from the word f is in L if and only if f is; that is,

$$f \in L \quad \Longleftrightarrow \quad u\,v_1\,v_2 \cdots v_j\,v_{k+1} \cdots v_h\,w \in L \ .$$

We can now state the desired result.

Theorem 3.3 [**Ehrenfeucht, Parikh & Rozenberg**]
Let L be a language of A^*. The following three conditions are equivalent:

(i) *L is recognisable;*
(ii) *there exists an integer h such that L satisfies $\mathfrak{E}'_h$;*
(iii) *there exists an integer h such that L satisfies $\mathfrak{E}_h$.*

3.4.2 Ramsey's Theorem

We are going to generalise the simplistic Pigeonhole Principle which we used to prove the above-mentioned star lemmas: *cf. p. 70*

> *Every partition into m classes of a set E whose cardinality is greater than or equal to $(r-1)\,m+1$ contains at least one class (a 'pigeonhole') with at least r elements.*

which we will first restate as follows:

> *For every number of classes m and every integer r, there exists an integer N,*
>
> $$N = (r-1)\,m + 1 \ ,$$
>
> *such that, if a set E is of cardinality greater than or equal to N, then for every partition $\mathcal{P}$ of E into m classes a subset of E of cardinality r can be found which is completely contained in a single class of $\mathcal{P}$.*

[35] $\mathfrak{E}$ is a gothic 'E'.

The generalisation rests on the fact that it is not the set E which is partitioned into m classes, but $\mathfrak{P}_{(k)}(E)$, the set of all the subsets of E with k elements, and is stated in the following terms.

Theorem 3.4 [**Ramsey**] *For all integers k (size of the subsets), m (number of classes), and r, there exists an integer $N = R(k,m,r)$ such that, if a set E has a cardinality greater than or equal to N, then for every partition $\mathcal{P}$ of $\mathfrak{P}_{(k)}(E)$ into m classes a subset F of E with cardinality r can be found such that $\mathfrak{P}_{(k)}(F)$ is completely contained in a single class of $\mathcal{P}$.*

The Pigeonhole Principle is an instance of Ramsey's Theorem for $k = 1$, $\mathfrak{P}_{(1)}(E)$ being identified naturally with E, and $R(1,m,r) = (r-1)\,m+1$. We shall not give the proof of Theorem 3.4, which can be found in many standard works on combinatorics.

see Notes, p. 214

3.4.3 Proof of Theorem 3.3

If L is recognisable, it is recognised by a deterministic complete automaton with h states, for some h. The reasoning that proves the Block Star Lemma shows that L satisfies $\mathfrak{E}'_h$. Clearly, $\mathfrak{E}'_h$ implies $\mathfrak{E}_h$. It remains to show that (iii) implies (i).

Lem. 1.14, p. 71

Let A be an alphabet and h a given integer. We shall show:

(a) that every (left) quotient of a language in $\mathfrak{E}_h(A^*)$ is a language in $\mathfrak{E}_h(A^*)$ (which is easy);

(b) and that the family $\mathfrak{E}_h(A^*)$ is finite (which is the core of the proof, and needs Ramsey's Theorem).

Properties (a) and (b) imply that every language which satisfies $\mathfrak{E}_h$ has a finite number of left quotients (less than the cardinality of $\mathfrak{E}_h(A^*)$) and is therefore recognisable.

Prop. 3.10, p. 114

(a) Let $L \in \mathfrak{E}_h(A^*)$, t be in A^*, and $f = u\,v_1\,v_2 \cdots v_h\,w$ a factorisation of a word f in A^*. Consider the word $f' = t\,f$; by hypothesis on L, there exists a pair of indices (j,k), $0 \leqslant j < k \leqslant h$, such that the word g', obtained by deleting in f' the factor $v_{j+1} \cdots v_k$, is in L if and only if f' is; that is,

$$f' = t\,u\,v_1\,v_2 \cdots v_h\,w \in L \quad \Longleftrightarrow \quad g' = t\,u\,v_1\,v_2 \cdots v_j\,v_{k+1} \cdots v_h\,w \in L \ ,$$

and hence

$$f = u\,v_1\,v_2 \cdots v_h\,w \in t^{-1}\,L \quad \Longleftrightarrow \quad g = u\,v_1\,v_2 \cdots v_j\,v_{k+1} \cdots v_h\,w \in t^{-1}\,L \ ,$$

which means that $t^{-1}\,L \in \mathfrak{E}_h(A^*)$.

(b) Now let h be a given integer, and set

$$N = R(2,2,h+1) \ ,$$

then write E for the set

$$E = \{0,1,\ldots,N-1\} \ .$$

Let L and K be two languages in $\mathfrak{E}_h(A^*)$ which coincide for words of length less than $N-1$; that is,

$$|f| < N-1 \quad \Longrightarrow \quad [f \in L \Leftrightarrow f \in K] \ .$$

We will show, by induction on the length of words, that $K = L$. Suppose that L and K coincide for words of length less than M, $M \geqslant N-1$, and let $f = a_1\, a_2 \cdots a_{N-1}\, t$ be a word from A^* of length M (the a_i are letters of A). Let X_f and Y_f be subsets of $\mathfrak{P}_{(2)}(E)$ defined by

$$X_f = \{(j,k) \mid 0 \leqslant j < k \leqslant N-1 \quad \text{and} \quad a_1\, a_2 \cdots a_j\, a_{k+1} \cdots a_{N-1}\, t \in L\}$$
$$\text{and} \qquad Y_f = \mathfrak{P}_{(2)}(E) \setminus X_f \ .$$

By the induction hypothesis we also have

$$X_f = \{(j,k) \mid 0 \leqslant j < k \leqslant N-1 \quad \text{and} \quad a_1\, a_2 \cdots a_j\, a_{k+1} \cdots a_{N-1}\, t \in K\} \ .$$

By Ramsey's Theorem, there exists a subset $F = \{i_0 < i_1 < \cdots < i_h\}$ of E, with cardinality $h+1$, such that either

$$\mathfrak{P}_{(2)}(F) \subseteq X_f \ , \tag{3.18}$$

or

$$\mathfrak{P}_{(2)}(F) \subseteq Y_f \ . \tag{3.19}$$

Then let $f = u\, v_1\, v_2 \cdots v_h\, w\, t$ be the factorisation of f obtained by considering the h pairs of consecutive indices in F (*cf.* Figure 3.7).

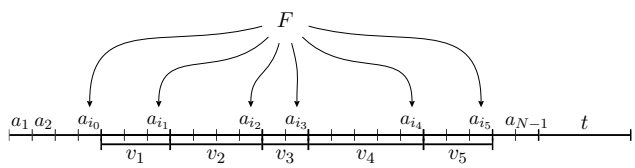


Figure 3.7: Ramsey factorisation

The choice of F implies that:

(i) either, for all $0 \leqslant j < k \leqslant h$, the word $u\, v_1\, v_2 \cdots v_j\, v_{k+1} \cdots v_h\, w\, t$ is in both L and K (if (3.18) holds);

(ii) or, for all $0 \leqslant j < k \leqslant h$, the word $u\, v_1\, v_2 \cdots v_j\, v_{k+1} \cdots v_h\, w\, t$ is in neither L nor K (if (3.19) holds).

Since L and K are in $\mathfrak{E}_h(A^*)$, there exist j' and k', $0 \leqslant j' < k' \leqslant h$, such that f and $f' = u\, v_1\, v_2 \cdots v_{j'}\, v_{k'+1} \cdots v_h\, w\, t$ are either both in L or both not in L, and there exist j'' and k'', $0 \leqslant j'' < k'' \leqslant h$, such that f and $f'' = u\, v_1\, v_2 \cdots v_{j''}\, v_{k''+1} \cdots v_h\, w\, t$ are either both in K or both not in K.

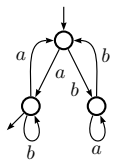
In case (i), f' and f'' are in L and K; and f is in L since f' is; and in K since f'' is. In case (ii), f' and f'' are in neither L nor K; and f is neither in L, like f', nor in K, like f''. In both cases, f has the same status with respect to L and K.

By induction, we have therefore shown that $K = L$ and hence that $\mathfrak{E}_h(A^*)$ is finite (indeed, that $\|\mathfrak{E}_h(A^*)\| \leqslant 2^{\|A\|^{R(2,2,h+1)}}$), and Theorem 3.3 has been proved. ■

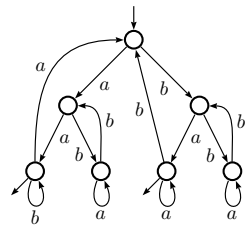
Exercises

3.22 Show that the following automata are minimal:

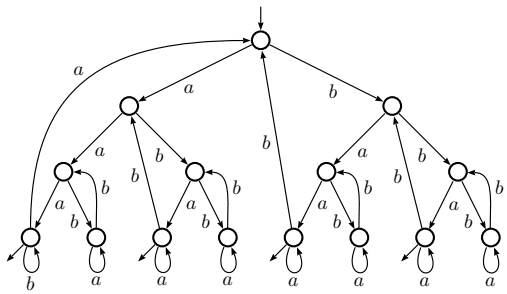
(a)



(b)



(c)



(d) Generalise the construction to an automaton with $2^n - 1$ states.

• 3.23 **Another characteristic iteration lemma.**
Imitating the preceding definitions, and for a given integer h, let us say that a language L over A^* satisfies the property[36] $\mathfrak{J}'_h$, or belongs to $\mathfrak{J}'_h(A^*)$, if, for every word f in A^h, there exists a factorisation $f = u\,v\,w$, with $|v| \geqslant 1$, such that every word g, obtained by iterating v in f (any number of times), has the same set of right contexts for L as f: that is,

$$\forall t \in A^*\,,\ \forall n \in \mathbb{N} \qquad f\,t \in L \quad \Longleftrightarrow \quad u\,(v)^n\,w\,t \in L\ .$$

Similarly, we will say that a language L over A^* satisfies the property $\mathfrak{J}_h$, or belongs to $\mathfrak{J}_h(A^*)$, if, for every word f in A^h, there exists a factorisation $f = u\,v\,w$, where $|v| \geqslant 1$, such that the word g, obtained by removing v from f, has the same set of right contexts for L as f: that is,

$$\forall t \in A^* \qquad f\,t \in L \quad \Longleftrightarrow \quad u\,w\,t \in L\ .$$

(a) Show (directly, that is without using Theorem 3.3):

Proposition 3.15 *Let L be a language of A^*. The following three conditions are equivalent:*

(i) *L is recognisable;*

(ii) *there exists an integer h such that L satisfies $\mathfrak{J}'_h$;*

(ii) *there exists an integer h such that L satisfies $\mathfrak{J}_h$.*

[36] $\mathfrak{J}$ is a gothic 'J'.

(b) Explain why Proposition 3.15 cannot be considered a combinatorial property of the words of L in the same way as Theorem 3.3.

(c) The answer to question (a) leads to the conclusion that if $L \subseteq A^*$ is recognised by a complete deterministic automaton with h states, it satisfies $\mathfrak{J}_h$ and that if L satisfies $\mathfrak{J}_h$, it is recognised by a complete deterministic automaton with $\|A\|^h$ states (if $\|A\| > 1$).

Show that this gap cannot be bridged: that is, that there exists, for every h, two languages over A^*, J_h and Y_h, which both satisfy $\mathfrak{J}_h$, and not $\mathfrak{J}_{h-1}$, and such that the minimal automaton of J_h has h states and that of Y_h has $\|A\|^h - 1$ states.

• 3.24 **An application of Ramsey's Theorem to free monoids.**

(a) Give a proof of the following proposition based on Ramsey's Theorem and comment.

Proposition 3.16 *For every integer m, and every integer r, there exists an integer $N = S(m,r)$ such that, for every partition $\mathcal{P}$ of A^+ into m classes, every word of length greater than N contains r consecutive factors in the same class of $\mathcal{P}$.*

(b) Show directly that $S(m,r) \leqslant r^m$, and that this bound is tight if $\|A\| = m$.

4 Rational expressions

We return now to the definition of rational languages using rational operations. Our initial definition – a language is rational 'if it belongs to the *smallest family* of rationally closed languages which contains the finite languages' – is not at all effective. After making this definition we showed how every rational language can be obtained from a series of union, product and star operations, described by a rational expression; this can be considered as a constructive definition of rational languages. This inductive definition is superficially apparent in the descriptions of the McNaughton–Yamada algorithm and of the state elimination method, even if it is not formally indispensable in those instances.

Def. 2.2, p. 86

cf. pp. 95–99

We thus give in this section a precise definition of rational expressions, with the same formalism as used, for example, for arithmetic expressions. Rational expressions will then become, like finite automata, a way of describing rational (or recognisable) languages *in finite terms*. They make explicit how a given language is created from finite sets with rational operations. Furthermore, two expressions are termed 'equivalent' if they represent two procedures that produce the same language.

This new definition will give us two new perspectives. The first is the study of the expressions themselves (that is, without reference to automata), and in particular that of syntactic transformations that preserve equivalence. The second sheds new light on Kleene's Theorem. We stated it in an abstract manner: an equality between two (infinite) families of languages (themselves potentially infinite objects). We can now see it as a concrete proposition about the existence of algorithms (finite procedures) which transform (finite) automata into rational expressions (also finite objects) and conversely of algorithms that compute an automaton from an expression. In both cases, the proof consists of exhibiting such algorithms, but it is clear that in the

Th. 2.1, p. 87

second case there is more interest in the performance of the algorithms, in terms of complexity or of the size of the results produced.

After the definition of rational expressions, we shall touch on the problem of determining the equivalence of expressions from a formal, or syntactic, point of view (that is, without referring to the construction of the corresponding automata). We will then use the ideas thus defined to make explicit the 'degree of closeness' between the different expressions obtained for the language recognised by a single finite automaton by applying the MNY algorithm or state elimination method, using the different possible orders on the states of the automaton. In the fourth part, we shall present the operation of *derivation* of expressions, which is the symbolic counterpart of the quotient of a language.

The transformation of a rational expression into a finite automaton, which is at the heart of many a problem in computer science, will be treated in the next section.

4.1 Rational expressions and languages

We begin by defining rational expressions on a fixed alphabet, and the languages they denote, then rational expressions on a set of variables and the result of interpreting such an expression.

4.1.1 Rational expressions over an alphabet

Let A be a given (non-empty) alphabet and let $\{\mathsf{0}, \mathsf{1}, +, \cdot, *\}$ be five function symbols. Naturally, the operations $+$ and $\cdot$ are binary, $*$ is unary, and $\mathsf{0}$ and $\mathsf{1}$ are nullary (they represent constants).

Definition 4.1 A *rational expression over A* is a formula obtained inductively from the letters of A and the functions $\{\mathsf{0}, \mathsf{1}, +, \cdot, *\}$ in the following manner:

(i) $\mathsf{0}$, $\mathsf{1}$, and a, for a in A, are rational expressions;

(ii) if E and F are rational expressions, then $(\mathsf{E} + \mathsf{F})$, $(\mathsf{E} \cdot \mathsf{F})$, and (E^*) are rational expressions.

We write $\mathsf{RatE}\, A^*$ for the set of rational expressions over A. □

Stating it in another way, rational expressions are *well-formed formulas*, made from the letters of A and $\mathsf{0}$ and $\mathsf{1}$ taken as *atomic formulas*, the binary operators $+$ and $\cdot$, and the unary operator $*$. For example,

$$\begin{gathered}((a \cdot b) + (b \cdot a)) \ , \quad ((((a + b)^*) \cdot (a \cdot b)) \cdot ((a + b)^*)) \ , \\ \text{and} \qquad ((a + (b \cdot ((((a \cdot (b^*)) \cdot a)^*) \cdot b)))^*)\end{gathered} \tag{4.1}$$

are rational expressions.

The definition of rational expressions calls for two other typographical symbols as well as the atoms: we needed the open and close parenthesis to ensure that expressions can be written unambiguously, since we have chosen an infix notation for reasons of legibility for the operators $+$ and $\cdot$. Without parentheses we would be unable to

distinguish the two expressions $((a+b)\cdot c)$ and $(a+(b\cdot c))$. (We could have used a postfix or prefix notation, in which case parentheses would not have been required.) *cf. Exer. 4.1*

As we see in the expressions (4.1), the parentheses rapidly become the most numerous symbols as the expressions grow in length, making reading difficult (rather compromising our aim of legibility!) and leading to frequent errors in writing. We therefore adopt an *operator precedence* convention: $*$ takes precedence over $\cdot$ which takes precedence over $+$. With this convention we write

$$a\cdot b+c^* \quad \text{for} \quad ((a\cdot b)+(c^*)) \ , \quad a\cdot(b+c)^* \quad \text{for} \quad (a\cdot((b+c)^*)) \ ,$$
$$\text{and} \qquad (a\cdot b+c)^* \quad \text{for} \quad (((a\cdot b)+c)^*) \ .$$

The expressions (4.1) become

$$a\cdot b+b\cdot a \ , \quad ((a+b)^*\cdot(a\cdot b))\cdot(a+b)^* \qquad \text{and} \qquad (a+b\cdot(((a\cdot b^*)\cdot a)^*\cdot b))^* \ .$$

Most of the definitions and propositions about rational expressions will consider the process of making formulas from expressions. This process can be followed by defining the *depth*[37] of an expression E which is an integer written $\mathsf{d}(\mathsf{E})$ and calculated in the following manner:

$$\mathsf{d}(0)=\mathsf{d}(1)=\mathsf{d}(a)=0 \ , \quad \text{for all } a \text{ in } A \ , \qquad \mathsf{d}((\mathsf{E}^*))=1+\mathsf{d}(\mathsf{E}) \ ,$$
$$\mathsf{d}((\mathsf{E}+\mathsf{F}))=\mathsf{d}((\mathsf{E}\cdot\mathsf{F}))=1+\mathsf{max}\big(\mathsf{d}(\mathsf{E}),\mathsf{d}(\mathsf{F})\big) \ .$$

Definition 4.2 To each rational expression E in $\mathsf{Rat}\mathsf{E}\,A^*$ we assign a corresponding language of A^*, written[38] $L[\mathsf{E}]$, or $|\mathsf{E}|$, and defined inductively – that is, by induction on the depth of expressions:

(i) for atomic expressions:

$$L[0]=\emptyset \ , \quad L[1]=1_{A^*} \ , \quad \text{and} \quad L[a]=a \quad \text{for all } a \text{ in } A \ ; \tag{4.2}$$

(ii) for composite expressions:

$$\begin{gathered} L[(\mathsf{E}+\mathsf{F})]=\{L[\mathsf{E}]\}\cup\{L[\mathsf{F}]\} \ , \quad L[(\mathsf{E}\cdot\mathsf{F})]=\{L[\mathsf{E}]\}\{L[\mathsf{F}]\} \ , \\ \text{and} \quad L[(\mathsf{E}^*)]=\{L[\mathsf{E}]\}^* \ . \end{gathered} \tag{4.3}$$

We say that E *denotes* the language $L[\mathsf{E}]$. □

For example,

$$L[(a+b)]=\{a,b\} \ , \tag{4.4}$$
$$L[((((a+b)^*)\cdot(a\cdot b))\cdot((a+b)^*))]=\{a,b\}^*a\,b\,\{a,b\}^* \ . \tag{4.5}$$

[37] I use the term 'depth' rather than 'height' (which is also possible and even more widely used) to avoid any confusion with 'star height', which will be defined later (*cf.* Section 6). Both terms use the image of a rational expression as a tree.

[38] I use square brackets in $L[\mathsf{E}]$ to be different from the parentheses used in expressions. This slightly cumbersome notation is used in the definitions to make them more legible. The more concise notation $|\mathsf{E}|$ will be used hereafter, when we are more accustomed to expressions.

Thus, $((((a+b)^*)\cdot(a\cdot b))\cdot((a+b)^*))$ denotes the language of words in A^* that contain at least one factor $a\,b$.

At first glance this last assertion, or equations such as (4.4) or (4.5), may seem utterly tautological. Why bother to make the subtle distinction between $(a+b)$ and $a\cup b$? To write pompously that $L[a]=a$? Are we not pointlessly adding unnecessary complexity? Actually, no. We must understand expressions as *syntactic* objects, and the languages which they denote as their *semantics*. For example, $(a+b)$ and $(b+a)$ are two *distinct* expressions which denote the *same* language $\{a,b\}$.

A procedure for automatic information processing – in short, a 'computer' – can only manipulate sequences of symbols and can therefore access only a syntactic description of the objects with which it deals. The role of programming is to ensure that objects which are syntactically different but semantically identical produce the same result.

If this argument has convinced the reader that I am not engaging in tetrapyloctomy for pure sadism, we can continue, starting by showing that the usage of the same adjective *rational* for languages and expressions is justified.

cf. p. 87 **Proposition 4.1** *A language of A^* is rational if and only if it is denoted by a rational expression over A.*

Proof. The family of languages denoted by a rational expression contains $\emptyset$, 1_{A^*}, and all the letters of A by (4.2); it is closed under union, product and star by (4.3): it therefore contains $\mathrm{Rat}\,A^*$.

Conversely, by induction on the depth of expressions, every language denoted by an expression is obtained from letters and the empty word by a finite series of unions, products and stars; it is thus contained in any rationaly closed family (that contains letters and empty word), and hence rational. ■

cf. pp. 87–94 We can therefore give another version of Proposition 2.1 and of its proofs:

cf. Sec. 5, p. 145 **Corollary 4.2** *Every rational expression* E *can effectively be transformed into an automaton that recognises the language denoted by* E. ■

Definition 4.3 Two rational expressions over A are *equivalent* if they denote the same language; that is, E and F are equivalent if $L[\mathsf{E}]=L[\mathsf{F}]$ and in this case we write $\mathsf{E}\equiv\mathsf{F}$. □

cf. Exam. 2.6, p. 96 **Example 4.1** In Section 2 we tried several methods for calculating the language recognised by the automaton $\mathcal{R}_1$ and obtained different expressions that denote the same language and are therefore equivalent:

$$(a^*\cdot b)^+\cdot a^+ + a^* \quad\equiv\quad 1_{A^*}+(a^*\cdot b)^*\cdot a^+ \quad\equiv\quad (b^*\cdot a)^* \ ,$$

where E^+ is an abbreviation for $\mathsf{E}\cdot\mathsf{E}^*\equiv\mathsf{E}^*\cdot\mathsf{E}$. □

Corollaries 3.6 and 4.2 immediately give us:

Theorem 4.1 *The equivalence of two rational expressions over A is decidable.* ■

Remark 4.1 This result, which 'falls out' without needing a particular proof, is nevertheless fundamental. It is also necessary to realise that *no other proof is known* of the decidability of the equivalence of expressions than by going via automata,[39] and hence by using Kleene's Theorem. □

4.1.2 Rational expressions over a set of variables

It is useful to give a more abstract definition of rational expressions in order to be able to distinguish atomic formulas from letters of the alphabet. Let $\mathbf{X} = \{\mathsf{x}, \mathsf{y}, \ldots\}$ be a set (*a priori* infinite, but countable) of symbols, called *variables*. A rational expression over $\mathbf{X}$ is, as in Definition 4.1, a well-formed formula constructed from $\mathsf{0}$, $\mathsf{1}$ and the elements of $\mathbf{X}$ taken as atomic formulas, and the operations $\{+, \cdot, *\}$. We write $\operatorname{Var}\mathsf{E}$ for the set of variables that appear in the expression E. Thus for example

$$\mathsf{E}_1 = (\mathsf{x} + \mathsf{y} \cdot (((\mathsf{x} \cdot \mathsf{t}^*) \cdot \mathsf{z})^* \cdot \mathsf{y})^*)^*$$

is a rational expression, and $\operatorname{Var}\mathsf{E}_1 = \{\mathsf{x}, \mathsf{y}, \mathsf{z}, \mathsf{t}\}$.

Definition 4.4 An *interpretation* of a rational expression E over A^* is a map τ from $\operatorname{Var}\mathsf{E}$ to $\mathfrak{P}(A^*)$, and hence a morphism from $(\operatorname{Var}\mathsf{E})^*$ to $\mathfrak{P}(A^*)$. The *result* of the interpretation τ of E, written $[\mathsf{E}]\tau$, is defined by induction on the depth of E:

(i) for atomic expressions:

$$[\mathsf{0}]\tau = \emptyset \ , \quad [\mathsf{1}]\tau = 1_{A^*} \ , \quad \text{and} \quad [\mathsf{x}]\tau = \mathsf{x}\tau \quad \text{for all } \mathsf{x} \text{ in } \operatorname{Var}\mathsf{E} \ ; \tag{4.6}$$

(ii) for composite expressions:

$$\begin{gathered}[(\mathsf{F} + \mathsf{G})]\tau = \{[\mathsf{F}]\tau\} \cup \{[\mathsf{G}]\tau\} \ , \quad [(\mathsf{F} \cdot \mathsf{G})]\tau = \{[\mathsf{F}]\tau\}\{[\mathsf{G}]\tau\} \ , \\ \text{and} \quad [(\mathsf{F}^*)]\tau = \{[\mathsf{F}]\tau\}^* \ .\end{gathered} \tag{4.7}$$

Example 4.2 If τ_1 is the interpretation of $\operatorname{Var}\mathsf{E}_1$ over A^* defined by

$$[\mathsf{x}]\tau_1 = [\mathsf{z}]\tau_1 = a \quad \text{and} \quad [\mathsf{y}]\tau_1 = [\mathsf{t}]\tau_1 = b \ ,$$

it follows that $[\mathsf{E}_1]\tau_1 = (a \cup b\,(((a\,b^*)\,a)^* b))^*$. □

This more abstract definition allows us to broaden the use of rational expressions to semirings other than $\mathfrak{P}(A^*)$, and gives us some terminological abbreviations that we will use later. It also gives rise to a proposition more general than Proposition 4.1, with an identical proof.

[39] These sorts of assertion always cause objections to be raised. In particular, the computation of *derivatives* (and also of *derived terms*) that we shall see later (Sec. 4.4, Sec. 5.2, Sec. III.4.2) might give the impression that we could develop procedures for proving the equivalence of expressions entirely in the world of expressions. We shall see that the computation of derivatives corresponds in fact to the computation of an automaton.

Proposition 4.3 *Let $\mathcal{C}$ be a family of subsets of A^*. A subset R of A^* belongs to* $\mathrm{Rat}\,\mathcal{C}$ *if and only if there exists a rational expression* E *and an interpretation of* E *whose value is in $\mathcal{C}$ and whose result is R.* ■

Definition 4.3 extends naturally: two expressions over the same set of variables are equivalent if they have the same result *for every interpretation*. Amongst all the possible interpretations of a rational expression E, one plays a special role: the *free interpretation* of E, the interpretation of E in $(\mathrm{Var}\,\mathsf{E})^*$ which associates $\{\mathsf{x}\}$ with x and whose result is written $L[\mathsf{E}]$ (this notation is consistent with Definition 4.2). This free interpretation is the ancestor of all the other interpretations, in the sense that we have the following result immediately by induction on the depth of E.

Proposition 4.4 *Let* E *be a rational expression and τ any interpretation of* E *over A^*. Then* $[\mathsf{E}]\tau = (L[\mathsf{E}])\,\tau$. ■

cf. also Rem. II.1.1, p. 227

Hence we conclude:

Corollary 4.5 *Two rational expressions* E *and* F *are equivalent if and only if* $L[\mathsf{E}] = L[\mathsf{F}]$. ■

These results justify our giving ourselves the ability to confuse, *by abuse of terminology*, a language defined by rational operations with the corresponding rational expression: for example, $(a+b)^*a\,b\,(a+b)^*$ is both the language of words that contain at least one factor $a\,b$ and a particular expression that denotes it.

Exercises

- 4.1 (a) Give a definition of rational expressions using a postfix notation for the operators $+$, $\cdot$ and $*$. Give the expressions that correspond to the expressions (4.1) with this convention.

  (b) Do the same using prefix notation. (c) Draw conclusions.

- 4.2 Give a rational expression for the language of words that do not contain a factor $b\,a$.

4.2 Rational identities

Corollary 4.5 and Theorem 4.1 ensure that we can decide whether two rational expressions over a set of variables $\mathbf{X}$ are equivalent. We know in any case that this decision process implies the transcription of expressions into automata and assumes that the latter are determinised if necessary. Thus we are led to ask if we can determine whether two expressions are equivalent directly, working with the expressions themselves, without resorting to automata.

cf. Rem. 4.1, p. 127

We have already said that no other proof is known of the equivalence of expressions than via automata, and hence that the answer to the question above is negative. It is nonetheless interesting to try to solve it, as much to understand the properties that we can prove in this direction as to grasp the reasons that prevent us from

arriving at a complete solution and which, *a contrario*, shed new light on our appeal to automata.[40]

If we try to analyse the reasons why two expressions can be equivalent – that is, can denote the same language – there are some cases that seem obvious. For example, for all expressions E and F, we have

$$\mathsf{E}+\mathsf{F} \quad \equiv \quad \mathsf{F}+\mathsf{E} \tag{4.8}$$

since $|\mathsf{E}| \cup |\mathsf{F}| = |\mathsf{F}| \cup |\mathsf{E}|$: that is, since union is commutative. If we return to the two equivalences of Example 4.1:

$$(a^* \cdot b)^+ \cdot a^+ + a^* \quad \equiv \quad (a^* \cdot b)^* \cdot a^+ + 1_{A^*} \quad \equiv \quad (b^* \cdot a)^* \ ,$$

the first does not seem much more complicated than the previous example: it suffices to write $a^* \equiv 1_{A^*} + a^+$, to take out the factor a^+ and use the same identity, in the opposite direction (and for $u = a^* \cdot b$), to go from the left-hand expression to that in the middle. The second example is not so simple. And what of the identity

$$(a + b\,(a\,b^*a)^*b)^* \ \equiv \ a^* + a^*b\,(b\,a^*b)^*b\,a^* + a^*b\,(b\,a^*b)^*a\,(b + a\,(b\,a^*b)^*a)^*a\,(b\,a^*b)^*b\,a^* \ ,$$

obtained by two distinct computations[41] of the language recognised by $\mathcal{P}_2$, the divider by 3 ?

Our objective in this subsection is to formalise this intuition with a mathematical rigour that allows us to state some properties and prove them. An identity between rational expressions, or *rational identity*, is an equation such as (4.8) which expresses the equivalence of two expressions over a set of variables (the expressions themselves being variables written E, F, *etc.*). In general, an equivalence corresponds to a property of the interpretation of the expressions. We can then reuse identities by substituting them into other identities to obtain new equivalences.

4.2.1 Classical identities

Trivial and natural identities. The first series of identities expresses the fact that expressions will be *interpreted in a semiring* – that is, that $+$ and $\cdot$, $\mathsf{0}$ and $\mathsf{1}$ will satisfy the corresponding axioms: $+$ and $\cdot$ are associative, $+$ is commutative, $\mathsf{0}$ is an identity element for $+$ and a zero for $\cdot$, $\mathsf{1}$ is an identity element for $\cdot$, and $\cdot$ is left and right distributive over $+$:

$$\mathsf{E}+\mathsf{0} \equiv \mathsf{0}+\mathsf{E} \equiv \mathsf{E} \ , \qquad \mathsf{E}\cdot\mathsf{0} \equiv \mathsf{0}\cdot\mathsf{E} \equiv \mathsf{0} \ , \qquad \mathsf{E}\cdot\mathsf{1} \equiv \mathsf{1}\cdot\mathsf{E} \equiv \mathsf{E} \ , \tag{T}$$

$$(\mathsf{E}+\mathsf{F})+\mathsf{G} \equiv \mathsf{E}+(\mathsf{F}+\mathsf{G}) \quad \text{and} \quad (\mathsf{E}\cdot\mathsf{F})\cdot\mathsf{G} \equiv \mathsf{E}\cdot(\mathsf{F}\cdot\mathsf{G}) \ , \tag{A}$$

$$\mathsf{E}\cdot(\mathsf{F}+\mathsf{G}) \equiv \mathsf{E}\cdot\mathsf{F}+\mathsf{E}\cdot\mathsf{G} \quad \text{and} \quad (\mathsf{E}+\mathsf{F})\cdot\mathsf{G} \equiv \mathsf{E}\cdot\mathsf{G}+\mathsf{F}\cdot\mathsf{G} \ , \tag{D}$$

$$\mathsf{E}+\mathsf{F} \ \equiv \ \mathsf{F}+\mathsf{E} \ . \tag{C}$$

[40] This ambitious programme will not be fully realised in the course of this work.

[41] As the conscientious reader who has done Exercise 2.15 will remember.

The identities in the first line (**T**) may be described as trivial. Other than being obvious, they have the particular property that if, in any expression E, one of the six terms on the left or in the middle is replaced by the corresponding right-hand term, iteratively, until all occurrences of the six terms have been eliminated, a *unique* equivalent expression is obtained (that is, independent of the order in which the replacements were made) which is effectively computable,[42] which we call a *reduced expression*. From now on, all computations on expressions will be performed modulo reduction by trivial identities without further mention.

The three other lines of identities **A**, **D** and **C** are no less obvious, since they reflect the properties of operations in a semiring. With the foregoing they are called *natural identities*. They are also often used without explicit mention (but not always). In particular, we allow ourselves from now on to write $\mathsf{E}+\mathsf{F}+\mathsf{G}$ and $\mathsf{E}\cdot\mathsf{F}\cdot\mathsf{G}$. (Still aiming to lighten the notation, we will often replace the $\cdot$ operator by simple juxtaposition and write $\mathsf{E}\,\mathsf{F}\,\mathsf{G}$ for $\mathsf{E}\cdot\mathsf{F}\cdot\mathsf{G}$.) Equally, we can and shall be led to distinguish the associativity of the sum $\mathbf{A_s}$ from that of the product $\mathbf{A_p}$.

cf. Rem. 4.5, p. 140

The identities that really matter are those that express, or at least reflect, the properties of the operator $*$ and those of the semiring $\mathfrak{P}(A^*)$, or a combination of the two, that is, of the operator $*$ in $\mathfrak{P}(A^*)$.

cf. p. 85

Aperiodic and cyclic identities. The relation (2.5), which we have been able to see since the definition of the star operation, can now be written as a pair of identities:

$$\mathsf{E}^* \equiv 1+\mathsf{E}\cdot\mathsf{E}^* \qquad (\mathbf{U_l})$$
$$\mathsf{E}^* \equiv 1+\mathsf{E}^*\cdot\mathsf{E}\ . \qquad (\mathbf{U_r})$$

By the definition of star, we also have

$$0^* \equiv 1\ . \qquad (\mathbf{Y})$$

The following identities, called *aperiodic identities*,[43] are collected in a statement which requires proof.

Proposition 4.6 *For all rational expressions* E *and* F

$$(\mathsf{E}+\mathsf{F})^* \equiv \mathsf{E}^*\cdot(\mathsf{F}\cdot\mathsf{E}^*)^* \qquad (\mathbf{S_l})$$
$$(\mathsf{E}+\mathsf{F})^* \equiv (\mathsf{E}^*\cdot\mathsf{F})^*\cdot\mathsf{E}^* \qquad (\mathbf{S_r})$$
$$(\mathsf{E}\cdot\mathsf{F})^* \equiv 1+\mathsf{E}\cdot(\mathsf{F}\cdot\mathsf{E})^*\cdot\mathsf{F}\ . \qquad (\mathbf{P})$$

Proof. Corollary 4.5 implies that an identity between two expressions is true if and only if the free interpretations of these expressions are equal: that is, if the languages that they denote when the symbols E, F, *etc.* are replaced by distinct letters from an alphabet A are equal.

[42] By a simple algorithm, linear in the size of the expression E.

[43] This term was proposed by D. Krob, and is justified in the bibliographic notes p. 215.

To prove $\mathbf{S_l}$, we therefore need to show that $(a+b)^* = a^*(ba^*)^*$. Arden's Lemma tells us that $(a+b)^*$ is the unique solution of *Lem. 2.9, p. 100*

$$X = (a+b)\,X + 1_{A^*} \quad .$$

We can verify using $\mathbf{U_l}$ (and the natural identities), that $a^*(ba^*)^*$ is a solution of the same equation:

$$\begin{aligned}(a+b)\big(a^*(ba^*)^*\big) + 1_{A^*} &= a\,a^*(ba^*)^* + \underbrace{b\,a^*(ba^*)^* + 1_{A^*}}_{\text{by } (\mathbf{U_l})} \\ &= a\,a^*(ba^*)^* + (ba^*)^* = \underbrace{(a\,a^* + 1_{A^*})}_{\text{by } (\mathbf{U_l})}(ba^*)^* = a^*(ba^*)^* \quad .\end{aligned}$$

A symmetric calculation shows that $(a^*b)^*a^*$ is a solution of $X = X\,(a+b) + 1_{A^*}$, by which $\mathbf{S_r}$. Analogously, the language $1_{A^*} + a\,(ba)^*b$ is a solution of $X = a\,b\,X + 1_{A^*}$:

$$\begin{aligned}a\,b\,(1_{A^*} + a\,(ba)^*b) + 1_{A^*} &= a\,b + a\,b\,a\,(ba)^*b + 1_{A^*} \\ &= a\,\underbrace{(1_{A^*} + b\,a\,(ba)^*)}_{\text{by } (\mathbf{U_l})}\,b + 1_{A^*} = a\,(ba)^*b + 1_{A^*} \quad .\end{aligned}$$

This proves $\mathbf{P}$. ■

We will use the aperiodic identities in the next subsection to compare the different methods of computing the language accepted by a finite automaton. The following proposition introduces an infinity of identities, indexed by $\mathbb{N}$, and called *cyclic identities.*

Proposition 4.7 *For every rational expression* E*, and for every* n *in* $\mathbb{N}$*, we have*[44]

$$\mathsf{E}^* \quad \equiv \quad \mathsf{E}^{<\mathsf{n}} \cdot (\mathsf{E}^{\mathsf{n}})^* \quad . \qquad (\mathbf{Z}_n)$$

Proof. We reuse the method of the proof of Proposition 4.6. For each n, $a^{<n}(a^n)^*$ is a solution of $X = a\,X + 1_{A^*}$, since

$$\begin{aligned}a\,\left(a^{<n}(a^n)^*\right) + 1_{A^*} &= [a + \cdots + a^{n-1} + \underbrace{a^n](a^n)^* + 1_{A^*}}_{\text{by } (\mathbf{U_l})} \\ &= [a + \cdots + a^{n-1}]\,(a^n)^* + (a^n)^* \\ &= [1_{A^*} + a + \cdots + a^{n-1}](a^n)^* \quad .\end{aligned}$$

■

Idempotent identities. Union, which is addition in $\mathfrak{P}(A^*)$, is idempotent, which implies that star is also an idempotent operation in $\mathfrak{P}(A^*)$. These observations give us two more identities:

$$\mathsf{E} + \mathsf{E} \quad \equiv \quad \mathsf{E} \qquad (\mathbf{I})$$

$$(\mathsf{E}^*)^* \quad \equiv \quad \mathsf{E}^* \quad . \qquad (\mathbf{J})$$

[44] Admitting that $\mathsf{E}^{<\mathsf{n}}$ is a suitable abbreviation for the expression $1 + \mathsf{E} + \mathsf{E}^2 + \cdots + \mathsf{E}^{\mathsf{n}-1}$.

We shall call the collection of identities from **T** to **J** *classical identities.*[45]

4.2.2 A formal computation

From one or more identities we can prove new ones. We note for example that $\mathbf{U_l}$ and $\mathbf{U_r}$ are obtained by replacing F by $\mathsf{1}$ (resp. E by $\mathsf{1}$) in **P**.[46] Another example is the second equivalence in Example 4.1:

$$\begin{aligned}(a^*b)^*a^+ + 1_{A^*} &\equiv \underbrace{(a^*b)^*a^*}_{\text{by } (\mathbf{S_r})} a + 1_{A^*} \equiv (a+b)^*a + 1_{A^*} \\ &\equiv \underbrace{(b+a)^*}_{\text{by } (\mathbf{S_r})} a + 1_{A^*} \equiv \underbrace{(b^*a)^*b^*a + 1_{A^*}}_{\text{by } (\mathbf{U_l})} \equiv (b^*a)^* \quad .\end{aligned}$$

Notation To be able to express the degree of similarity of two expressions, we will use a notation familiar to logicians and write

$$\mathbf{X} \vdash \qquad \mathsf{E} \quad \equiv \quad \mathsf{F}$$

to mean that the identity $\mathsf{E} \equiv \mathsf{F}$ can be deduced from the identity, or set of identities, **X**. The preceding identity could therefore be written

$$\mathbf{S_r} \wedge \mathbf{U_l} \vdash \qquad (a^*b)^*a^+ + 1_{A^*} \equiv (b^*a)^*$$

(the symbol $\wedge$ denoting the conjunction of identities). We will write $\mathbf{U} = \mathbf{U_l} \wedge \mathbf{U_r}$ and $\mathbf{S} = \mathbf{S_l} \wedge \mathbf{S_r}$.

Remark 4.2 To be rigorous, we would have to specify that the above identities are valid for expressions interpreted in $\mathfrak{P}(A^*)$.

cf. Prop. III.2.5, p. 396 and Prop. III.5.3, p. 473

The proof of Propositions 4.6 and 4.7 show that the aperiodic and cyclic identities are satisfied when the operator $*$ and the semiring in which the expressions are interpreted satisfy the identities **U** and Arden's Lemma. □

We could continue our consideration of axioms and identities by reversing the question, and seeking to describe sets of axioms which allow us to find the identity of all equivalent expressions (complete systems) and, among them, the minimal sets (independent systems).

This is a rich area of study, replete with often difficult results, which we shall not explore. The most important point is that every complete axiomatic system is necessarily infinite, and that if we allow, as well as substitution, a stronger inference rule deduced from Arden's Lemma, then we can build a complete finite system (see bibliographic notes).

---

[45] These are the same as what J. H. Conway called *classical axioms* in his book *Regular Algebra and Finite Machines* [66].

[46] But this would only really have been interesting if **P** had been proved without using $\mathbf{U_l}$.

Exercises

• 4.3 Prove the following identities using the classical identities (but without Lemma 2.9).

(a) $(a+b)^* = a^* + a^*b\,(a+b)^*$; (b) $(a+b)^* = a^*(b^+a^+)^*b^*$;

(c) $(a+b)^+ = a^+ + a^*(b\,a^*)^+$.

• 4.4 Do the same for:

(a) $(a+b+c)^* = [a+b+c\,(c+b)^*a]^*[1+c\,(c+b)^*]$;

(b) $(x+y)^* = [x+y\,(y^+x)^*x]^*[1+y\,(y^+x)^*y^*]$. [Hint: use (a) to show (b).]

4.5 $\mathbf{P} \vdash \quad (\mathsf{E}\,\mathsf{F})^*\mathsf{E} \quad \equiv \quad \mathsf{E}\,(\mathsf{F}\,\mathsf{E})^*$.

4.6 Show that, for all integers n and m, $\mathbf{Z}_n \wedge \mathbf{Z}_m \vdash \quad \mathbf{Z}_{n\,m}$.

4.3 Expressions for the behaviour of a finite automaton

In Section 2 we described three methods, distinct in their formulation at least, for computing the language recognised by a finite automaton $\mathcal{A}$. We can now be more precise and say that we described three distinct methods for computing *an expression that denotes* $L(\mathcal{A})$, the computation and its result, depending, in each case, on a total order assigned to the states of $\mathcal{A}$. We also said that these three methods are three presentations of a *single proof*; we now have concepts adequate for justifying this assertion. We will now make explicit the degree of relationship between the various expressions that can be obtained by these methods and by the different ways of putting them into practice. In doing this we describe syntactic procedures that allow us to turn one such expression into another.

For the rest of this subsection let $\mathcal{A} = \langle\, Q, A, E, I, T \,\rangle$ be a finite automaton and ω a total order on Q which fixes the operation of the algorithms that calculate $L(\mathcal{A})$.

4.3.1 The state elimination and equation solution methods

Let us begin by verifying the exact identity of the state elimination method and the solution of a system of linear equations taken from the automaton (solution by Gaussian elimination). Recall the notation of Section 2 for writing this system: for p and q in Q, the *set of words* that are the label of a computation which goes from p to a final state of $\mathcal{A}$ is written *cf. p. 99*

$$L_p = \{f \mid \exists t \in T \quad p \xrightarrow[\mathcal{A}]{f} t\} \ ,$$

and we write $E_{p,q}$ for the *set of labels of transitions* that go from p to q and the symbol $\delta_{p,R}$ for a subset R of Q; this equals 1_{A^*} if p is in R and $\emptyset$ if not. The system of equations associated with $\mathcal{A}$ is written

$$L(\mathcal{A}) = \sum_{p\in I} L_p \qquad = \sum_{p\in Q} \delta_{p,I}\, L_p \tag{4.9}$$

$$\forall p \in Q \qquad L_p = \sum_{q\in Q} E_{p,q}\, L_q + \delta_{p,T} \ . \tag{4.10}$$

After the elimination of a certain number of unknowns L_p – we write Q' for the set of indices of those which have not been eliminated – we obtain a system of the form

$$L(\mathcal{A}) = \sum_{p \in Q'} G_p L_p + H \tag{4.11}$$

$$\forall p \in Q' \qquad L_p = \sum_{q \in Q'} F_{p,q} L_q + K_p \ . \tag{4.12}$$

We can make a generalised automaton $\mathcal{B}'$ corresponding to such a system, whose set of states is $Q' \cup \{i, t\}$, where i and t do not belong to Q', and such that, for all p and q in Q':

- the transition from p to q is labelled $F_{p,q}$;
- the transition from p to t is labelled K_p;
- the transition from i to p is labelled G_p;
- and the transition from i to t is labelled H.

cf. p. 97 Note that this definition applied to the system (4.9)–(4.10) characterises the automaton constructed in the first phase of the state elimination method applied to $\mathcal{A}$.

The elimination in the system (4.11)–(4.12) of the unknown L_p by substitutions and the application of Arden's Lemma give the system

$$L(\mathcal{A}) = \sum_{r \in Q' \setminus p} \left[G_r + G_p F^*_{p,p} F_{p,r}\right] L_r + \left[H + G_p F^*_{p,p} K_p\right] \tag{4.13}$$

$$\forall r \in Q' \setminus p \qquad L_r = \sum_{q \in Q' \setminus p} \left[F_{r,q} + F_{r,p} F^*_{p,p} F_{r,q}\right] L_q + \left[K_r + F_{r,p} F^*_{p,p} K_p\right] \ , \tag{4.14}$$

whose coefficients are exactly the transition labels of the generalised automaton obtained by removing the state p from $\mathcal{B}'$.

Thus, since the starting points correspond and since each step maintains the correspondence, the expression obtained for $L(\mathcal{A})$ by the state elimination method is the same as that obtained by the solution of the system (4.9)–(4.10). More precisely, we can say that the state elimination method, or BMC algorithm, reproduces in the automaton $\mathcal{A}$ the computations corresponding to the solution of the system.

4.3.2 The BMC and MNY algorithms, identical orders

The order ω fixes the operation of the state elimination method whose result is a rational expression over A^*, written[47] $\mathsf{E}_{\mathrm{BMC}}(\mathcal{A}, \omega)$. For greater precision, we write the result of this algorithm $\mathsf{E}_{\mathrm{BMC}}(\mathcal{A}, \omega, (p, q))$ when we take p as the initial state and q as the final state.

On the other hand, we will write $\mathsf{M}_{\mathrm{MNY}}(\mathcal{A}, \omega)$ for the matrix of rational expressions obtained when we apply the McNaughton–Yamada algorithm to the automaton $\mathcal{A}$ whose states are ordered by ω. It then follows that:

[47] A reminder that this algorithm is due to J. Brzozowski and E. McCluskey.

Proposition 4.8 *Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ an automaton over A^*. For every (total) order ω on Q and all p and q in Q, we have*

$$\mathbf{U} \vdash \qquad [\mathsf{M}_{MNY}(\mathcal{A}, \omega)]_{p,q} \quad \equiv \quad \mathsf{E}_{BMC}(\mathcal{A}, \omega, (p, q)) \ .$$

Proof. This result is not so surprising: to prove it, we will show a correspondence between the operations performed by the two algorithms. The difficulty, if it can be called that, is that we have to compare two objects whose form and mode of construction are rather different: on one hand a $Q \times Q$ matrix obtained by successive transformations, from which we choose one entry; and on the other an expression obtained by repeated modification of an automaton, hence of a matrix, but one whose size decreases at each step.

In the following, $\mathcal{A}$ and ω are fixed and remain implicit. The automaton $\mathcal{A}$ has n states, identified by integers from 1 to n; the two algorithms perform n steps starting in a situation called 'step 0', the kth step of the state elimination method consisting of the removal of state k, and that of the MNY algorithm consisting of calculating the labels of paths that do not include nodes (strictly) greater than k. We write

$$\mathsf{E}^{(k)}(r, s)$$

for the label of the transition from r to s in the automaton obtained from $\mathcal{A}$ (and ω) at the kth step of the state elimination method; necessarily, in this notation, $k + 1 \leqslant r$ and $k + 1 \leqslant s$ (abbreviated to $k + 1 \leqslant r, s$). We write

$$\mathsf{M}^{(k)}_{r,s}$$

for the entry r, s of the $n \times n$ matrix computed by the kth step of the MNY algorithm. At step 0, the automaton $\mathcal{A}$ has not been modified and we have

$$\forall r, s\,,\ 1 \leqslant r, s \leqslant n \qquad \mathsf{M}^{(0)}_{r,s} = \mathsf{E}^{(0)}(r, s) \ , \tag{4.15}$$

which will be the base case of the inductions to come. Algorithm MNY is written:

$$\forall k\,,\ 0 < k \leqslant n\,,\ \forall r, s\,,\ 1 \leqslant r, s \leqslant n$$
$$\mathsf{M}^{(k)}_{r,s} = \mathsf{M}^{(k-1)}_{r,s} + \mathsf{M}^{(k-1)}_{r,k} \cdot [\mathsf{M}^{(k-1)}_{k,k}]^* \cdot \mathsf{M}^{(k-1)}_{k,s} \ . \tag{4.16}$$

The state elimination algorithm is written

$$\forall k\,,\ 0 < k \leqslant n\,,\ \forall r, s\,,\ k < r, s \leqslant n$$
$$\mathsf{E}^{(k)}(r, s) = \mathsf{E}^{(k-1)}(r, s) + \mathsf{E}^{(k-1)}(r, k) \cdot [\mathsf{E}^{(k-1)}(k, k)]^* \cdot \mathsf{E}^{(k-1)}(k, s) \ . \tag{4.17}$$

Hence we conclude, for given r and s and by induction on k, that

$$\forall r, s\,,\ 1 \leqslant r, s \leqslant n\,,\ \forall k\,,\ 0 \leqslant k < \mathsf{min}(r, s) \qquad \mathsf{M}^{(k)}_{r,s} = \mathsf{E}^{(k)}(r, s) \ . \tag{4.18}$$

We see in fact (as there is, even so, something to see) that if $k < \mathsf{min}(r, s)$ then all integer triples (l, u, v) such that $\mathsf{M}^{(l)}_{u,v}$ occurs in the computation of $\mathsf{M}^{(k)}_{r,s}$ by the (recursive) use of (4.16), are such that $l < \mathsf{min}(u, v)$.

Suppose now that we have p and q, also fixed, such that $1 \leqslant p < q \leqslant n$ (the other cases are dealt with similarly). We call the initial and final states added to $\mathcal{A}$, in the first phase of the state elimination method, i and t respectively; i and t are not integers between 1 and n. The transition from i to p and that from q to t are labelled 1_{A^*}. Now let us consider step p of each algorithm. For every state s, $p < s$, $\mathsf{M}^{(p)}_{p,s}$ is given by (4.16):

cf. the proof of Prop. 2.8, p. 97

$$\mathsf{M}^{(p)}_{p,s} = \mathsf{M}^{(p-1)}_{p,s} + \mathsf{M}^{(p-1)}_{p,p} \cdot [\mathsf{M}^{(p-1)}_{p,p}]^* \cdot \mathsf{M}^{(p-1)}_{p,s} \quad ,$$

and $\mathsf{E}^{(p)}(i,s)$ by

$$\mathsf{E}^{(p)}(i,s) = [\mathsf{E}^{(p-1)}(p,p)]^* \cdot \mathsf{E}^{(p-1)}(p,s) \quad ;$$

and hence, by (4.18)

$$\forall s\,,\ p < s \leqslant n \qquad \mathbf{U} \vdash \qquad \mathsf{M}^{(p)}_{p,s} \equiv \mathsf{E}^{(p)}(i,s) \quad . \tag{4.19}$$

Next we consider the steps following p (and row p of the matrices $\mathsf{M}^{(k)}$). For all k, $p < k$, and all s, $k < s \leqslant n$, $\mathsf{M}^{(k)}_{p,s}$ is still computed by (4.16) and $\mathsf{E}^{(k)}(i,s)$ by

$$\mathsf{E}^{(k)}(i,s) = \mathsf{E}^{(k-1)}(i,s) + \mathsf{E}^{(k-1)}(i,k) \cdot [\mathsf{E}^{(k-1)}(k,k)]^* \cdot \mathsf{E}^{(k-1)}(k,s) \quad . \tag{4.20}$$

From (4.19), and based on an observation analogous to the previous one, we conclude from the term-by-term correspondence of (4.16) and (4.20) that

$$\forall k\,,\ p < k\,,\ \forall s\,,\ p < s \leqslant n \qquad \mathbf{U} \vdash \qquad \mathsf{M}^{(k)}_{p,s} \equiv \mathsf{E}^{(k)}(i,s) \quad . \tag{4.21}$$

The analysis of step q gives a similar, and symmetric, result to that which we have just obtained from the analysis of step p: for all r, $q < r$, we have

$$\mathsf{M}^{(q)}_{r,q} = \mathsf{M}^{(q-1)}_{r,q} + \mathsf{M}^{(q-1)}_{r,q} \cdot [\mathsf{M}^{(q-1)}_{q,q}]^* \cdot \mathsf{M}^{(q-1)}_{q,q}$$

and

$$\mathsf{E}^{(q)}(r,t) = \mathsf{E}^{(q-1)}(r,q) \cdot [\mathsf{E}^{(q-1)}(q,q)]^*$$

and hence

$$\forall r\,,\ q < r \leqslant n \qquad \mathbf{U} \vdash \qquad \mathsf{M}^{(q)}_{r,q} \equiv \mathsf{E}^{(q)}(r,t) \quad . \tag{4.22}$$

The steps following q give rise to an equation symmetric to (4.21) (for column q of the matrices $\mathsf{M}^{(k)}$):

$$\forall k\,,\ q < k\,,\ \forall r\,,\ q < r \leqslant n \qquad \mathbf{U} \vdash \qquad \mathsf{M}^{(k)}_{r,q} \equiv \mathsf{E}^{(k)}(r,t) \quad . \tag{4.23}$$

Finally, from

$$\mathsf{M}^{(k)}_{p,q} = \mathsf{M}^{(k-1)}_{p,q} + \mathsf{M}^{(k-1)}_{p,k} \cdot [\mathsf{M}^{(k-1)}_{k,k}]^* \cdot \mathsf{M}^{(k-1)}_{k,q}$$

and

$$\mathsf{E}^{(k)}(i,t) = \mathsf{E}^{(k-1)}(i,t) + \mathsf{E}^{(k-1)}(i,k) \cdot [\mathsf{E}^{(k-1)}(k,k)]^* \cdot \mathsf{E}^{(k-1)}(k,t) \quad ,$$

Equations (4.18), (4.21) and (4.23) together allow us to conclude, by induction on k, that

$$\forall k\,,\ q \leqslant k \leqslant n \qquad \mathbf{U} \vdash \qquad \mathsf{M}^{(k)}_{p,q} \equiv \mathsf{E}^{(k)}(i,t) \quad . \tag{4.24}$$

When we reach $k = n$ in this equation we obtain the identity we want. ■

4.3.3 The BMC and MNY algorithms, distinct orders

Having compared the state elimination method and algorithm MNY, we can compare the results of these algorithms for different execution conditions.

Theorem 4.2 *Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton over A^*. The expressions denoting $L(\mathcal{A})$ computed by the McNaughton–Yamada algorithm, like those computed by the state elimination method or the solution of a system of equations, are all equivalent modulo* **S** *and* **P**, *that is, for all orders ω and ω' on Q and all p and q in Q*

$$\begin{array}{lcl} \mathbf{S} \wedge \mathbf{P} \quad \vdash & [\mathsf{M}_{MNY}(\mathcal{A}, \omega)]_{p,q} \equiv & \left[\mathsf{M}_{MNY}(\mathcal{A}, \omega')\right]_{p,q} \ , \\ \mathbf{S} \wedge \mathbf{P} \quad \vdash & \mathsf{E}_{BMC}(\mathcal{A}, \omega, (p,q)) \equiv & \mathsf{E}_{BMC}(\mathcal{A}, \omega', (p,q)) \ . \end{array}$$

Proof. The previous proposition allows us to show the property for expressions computed by the state elimination method, which is easier to deal with (remembering that $\mathbf{P} \vdash \mathbf{U}$). Furthermore, we can go from an order ω to any other order ω', a permutation of Q, by a series of transpositions.

We therefore arrive at the situation illustrated in Figure 4.1(a) and need to show that the expressions obtained by the state elimination method when we first remove the state r and then r' are equivalent to those obtained from removing first r' and then r, modulo $\mathbf{S} \wedge \mathbf{P}$.

The removal of state r gives the expressions in Figure 4.1(b). The removal of state r' gives the expression

$$\mathsf{E} = K\,L^*H + (K\,L^*G + K')\,\left[G'\,L^*G + L'\right]^*(G'\,L^*H + H') \ ,$$

which using **S** (and the natural identities) becomes

$$\begin{aligned} \mathsf{E} \equiv K\,L^*H &+ K\,L^*G\,\left[L'^*G'\,L^*G\right]^* L'^*G'\,L^*H \\ &+ K'\,\left[L'^*G'\,L^*G\right]^* L'^*G'\,L^*H \\ &\quad + K\,L^*G\,\left[L'^*G'\,L^*G\right]^* L'^*H' + K'\,\left[L'^*G'\,L^*G\right]^* L'^*H' \ . \end{aligned}$$

We write

$$K'\,\left[L'^*G'\,L^*G\right]^* L'^*H' \equiv K'\,L'^*H' + K'\,L'^*G'\,L^*\left[G\,L'^*G'\,L^*\right]^* G\,L'^*H'$$

by using **P** then, by 'switching the brackets' (using the identity $(\mathsf{X}\mathsf{Y})^*\mathsf{X} \equiv \mathsf{X}(\mathsf{Y}\mathsf{X})^*$ which is also a consequence of **P**), we obtain *cf. Exer. 4.5*

$$\begin{aligned} \mathsf{E} \equiv K\,L^*H &+ K\,L^*G\,\left[L'^*G'\,L^*G\right]^* L'^*G'\,L^*H \\ &+ K'\,L'^*G'\,\left[L^*G\,L'^*G'\right]^* L^*H + K\,L^*G\,\left[L'^*G'\,L^*G\right]^* L'^*H' \\ &\quad + K'\,L'^*G'\,\left[L^*G\,L'^*G'\right]^* L^*G\,L'^*H' + K'\,L'^*H' \end{aligned}$$

an expression that is perfectly symmetric in the letters with and without primes, which shows that we would have obtained the same result if we had started by removing r' then r. ∎

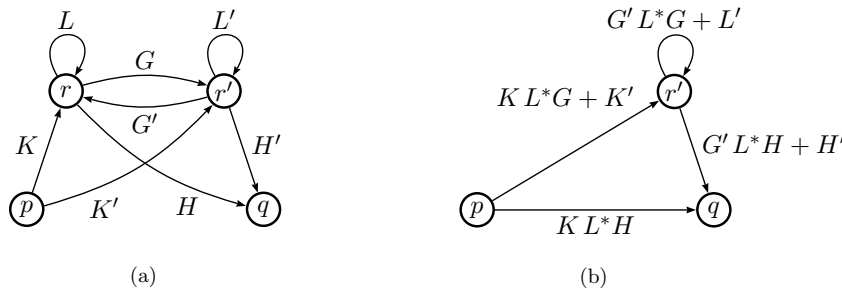


Figure 4.1: A step of the state elimination method

Remark 4.3 We conclude in particular from Theorem 4.2 that, since the identities **P** and **S** were obtained without using the idempotence of union in $\mathfrak{P}(A^*)$, the identities deduced from the computation of the language recognised by a fixed automaton by the state elimination method (or by algorithm MNY), with two different orders on the states, are valid even if the expressions are interpreted in a semiring that is not necessarily idempotent.

This property is also the consequence of the more general result, which we shall see in Chapter III, that a rational expression obtained from an automaton describes not only the set of words accepted but also the *weight* with which each word is accepted in the automaton. □

cf. also Prop. II.1.11, p. 230

4.4 Derivation of expressions

A (rational) expression denotes a (rational) language; these two notions describe the same object at two levels: let us say that the language is the reality and that the expression is the symbolic, or syntactic, representation of this reality. In this subsection we will bring to the level of expressions a technique which has been very successful with languages: taking the quotient.

This construction, which we will call *derivation* to distinguish the two approaches, is of a different nature, but ensures a perfect parallel between the two levels: the quotient of a language L by a word f is a language, $f^{-1}L$, while the derivative with respect to f of an expression E that denotes L is an expression that denotes $f^{-1}L$; a rational language has only a finite number of quotients, while a rational expression, modulo certain conventions, has only a certain number of derivatives; the quotients of L are the states of the minimal automaton that recognises L, while the derivatives of E are the states of a deterministic automaton that recognises the language denoted by E. The interest of this construction is as much practical as theoretical: we shall see that, with suitable alterations, it forms the basis of powerful algorithms for computing an automaton from an expression.

cf. Sec. 5.2, p. 149

4.4.1 Derivatives of an expression

We begin by defining the concept corresponding to the *constant term of a language.*

Definition 4.5 The *constant term* of a rational expression E over A, written $\mathsf{c}(\mathsf{E})$, is the function $\mathsf{1}$ or $\mathsf{0}$, computed by induction on the depth of E by the following formulas:

$$\begin{aligned} \mathsf{c}(\mathsf{1}) = \mathsf{1} \ , \qquad \mathsf{c}(\mathsf{0}) = \mathsf{c}(a) = \mathsf{0} \qquad \forall a \in A \ , \\ \mathsf{c}(\mathsf{E}+\mathsf{F}) = \mathsf{c}(\mathsf{E}) + \mathsf{c}(\mathsf{F}) \ , \qquad \mathsf{c}(\mathsf{E}\cdot\mathsf{F}) = \mathsf{c}(\mathsf{E})\cdot\mathsf{c}(\mathsf{F}) \qquad \text{and} \qquad \mathsf{c}(\mathsf{E}^*) = \mathsf{1} \ . \end{aligned} \tag{4.25}$$

We can easily verify that this definition makes sense:

Property 4.1 *The* constant term *of an expression* E *is* $\mathsf{1}$ *or* $\mathsf{0}$ *according to whether* 1_{A^*} *is in the language denoted by* E *or not.*[48] ■

Remark 4.4 It might seem simpler to take the statement of Property 4.1 as a definition and then verify the formulas (4.25) which would then become properties. It is not. The formulas (4.25) enable us to *compute effectively* $\mathsf{c}(\mathsf{E})$ – and hence $\mathsf{c}(|\mathsf{E}|)$ – from E: it is $|\mathsf{E}|$ which is obtained from E, not the other way around. □

Definition 4.6 Let E be a rational expression over A and a a letter of A. The *derivative* of E with respect to a, written $\frac{\partial}{\partial a}\mathsf{E}$, is a rational expression over A, defined recursively by the following formulas:

$$\frac{\partial}{\partial a}\mathsf{0} = \frac{\partial}{\partial a}\mathsf{1} = \frac{\partial}{\partial a} b = \mathsf{0} \qquad \forall b \in A\,,\ b \neq a$$

$$\frac{\partial}{\partial a} a = \mathsf{1} \tag{4.26}$$

$$\frac{\partial}{\partial a}(\mathsf{E}+\mathsf{F}) = \frac{\partial}{\partial a}\mathsf{E} + \frac{\partial}{\partial a}\mathsf{F} \tag{4.27}$$

$$\frac{\partial}{\partial a}(\mathsf{E}\cdot\mathsf{F}) = \left[\frac{\partial}{\partial a}\mathsf{E}\right]\cdot\mathsf{F} + \mathsf{c}(\mathsf{E})\cdot\frac{\partial}{\partial a}\mathsf{F} \tag{4.28}$$

$$\frac{\partial}{\partial a}\mathsf{E}^* = \left[\frac{\partial}{\partial a}\mathsf{E}\right]\cdot\mathsf{E}^* \ . \tag{4.29}$$

The (vague) resemblance of (4.27) and (4.28) to the formulas for the derivation of a sum or product of functions could be considered as justifying the terminology.

Example 4.3 The derivatives, with respect to a and b, of the expression

$$\mathsf{E}_1 = (a+b)^*\cdot\Big(a\cdot\big(b\cdot(a+b)^*\big)\Big)$$

[48] In fact, it would be more accurate to write $\mathsf{c}(|\mathsf{E}|) = \mathsf{c}(\mathsf{E})$ but we will stick to an elementary version of the constant term of a language (*cf.* Note 25, p. 89).

are obtained by the following calculation:

$$\begin{aligned}
\frac{\partial}{\partial a}\,\mathsf{E}_1 &= \left[\frac{\partial}{\partial a}(a+b)^*\right]\cdot\Big(a\cdot\big(b\cdot(a+b)^*\big)\Big) + \mathsf{c}\,((a+b)^*)\cdot\left[\frac{\partial}{\partial a}\Big(a\cdot\big(b\cdot(a+b)^*\big)\Big)\right] \\
&= \left[\frac{\partial}{\partial a}(a+b)\right]\cdot(a+b)^*\cdot\Big(a\cdot\big(b\cdot(a+b)^*\big)\Big) \\
&\qquad + \left[\frac{\partial}{\partial a}\,a\right]\cdot\big(b\cdot(a+b)^*\big) + \mathsf{c}\,(a)\cdot\left[\frac{\partial}{\partial a}\big(b\cdot(a+b)^*\big)\right] \\
&= \mathsf{E}_1 + \big(b\cdot(a+b)^*\big) \\
\frac{\partial}{\partial b}\,\mathsf{E}_1 &= \left[\frac{\partial}{\partial b}(a+b)^*\right]\cdot\Big(a\cdot\big(b\cdot(a+b)^*\big)\Big) + \mathsf{c}\,((a+b)^*)\cdot\left[\frac{\partial}{\partial b}\Big(a\cdot\big(b\cdot(a+b)^*\big)\Big)\right] \\
&= \left[\frac{\partial}{\partial b}(a+b)\right]\cdot(a+b)^*\cdot\Big(a\cdot\big(b\cdot(a+b)^*\big)\Big) \\
&\qquad + \left[\frac{\partial}{\partial b}\,a\right]\cdot\big(b\cdot(a+b)^*\big) + \mathsf{c}\,(a)\cdot\left[\frac{\partial}{\partial b}\big(b\cdot(a+b)^*\big)\right] = \mathsf{E}_1 \quad . \qquad \square
\end{aligned}$$

These computations call for two remarks.

Remarks 4.5 (i) The calculations were conducted modulo the trivial identities **T**: at each step of the calculation, the expression obtained was replaced by the *equivalent reduced expression.*

cf. p. 130 (ii) The expression E_1 is 'completely bracketed', contrary to the usage which we adopted in Subsection 4.2. The reason is that the derivation of expressions *is not invariant* modulo the associativity of the product operation. $\square$

cf. Exer. 4.10

The derivative of an expression with respect to a letter a is consistent with the (left) quotient of languages by the same letter a since we can easily verify the analogues of the formulas (4.26)–(4.29):

Properties 4.2 $\forall a \in A\,,\ \forall L, K \subseteq A^*$ (i) $a^{-1}\,(L \cup K) = a^{-1}L \,\cup\, a^{-1}K\,,$
(ii) $a^{-1}\,(L\,K) = \big(a^{-1}L\big)\,K \,\cup\, \mathsf{c}\,(L)\,a^{-1}K\,,$ (iii) $a^{-1}\,(L^*) = \big(a^{-1}L\big)\,L^*\,.$ ■

From which we conclude, by induction on the depth of expressions:

Property 4.3 $\forall \mathsf{E} \in \mathsf{Rat}\mathsf{E}\,A\,,\ \forall a \in A \qquad \left|\frac{\partial}{\partial a}\,\mathsf{E}\right| = a^{-1}\,|\mathsf{E}|\,.$ ■

Definition 4.7 Let E be a rational expression over A and g a non-empty word from A^*: that is, $g = f\,a$ with a in A. The *derivative* of E with respect to g, written $\dfrac{\partial}{\partial g}\,\mathsf{E}$, is the rational expression over A, defined recursively by the formulas (4.26)–(4.29) and by

$$\forall f \in A^*\,,\ \forall a \in A \qquad \frac{\partial}{\partial f a}\,\mathsf{E} = \frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}\,\mathsf{E}\right) \quad . \tag{4.30}$$

Example 4.3 (continued) The derivatives of E_1 with respect to words of length 2 are

$$\frac{\partial}{\partial a^2}\mathsf{E}_1 = \frac{\partial}{\partial a}[\mathsf{E}_1 + b(a+b)^*] = \frac{\partial}{\partial a}\mathsf{E}_1 + \frac{\partial}{\partial a}(b(a+b)^*) = \frac{\partial}{\partial a}\mathsf{E}_1 \ ,$$

$$\frac{\partial}{\partial ab}\mathsf{E}_1 = \frac{\partial}{\partial b}[\mathsf{E}_1 + b(a+b)^*] = \frac{\partial}{\partial b}\mathsf{E}_1 + \frac{\partial}{\partial b}(b(a+b)^*) = \mathsf{E}_1 + (a+b)^* \ ,$$

$$\frac{\partial}{\partial ba}\mathsf{E}_1 = \frac{\partial}{\partial a}\mathsf{E}_1 \qquad \text{and} \qquad \frac{\partial}{\partial b^2}\mathsf{E}_1 = \frac{\partial}{\partial b}\mathsf{E}_1 = \mathsf{E}_1 \ . \qquad \square$$

From Property 4.3 we then deduce, by induction on the length of f, the close relationship that exists between the derivative of an expression by f and the (left) quotient of a language by f:

Proposition 4.9 $\forall \mathsf{E} \in \mathsf{Rat}\mathsf{E}\,A\,,\ \forall f \in A^* \qquad \left|\frac{\partial}{\partial f}\mathsf{E}\right| = f^{-1}|\mathsf{E}|\,.$ ■

The set of quotients of a rational language is finite; the same is not true of the set of derivatives of a rational expression, *but it is almost true.* This is what we shall now see.

4.4.2 A theorem of J. Brzozowski

We start by observing that the formulas (4.26)–(4.30) can produce an infinite number of expressions derived from a single expression E.

Example 4.3 (continued) If we calculate the derivatives of E_1 with respect to words of the form $(a\,b)^*$ and $(a\,b)^*a$, we obtain the following expressions:

$$\frac{\partial}{\partial aba}\mathsf{E}_1 = \frac{\partial}{\partial a}[\mathsf{E}_1 + (a+b)^*] = \frac{\partial}{\partial a}\mathsf{E}_1 + \frac{\partial}{\partial a}((a+b)^*) = \mathsf{E}_1 + b(a+b)^* + (a+b)^*\,,$$

$$\frac{\partial}{\partial (ab)^2}\mathsf{E}_1 = \frac{\partial}{\partial b}\mathsf{E}_1 + \frac{\partial}{\partial b}(b(a+b)^*) + \frac{\partial}{\partial b}((a+b)^*) = \mathsf{E}_1 + (a+b)^* + (a+b)^*\,,$$

and more generally

$$\frac{\partial}{\partial (ab)^{n\text{-}1}a}\mathsf{E}_1 = \mathsf{E}_1 + b(a+b)^* + \underbrace{(a+b)^* + \cdots + (a+b)^*}_{n-1 \text{ times}}\,,$$

$$\frac{\partial}{\partial (ab)^n}\mathsf{E}_1 = \mathsf{E}_1 + \underbrace{(a+b)^* + \cdots + (a+b)^*}_{n \text{ times}} \ ;$$

these expressions are all distinct. But we also see that all of them – the $\frac{\partial}{\partial (ab)^n a}\mathsf{E}_1$ on one hand and the $\frac{\partial}{\partial (ab)^n}\mathsf{E}_1$ on the other – are equivalent modulo **I**. This is the property that we will now prove in general. $\square$

Theorem 4.3 *The set of derivatives of a rational expression is finite modulo the identities* $\mathbf{A_s}$, $\mathbf{C}$ *and* $\mathbf{I}$, *and it is effectively computable.*[49]

First, let us prove an elementary property.

Lemma 4.10 *Let* $\mathcal{F}$ *be a finite set of expressions. The set of expressions formed by sums of elements of* $\mathcal{F}$ *is finite modulo* $\mathbf{A_s CI}$ *and its cardinal is bounded by* $2^{\|\mathcal{F}\|}$.

Proof. In any sum of elements of $\mathcal{F}$, we use the commutativity and associativity of the sum to regroup identical expressions, which are the same modulo $\mathbf{I}$. There is therefore a bijection between the sums of elements of $\mathcal{F}$, where each element appears at most once, and the subsets of $\mathcal{F}$. ■

Proof of Theorem 4.3. Write $\mathbf{nd}(\mathsf{E})$ for the number of distinct derivatives of E modulo $\mathbf{A_s}$, $\mathbf{C}$, and $\mathbf{I}$. We shall show by induction on the depth of expressions that $\mathbf{nd}(\mathsf{E})$ is finite.

(a) Atoms. If E is an atomic expression, $\mathbf{nd}(\mathsf{E})$ is either 1 or 2.

(b) Sum. Formulas (4.30) and (4.27) imply

$$\forall f \in A^* \qquad \frac{\partial}{\partial f}(\mathsf{E}+\mathsf{F}) = \frac{\partial}{\partial f}\,\mathsf{E} + \frac{\partial}{\partial f}\,\mathsf{F} \ .$$

There are $\mathbf{nd}(\mathsf{E})$ possible derivatives for E and $\mathbf{nd}(\mathsf{F})$ possible derivatives for F, so

$$\mathbf{nd}(\mathsf{E}+\mathsf{F}) \leqslant \mathbf{nd}(\mathsf{E})\,\mathbf{nd}(\mathsf{F}) \ .$$

The inequality comes from the fact that some of these combinations may be identical.

(c) Product. Let $f = a_1\,a_2 \cdots a_n$ be a word of length n. By expanding (4.30), we obtain

$$\frac{\partial}{\partial a_1 a_2 \cdots a_n}(\mathsf{E}\cdot\mathsf{F}) = \left[\frac{\partial}{\partial a_1 a_2 \cdots a_n}(\mathsf{E})\right]\cdot\mathsf{F} + \mathsf{c}\left(\frac{\partial}{\partial a_1 a_2 \cdots a_{n\text{-}1}}(\mathsf{E})\right)\cdot\frac{\partial}{\partial a_n}\,\mathsf{F} + \\ \mathsf{c}\left(\frac{\partial}{\partial a_1 a_2 \cdots a_{n\text{-}2}}(\mathsf{E})\right)\cdot\frac{\partial}{\partial a_{n\text{-}1} a_n}\,\mathsf{F} + \cdots + \mathsf{c}\,(\mathsf{E})\cdot\frac{\partial}{\partial a_1 a_2 \cdots a_n}\,\mathsf{F} \ .$$

Thus, $\frac{\partial}{\partial f}(\mathsf{E}\cdot\mathsf{F})$ is the sum of $\left(\frac{\partial}{\partial f}\,\mathsf{E}\right)\cdot\mathsf{F}$ and at most $|f|$ derivatives of F. By Lemma 4.10 this sum, with its arbitrarily large number of terms, must, by the inductive hypothesis that $\mathbf{nd}(\mathsf{F})$ is finite, be bounded, and we have

$$\mathbf{nd}(\mathsf{E}\cdot\mathsf{F}) \leqslant \mathbf{nd}(\mathsf{E})\,2^{\mathbf{nd}(\mathsf{F})} \ .$$

[49] We do not mention the trivial identities $\mathbf{T}$ explicitly since in any case every computation on derivatives is performed modulo $\mathbf{T}$; *cf.* Rem. 4.5 (i).

(d) Star. Equations (4.30) and (4.29) give, where a, b and c are letters in A:

$$\frac{\partial}{\partial ab}(\mathsf{E}^*) = \left[\frac{\partial}{\partial ab}\,\mathsf{E}\right]\cdot\mathsf{E}^* + \mathsf{c}\left(\frac{\partial}{\partial a}\,\mathsf{E}\right)\cdot\frac{\partial}{\partial b}\,\mathsf{E}\cdot\mathsf{E}^* \ ,$$

$$\frac{\partial}{\partial abc}(\mathsf{E}^*) = \left[\frac{\partial}{\partial abc}\,\mathsf{E}\right]\cdot\mathsf{E}^* + \mathsf{c}\left(\frac{\partial}{\partial ab}\,\mathsf{E}\right)\cdot\frac{\partial}{\partial b}\,\mathsf{E}\cdot\mathsf{E}^* + \mathsf{c}\left(\frac{\partial}{\partial a}\,\mathsf{E}\right)\cdot\frac{\partial}{\partial bc}\,\mathsf{E}\cdot\mathsf{E}^* + \mathsf{c}\left(\frac{\partial}{\partial a}\,\mathsf{E}\right)\cdot\mathsf{c}\left(\frac{\partial}{\partial b}\,\mathsf{E}\right)\cdot\frac{\partial}{\partial c}\,\mathsf{E}\cdot\mathsf{E}^* \ .$$

We conclude that in general $\frac{\partial}{\partial f}(\mathsf{E}^*)$ is a sum with (at most $2^{|f|-1}$) terms of the form $\frac{\partial}{\partial g}\,\mathsf{E}\cdot\mathsf{E}^*$ (where g is a suffix of f) and hence:

$$\mathbf{nd}(\mathsf{E}^*) \leqslant 2^{\mathbf{nd}(\mathsf{E})} \ .$$

The assertion that D_{E}, the set of derivatives of E modulo $\mathbf{A_sCI}$, is *effectively computable*, does not follow directly from $\mathbf{nd}(\mathsf{E})$ being finite, but is easily deducible. We note first that computations modulo $\mathbf{A_sCI}$ are effective: by, for example, ordering the expressions lexicographically we can arrange that each expression is equivalent to a *unique expression* of minimal length.

If we then write $D_{\mathsf{E}}^{(k)}$ for the set (modulo $\mathbf{A_sCI}$) of derivatives of E with respect to all the words in A^* of length less than or equal to k, we obtain an increasing sequence of sets of expressions. If for some l, $D_{\mathsf{E}}^{(l)} = D_{\mathsf{E}}^{(l+1)}$ then, by a chain of reasoning which is now familiar to us *cf. Prop. 3.12, p. 115*

$$D_{\mathsf{E}}^{(l+2)} = \left\{\frac{\partial}{\partial a}\,\mathsf{F}\,\middle|\,\mathsf{F}\in D_{\mathsf{E}}^{(l+1)} \quad a\in A\right\} = \left\{\frac{\partial}{\partial a}\,\mathsf{F}\,\middle|\,\mathsf{F}\in D_{\mathsf{E}}^{(l)} \quad a\in A\right\} = D_{\mathsf{E}}^{(l+1)} = D_{\mathsf{E}}^{(l)} \ ,$$

and we see that $D_{\mathsf{E}}^{(l+p)} = D_{\mathsf{E}}^{(l)}$ for all p and hence $D_{\mathsf{E}} = D_{\mathsf{E}}^{(l)}$. ■

4.4.3 Derivative automaton

Theorem 4.3 gives us as a corollary an effective procedure for computing from an expression E a finite deterministic automaton which accepts the language denoted by E.

Definition 4.8 Let E be a rational expression over A and D_{E} the set of its derivatives modulo $\mathbf{A_sCI}$. The automaton $\mathcal{B}_{\mathsf{E}} = \langle\, D_{\mathsf{E}}, A, \delta_{\mathsf{E}}, \{\mathsf{E}\}, T_{\mathsf{E}}\,\rangle$ defined by

$$\forall \mathsf{F}\in D_{\mathsf{E}}\,,\ \forall a\in A \qquad (\mathsf{F},a)\delta_{\mathsf{E}} = \frac{\partial}{\partial a}\,\mathsf{F} \qquad \text{and} \qquad T_{\mathsf{E}} = \{\mathsf{F}\in D_{\mathsf{E}}\,\big|\,\mathsf{c}(\mathsf{F}) = 1\}$$

is a *finite* deterministic automaton effectively computable from E. It is called the *derivative automaton* of E. □

From Proposition 4.9 and Proposition 3.7 there follows:

Proposition 4.11 *The automaton $\mathcal{B}_{\mathsf{E}}$ recognises* $|\mathsf{E}|$. ■

Example 4.3 (continued) Consider $\mathsf{E}_1 = (a+b)^* a\, b\, (a+b)^*$ once more and note

$$\mathsf{E}_2 = \frac{\partial}{\partial a}\,\mathsf{E}_1 = \mathsf{E}_1 + b\,(a+b)^* \ , \qquad \mathsf{E}_3 = \frac{\partial}{\partial ab}\,\mathsf{E}_1 = \mathsf{E}_1 + (a+b)^* \ ,$$

$$\mathsf{E}_4 = \frac{\partial}{\partial aba}\,\mathsf{E}_1 = \mathsf{E}_1 + b\,(a+b)^* + (a+b)^* \ .$$

From the above calculations, we deduce that $D_{\mathsf{E}_1} = \{\mathsf{E}_1, \mathsf{E}_2, \mathsf{E}_3, \mathsf{E}_4\}$, and the automaton $\mathcal{B}_{\mathsf{E}_1}$ shown in Figure 4.2. □

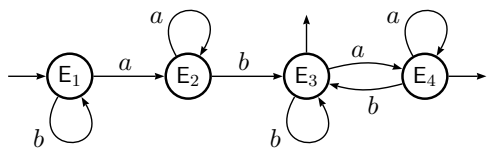


Figure 4.2: The automaton $\mathcal{B}_{\mathsf{E}_1}$

We see from this example that the derivative automaton of an expression E is not, in general, equal (isomorphic) to the automaton of quotients, the minimal automaton of $|\mathsf{E}|$. We will return in the next section to the problem of the computation of an automaton from a rational expression.

Exercises

4.7 Verify Property 4.1.

4.8 Verify Property 4.2.

• 4.9 Calculate the derivative automaton of the following expressions (we write $\mathsf{E}\,\mathsf{F}\,\mathsf{G}$ for $\mathsf{E}(\mathsf{F}\mathsf{G})$):

(a) $(a^*b + b\,b^*a)^*$; (b) $a^*b\,(a\,b^*a + b\,a^*b)^*b\,a^* + a^*$; (c) $(a + b\,(a\,b^*a)^*b)^*$;

(d) $a^*a\,a^*a\,a^*$.

• 4.10 **Derivation and associativity of product.**

(a) Calculate the derivative automaton of $\mathsf{E}'_1 = \Big(\big((a+b)^* \cdot a\big) \cdot b\Big) \cdot (a+b)^*$.

(b) Do the same for $\mathsf{E}'_2 = \Big(\big((a^* \cdot a) \cdot a^*\big) \cdot a\Big) \cdot a^*$.

(c) And again for $\Big((a \cdot b) \cdot \big(c \cdot (a \cdot b)^*\big)\Big)$ and $\Big(a \cdot \Big(b \cdot \big(c \cdot (a \cdot b)^*\big)\Big)$.

(d) Comment and draw conclusion.

DEEPER & BROADER

It is rather arbitrary to classify topics in a domain as principal and complementary, and anyone whose speciality or favourite tool has been relegated to a seemingly inferior category will take it as a mark of bad taste or want of discernment. There is no question of hierarchy or value judgement in our mode of presentation, but rather of a change of tone. In the preceding sections, I attempted to be comprehensive and didactic; in those which follow I will allow myself more personal choices, a more elliptic style, and references to other works and articles.

As promised at the start of the chapter, this part comprises four sections. The first two can be seen as two completely different variations on Kleene's Theorem. Section 5 presents the construction of automata whose interest lies in their translation into algorithms which are used ubiquitously. Section 6 tackles one of the hardest problems in the study of rational languages, whose solution still holds many mysteries; the fascination that it excites is of a rather theoretical order. Section 7 presents different models of 'machines' equivalent to finite automata, and Section 8 describes, in the form of problems, several properties of rational languages.

5 From expressions to automata

The computation of an automaton from a rational expression – that is, the algorithmic version of one direction of Kleene's Theorem – is a problem which has been tackled many times. Several works deal only with this aspect of automata. This is because it was the central question in one of the first applications of finite automata: the lexical analysis of programming languages and searching for sequences or patterns, which are abilities used by all word processors and many other utility programs.

In the first subsection we shall revisit the construction of a standard automaton from an expression. The variety of methods that all lead, directly or indirectly, to the same result gives this object a fundamental character. The second subsection describes another construction based on a variant of the notion of derivation of an expression, which produces a different result from the standard automaton. The last subsection is devoted to the problem of finding a word in a text, which comes down to considering some rather special rational expressions.

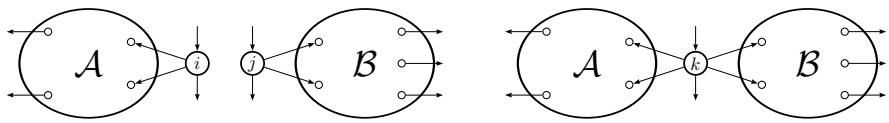
5.1 The standard automaton of an expression

We discussed in Section 2.2 how the proof of Proposition 2.2 not only demonstrates that $\operatorname{Rec} A^*$ is rationally closed but that it is itself an *algorithm* that produces an *cf. p. 92 and p. 94*

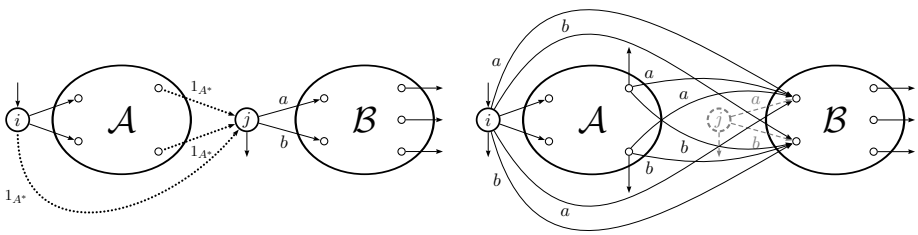
automaton from a rational expression.[50] The automaton thus constructed clearly depends on the method used: we imagined that it would be a normalised or standard automaton.

cf. Exam. 2.3, p. 90

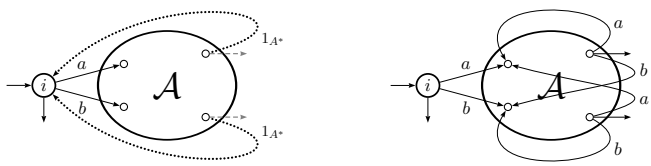
The use of normalised automata leads to an almost absurd explosion in the number of states, which does not invalidate the idea of a normalised automaton itself but does make it unsuitable for making an expression into an automaton. We shall see that standard automata, on the other hand, are reasonably compact, and susceptible to a variety of interpretations that testify to the canonical nature of the link they create between rational expressions and automata.



(a) for union



(b) for product



(c) for star

Figure 5.1: Construction of standard automata

5.1.1 Direct construction

The two constructions of standard automata for the product and star of languages recognised by standard automata are shown again in Figure 5.1(b) and (c); the construction for the union consists of simply merging the initial states (Figure 5.1(a)). Combined with the trivial observation that a single letter and the empty word are

cf. Fig. 2.7, p. 93 and Fig. 2.8, p. 94

[50] But at the time we had not yet given a precise definition of rational expressions.

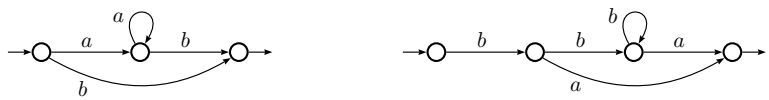
both languages recognised by standard automata, they enable us to associate, with every rational expression E, a corresponding standard automaton $\mathcal{S}_{\mathsf{E}}$, unique and well defined.

Definition 5.1 The *literal length* of an expression E, written $\ell(\mathsf{E})$, is the number of occurrences of letters of A in E. □

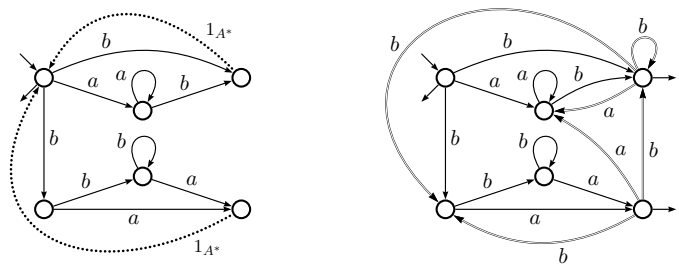
We observe that $\ell((\mathsf{E}^*)) = \ell(\mathsf{E})$ and that $\ell((\mathsf{E}+\mathsf{F})) = \ell((\mathsf{E}\cdot\mathsf{F})) = \ell(\mathsf{E}) + \ell(\mathsf{F})$. These formulas, combined with an obvious induction, allow us to show:

Proposition 5.1 *The standard automaton $\mathcal{S}_{\mathsf{E}}$ associated with the rational expression E has $\ell(\mathsf{E})+1$ states.* ■

Example 2.5 anticipated this proposition. Figure 5.2 shows the construction of the standard automaton of the expression $\mathsf{E}_2 = (a^*b + bb^*a)^*$. *cf. p. 94 and Exer. 2.13*



(a) standard automata for a^*b and bb^*a



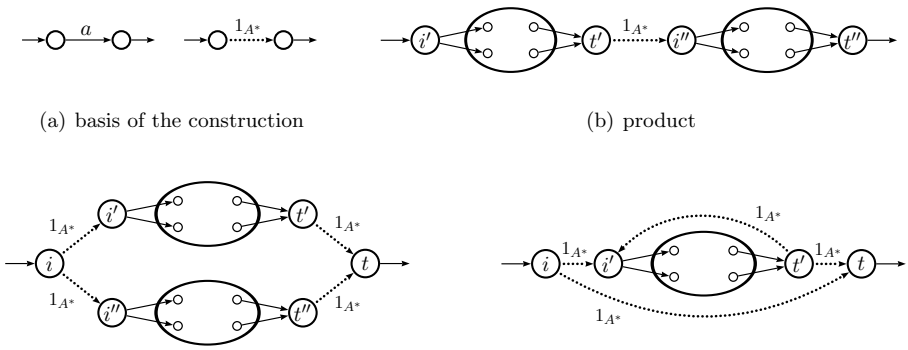
(b) Last step: star of $a^*b + bb^*a$. The transitions generated by the removal of spontaneous transitions are drawn with double lines.

Figure 5.2: Construction of $\mathcal{S}_{\mathsf{E}_2}$

5.1.2 Thompson's construction

Another way to construct an automaton from a rational expression is based on the use of spontaneous transitions and produces, in fact, an automaton with spontaneous transitions. The method is due to K. Thompson [243]. Conceived to be directly implementable as a program, it produces a representation that is most efficient for scanning text. Figure 5.3 shows the construction, which allows every rational expression E to be associated with a unique, well-defined automaton with spontaneous transitions $\mathcal{T}_{\mathsf{E}}$. Figure 5.4 shows the construction applied to the expression $\mathsf{E}_2 = (a^*b + bb^*a)^*$.

Notice that this construction does not encode the associativity of union, so $(\mathsf{E}+\mathsf{F})+\mathsf{G}$ and $\mathsf{E}+(\mathsf{F}+\mathsf{G})$ produce two different automata. We also observe the following



(a) basis of the construction (b) product

(c) union (d) star

Figure 5.3: Thompson's construction

properties:

Property 5.1 *If the rational expression* E *comprises* n *symbols (letters + operators), the automaton* $\mathcal{T}_{\mathsf{E}}$ *has at most* $2n$ *states.* ■

Property 5.2 *Every state of* $\mathcal{T}_{\mathsf{E}}$ *is the source (resp. destination) of one or two transitions, except for the final state (resp. the initial state).* ■

It is remarkable that this construction corresponds to another way of defining the standard automaton, a result whose proof we shall leave to an exercise.

Proposition 5.2 *Let* E *be a rational expression. The automaton obtained from* $\mathcal{T}_{\mathsf{E}}$ *by* backward closure *is equal to* $\mathcal{S}_{\mathsf{E}}$. ■

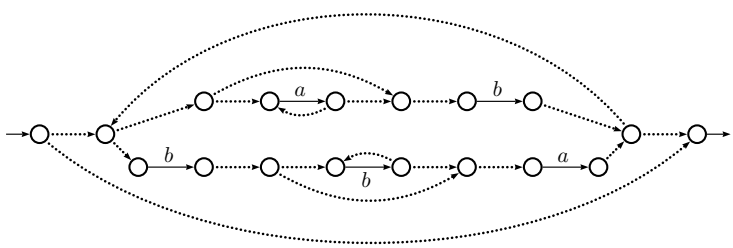


Figure 5.4: The automaton $\mathcal{T}_{\mathsf{E}_2}$

Exercises

- 5.1 Compute the standard automaton of the expression $(a + b\,(a\,b^*a)^*b)^*$.

5.2 Verify Proposition 5.1.

5.3 Verify Proposition 5.2.

5.2 The derived term automaton

We have seen that the derivation of rational expressions is also a way of associating an automaton with an expression. This automaton has the disadvantage of being deterministic, which is not a defect in itself (on the contrary) but which raises the possibility of an exponential explosion. The process of computation itself assumes that expressions can be compared for equality modulo the identities $\mathbf{A_sCI}$, which is algorithmically expensive.

cf. Def. 4.8, p. 143

cf. Th. 4.3, p. 142

A modification to the definition of derivation will overcome these disadvantages: first, the expressions computed will no longer have to be considered modulo any identities, and second, the automata produced will be no larger than the standard automata of the expressions.

The idea underlying this transformation is that the derivative of a sum, instead of producing an expression which is the sum of the derivatives of the terms of the sum, produces a *set of expressions* that is the set of derivatives of the terms of the sum. This idea may seem baroque and artificial, excused only by the quality of the result. It is nothing of the sort: it has a solid theoretical foundation and will be fully justified when we have defined the derivation of *weighted expressions* in Chapter III.

cf. Sec. III.4.2, p. 443

5.2.1 Derived terms

Definition 5.2 Let E be a rational expression over A and a a letter in A. The $\mathbb{B}$-*derivation* of E with respect to a, written $\dfrac{\partial'}{\partial' a}\mathsf{E}$, is *a set of rational expressions over* A, defined recursively by the following formulas:

$$\frac{\partial'}{\partial' a}0 = \frac{\partial'}{\partial' a}1 = 0 \ , \qquad \forall a, b \in A \quad \frac{\partial'}{\partial' a}b = \begin{cases} 1 & \text{if} \quad b = a \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial'}{\partial' a}(\mathsf{E}+\mathsf{F}) = \left\{\frac{\partial'}{\partial' a}\mathsf{E}\right\} \cup \left\{\frac{\partial'}{\partial' a}\mathsf{F}\right\} \tag{5.1}$$

$$\frac{\partial'}{\partial' a}(\mathsf{E}\cdot\mathsf{F}) = \left\{\left[\frac{\partial'}{\partial' a}\mathsf{E}\right]\cdot\mathsf{F}\right\} \cup \left\{\mathsf{c}(\mathsf{E})\ \frac{\partial'}{\partial' a}\mathsf{F}\right\} \tag{5.2}$$

$$\frac{\partial'}{\partial' a}(\mathsf{E}^*) = \left[\frac{\partial'}{\partial' a}\mathsf{E}\right]\cdot\mathsf{E}^* \tag{5.3}$$

As for derivation, all the calculations are performed modulo the trivial identities $\mathbf{T}$. Furthermore, the induction implied by (5.2) and (5.3) must be realised by distributing $\mathbb{B}$-derivation over union:

cf. Rem. 4.5, p. 140

$$\frac{\partial'}{\partial' a}\left(\bigcup_{i\in I}\mathsf{E}_i\right) = \bigcup_{i\in I}\frac{\partial'}{\partial' a}\mathsf{E}_i \ ,$$

and product over union:

$$\left(\bigcup_{i\in I}\mathsf{E}_i\right)\cdot\mathsf{F} = \bigcup_{i\in I}(\mathsf{E}_i\cdot\mathsf{F}) \ . \tag{$\mathbf{D}$}$$

Definition 5.3 Let E be a rational expression over A and g a non-empty word from A^*: that is, $g = f\,a$ with a in A. The $\mathbb{B}$-*derivation* of E with respect to g, written $\dfrac{\partial'}{\partial' g}\mathsf{E}$, is a set of rational expressions over A, defined recursively by the formulas (5.1)–(5.3), and by

$$\forall f \in A^*\,,\ \forall a \in A \qquad \frac{\partial'}{\partial' fa}\mathsf{E} = \frac{\partial'}{\partial' a}\left(\frac{\partial'}{\partial' f}\mathsf{E}\right) \quad . \tag{5.4}$$

We will call any rational expression which belongs to a set $\frac{\partial'}{\partial' g}\mathsf{E}$ for some g in A^* a *derived term* of E. □

One thing that emerges from Definitions 5.2 and 5.3 is that the computation of derived terms of an expression E corresponds to taking the $\mathbb{B}$-derivation of some sub-expressions of derivatives of E.

Example 5.1 The $\mathbb{B}$-derivation with respect to a and b of the expression

$$\mathsf{E}_1 = (a+b)^* \cdot \Big(a \cdot \big(b \cdot (a+b)^*\big)\Big)$$

yields $\frac{\partial'}{\partial' a}\mathsf{E}_1 = \{\mathsf{E}_1, b\cdot(a+b)^*\}$ and $\frac{\partial'}{\partial' b}\mathsf{E}_1 = \mathsf{E}_1$. The derived terms of E_1 are therefore $\{\mathsf{E}_1, b\cdot(a+b)^*, (a+b)^*\}$. □

Having acknowledged that the language denoted by a set of rational expressions is the union of the languages denoted by the expressions, we immediately conclude from Properties 4.2 and 4.3 and from Proposition 4.9 the corresponding result for $\mathbb{B}$-derivation:

cf. p. 140

Proposition 5.3 $\forall \mathsf{E} \in \mathsf{Rat}\mathsf{E}\,A^*\,,\ \forall f \in A^* \qquad \left|\dfrac{\partial'}{\partial' f}\mathsf{E}\right| = f^{-1}\,|\mathsf{E}|\,.$ ■

5.2.2 A theorem of V. Antimirov

The interest of $\mathbb{B}$-derivation (with respect to the derivation seen in Section 4.4) is that it enables us to consider derived terms directly, rather than via the identities $\mathbf{A_sCI}$.

Theorem 5.1 *A rational expression* E *has at most* $\ell(\mathsf{E})$ *derived terms.*

We shall prove a more precise proposition, but with a different formulation: instead of considering $\mathbb{B}$-derivation with respect to all the words in A^*, we shall look just at $\mathbb{B}$-derivation with respect to the letters of A; in return we shall require global closure.

Theorem 5.2 *Let* E *be in* $\mathsf{Rat}\mathsf{E}\,A^*$ *. The derived terms of* E *form a set* P *of expressions* $\{\mathsf{K}_p\}_{p\in P}$ *, such that both* $\|P\| \leqslant \ell(\mathsf{E})$ *, and for each letter* a *in* A*, there exists a subset* $Q^{(a)}$ *of* P *and subsets* $\{H_p^{(a)}\}_{p\in P}$ *of* P*, such that*

(i) $\dfrac{\partial'}{\partial' a}\mathsf{E} = \displaystyle\bigcup_{p\in Q^{(a)}} \mathsf{K}_p$ *and* (ii) $\forall p \in P \quad \dfrac{\partial'}{\partial' a}\mathsf{K}_p = \displaystyle\bigcup_{j\in H_p^{(a)}} \mathsf{K}_j\,.$

Proof. A set P which satisfies (i) and (ii), and where $\|P\| \leqslant \ell(\mathsf{E})$, is constructed by induction on the depth of the expression E. The fact that it constitutes the set of derived terms is shown by induction on the length of words at each step of the preceding induction. The proposition is clearly true for $\mathsf{E} = 1$ and $\mathsf{E} = a$, for each a in A (the expression 0 is not counted as a derived term).

(a) If it is true for E and F, it is true for $(\mathsf{E} + \mathsf{F})$. In fact, we have

$$\frac{\partial'}{\partial' a}(\mathsf{E}+\mathsf{F}) = \left\{ \frac{\partial'}{\partial' a}\,\mathsf{E},\ \frac{\partial'}{\partial' a}\,\mathsf{F} \right\} = \bigcup_{p \in Q^{(a)}} \mathsf{K}_p \cup \bigcup_{s \in R^{(a)}} \mathsf{L}_s$$

using the obvious notation. The set $\bigcup_{p\in P} \mathsf{K}_p \cup \bigcup_{s\in S} \mathsf{L}_s$ satisfies condition (ii) and is clearly the set of derived terms of $(\mathsf{E} + \mathsf{F})$.

(b) If it is true for E and F, it is true for $(\mathsf{E} \cdot \mathsf{F})$. It follows that

$$\frac{\partial'}{\partial' a}(\mathsf{E} \cdot \mathsf{F}) = \left(\left[\frac{\partial'}{\partial' a}\,\mathsf{E} \right] \cdot \mathsf{F} \right) \cup \mathsf{c}(\mathsf{E})\ \frac{\partial'}{\partial' a}\,\mathsf{F} = \bigcup_{p \in Q^{(a)}} (\mathsf{K}_p \cdot \mathsf{F}) \cup \bigcup_{s \in R^{(a)}} \mathsf{c}(\mathsf{E})\ \mathsf{L}_s \ ,$$

and, for each p in P, $$\frac{\partial'}{\partial' a}(\mathsf{K}_p \cdot \mathsf{F}) = \bigcup_{j \in H_p^{(a)}} (\mathsf{K}_j \cdot \mathsf{F}) \cup \bigcup_{s \in R^{(a)}} \mathsf{c}(\mathsf{K}_p)\ \mathsf{L}_s \ ,$$

from which we see that the set $\bigcup_{p\in P}(\mathsf{K}_p \cdot \mathsf{F}) \cup \bigcup_{s\in S} \mathsf{L}_s$ satisfies condition (ii) and, by induction on the length of words, that it is the set of derived terms of $(\mathsf{E} \cdot \mathsf{F})$.

(c) If it is true for E, it is true for (E^*). We have

$$\frac{\partial'}{\partial' a}(\mathsf{E}^*) = \left(\left[\frac{\partial'}{\partial' a}\,\mathsf{E} \right] \cdot (\mathsf{E}^*) \right) = \bigcup_{p \in Q^{(a)}} (\mathsf{K}_p \cdot \mathsf{E}^*) \ ,$$

and, for each p in P, $$\frac{\partial'}{\partial' a}(\mathsf{K}_p \cdot \mathsf{E}^*) = \bigcup_{j \in H_p^{(a)}} (\mathsf{K}_j \cdot \mathsf{E}^*) \cup \bigcup_{j \in Q^{(a)}} \mathsf{c}(\mathsf{K}_p)\ (\mathsf{K}_j \cdot \mathsf{E}^*) \ ,$$

from which we see that the set $\bigcup_{p\in P}(\mathsf{K}_p \cdot \mathsf{E}^*)$ satisfies condition (ii) and, by induction on the length of words, that it is the set of derived terms of (E^*).

At each of these three steps, we verify that the number of derived terms computed is really less than the literal length of the expression. ■

The statement of Theorem 5.2 can itself be seen as the definition of an automaton.

Definition 5.4 Let E be in $\mathsf{Rat}\mathsf{E}\,A^*$. Let P_{E} be the set of its derived terms and P'_{E} the union of P_{E} and E (that is, $P'_{\mathsf{E}} = P_{\mathsf{E}}$ if E is a derived term of itself). Set $\mathsf{E} = \mathsf{K}_i$ and $Q^{(a)} = H_i^{(a)}$ (and we have that $i \in P'$). The automaton

$$\mathcal{A}_{\mathsf{E}} = \langle\, P'_{\mathsf{E}}, A, \tau_{\mathsf{E}}, \{\mathsf{E}\}, U_{\mathsf{E}} \,\rangle$$

defined by $$\forall p \in P'_{\mathsf{E}}\,,\ \forall a \in A \qquad (\mathsf{K}_p, a)\tau_{\mathsf{E}} = \{\mathsf{K}_q \mid q \in H_p^{(a)}\}$$ and

$$U_{\mathsf{E}} = \{\mathsf{K}_p \in P'_{\mathsf{E}} \mid \mathsf{c}(\mathsf{K}_p) = 1_{A^*}\}$$

is called the *derived term automaton* of E. □

From Proposition 5.3 it follows that:

cf. also Propty 3.7, p. 112

Proposition 5.4 *Let* E *be in* $\mathsf{Rat}\mathsf{E}\,A^*$ *. The derived term automaton of* E *recognises the language denoted by* E*:* $|\mathcal{A}_{\mathsf{E}}| = |\mathsf{E}|$ *.* ■

Example 5.1 (continued) Consider again $\mathsf{E}_1 = (a+b)^* \cdot \Big(a \cdot \big(b \cdot (a+b)^*\big)\Big)$. The derived terms of E_1 are E_1 , $b\,(a+b)^*$, and $(a+b)^*$, and the automaton of derived terms of E_1 is shown in Figure 5.5; we recognise it as the automaton $\mathcal{A}_1$. □

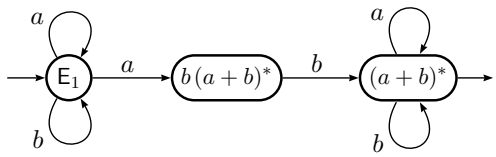


Figure 5.5: The automaton $\mathcal{A}_{\mathsf{E}_1}$

With no further assumption about the expression, the derived term automaton (its determinisation in fact) and the derivative automaton have no specific relationship as can be seen via simple examples (see below). We leave to Chapter II the description of the link of the derived term automaton with the standard automaton.

cf. Exer. II.3.8, p. 264

Exercises

- 5.4 Compute the derived term automaton of the expression $\mathsf{E}_2 = (a^*b + b\,b^*a)^*$.

cf. Exer. 2.15, p. 101 and Fig. 6.1 (c), p. 158

- 5.5 Do the same for the expression
$$\mathsf{E}_3 = a^* + a^*b\,(b\,a^*b)^*b\,a^* + a^*b\,(b\,a^*b)^*a\,(b + a\,(b\,a^*b)^*a)^*a\,(b\,a^*b)^*b\,a^* \,.$$

- 5.6 Verify from the expressions $\mathsf{F}_1 = a\,(a+b)^* + b\,(b+a)^*$ and $\mathsf{F}_2 = (a\,a + a\,a\,a)^*$ that the determinisation of the derived term automaton may be larger, or smaller, than the derivative automaton.

5.3 String matching

The problem of finding a word in a text appears in different forms in various domains. This type of algorithm is used for online search engines, for the search and replace function of word processors, in spelling checkers which must look words up in dictionaries, as well as in lexical analysers, which, as the first pass of a compiler, break the stream of characters forming a program into a sequence of lexemes (keywords, identifiers *etc.*). Syntactic analysis and translation into another language are performed by other parts of the compiler.

So many situations, so many different algorithms, but the fundamental problem remains of finding occurrences of words belonging to a finite set X in a text written in an alphabet A; that is, to recognise the language A^*X, a rational language, and hence to construct the corresponding automaton. Here we will only study the case where X is reduced to a single word. The case where X is an arbitrary finite set gives rise to standard generalisations.[51]

[51] *Cf.* Notes and references section. § 5.3.1 is adapted from [183], and § 5.3.2 from [23, Ch. 10].

5.3.1 Automaton for finding a word

Let $f = a_1\, a_2 \cdots a_n$ be a word in A^* (with the a_i in A) and let $\mathcal{A}_f$ be the automaton below which recognises the language $A^* f$; we write P_f for the set of *prefixes* of f:

$$P_f = \{1_{A^*}, a_1, a_1\, a_2, \ldots, a_1\, a_2 \cdots a_{n-1}, a_1\, a_2 \cdots a_{n-1}\, a_n\}$$

The set of states of $\mathcal{A}_f$ can be identified with P_f (via the identification between the integer i and the prefix $a_1\, a_2 \cdots a_i$). We write $\mathcal{D}_f$ for the determinisation of $\mathcal{A}_f$. *A priori*, the set of states of $\mathcal{D}_f$ is a subset of the *set of subsets* of P_f. We will show that the set of states of $\mathcal{D}_f$ can also be identified with P_f.

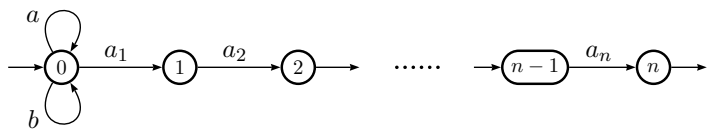


Figure 5.6: The automaton $\mathcal{A}_f$

Let w be an arbitrary word in A^*. A computation $0 \xrightarrow{w} k$ in $\mathcal{A}_f$ necessarily decomposes into

$$0 \xrightarrow{u} 0 \xrightarrow{a_1} 1 \xrightarrow{a_2} 2 \cdots \xrightarrow{a_k} k \tag{5.5}$$

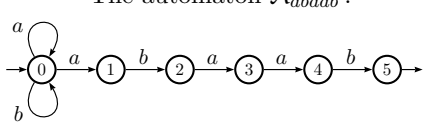
where $w = u\, a_1\, a_2 \cdots a_k$. Conversely, if $a_1\, a_2 \cdots a_k$ is a suffix of w, (5.5) is a computation in $\mathcal{A}_f$ with label w.

Let v be the *longest suffix* of w which is an element of P_f; that is, which is a prefix of f: the suffixes of w that are prefixes of f are the suffixes of v which are prefixes of f. Conversely, each element v of P_f takes $\mathcal{D}_f$ to the state which is the set of suffixes of v that belong to P_f and of which v is the greatest element: there is a bijection between the states of $\mathcal{D}_f$ and P_f, and $\mathcal{D}_f$ has $|f| + 1$ states (like $\mathcal{A}_f$).

Furthermore, as $\mathcal{A}_f$ is co-deterministic and co-accessible, $\mathcal{D}_f$ is the *minimal automaton* of $A^* f$. *cf. Prop. 3.13, p. 116*

Example 5.2 The automaton $\mathcal{A}_{abaab}$ and the construction of $\mathcal{D}_{abaab}$ by the subset construction are shown in Figure 5.7. □

The automaton $\mathcal{A}_{abaab}$:



The table for the determinisation of $\mathcal{A}_{abaab}$:

	$\{0\}$	$\{0,1\}$	$\{0,2\}$	$\{0,1,3\}$	$\{0,1,4\}$	$\{0,2,5\}$
a	$\{0,1\}$	$\{0,1\}$	$\{0,1,3\}$	$\{0,1,4\}$	$\{0,1\}$	$\{0,1,3\}$
b	$\{0\}$	$\{0,2\}$	$\{0\}$	$\{0,2\}$	$\{0,2,5\}$	$\{0\}$

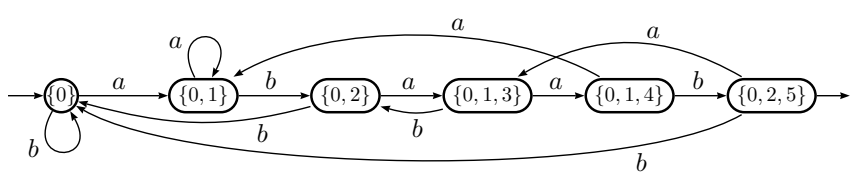


Figure 5.7: The automaton $\mathcal{D}_{abaab}$

The automaton $\mathcal{D}_f$ 'finds' the word f in the sense that if a word t in A^* (the 'text') is 'read' (from left to right) by $\mathcal{D}_f$, each letter of t which takes $\mathcal{D}_f$ to its unique final state is the last letter of an occurrence of a factor f in t, even if the factor f overlaps other preceding occurrences. It is not quite enough to have defined $\mathcal{D}_f$; we must also show how to compute it from f without just reusing the subset construction.

Let us start with a definition: the *border* of a non-empty word u is the longest *proper* prefix of u which is also a suffix of u. Thus, ab is the border of $abaab$, and $abab$ is the border of $ababab$.[52] Then $\beta\colon A^* \to A^*$ is the function which takes each word to its border (β is undefined for 1_{A^*}). From this definition, and using the preceding notations, it follows that for all p in P_f, $(p)\beta$ is the longest proper suffix of p which is in P_f. The following table gives the values of β for $f = abaab$:

p	a	ab	aba	$abaa$	$abaab$
$(p)\beta$	1_{A^*}	1_{A^*}	a	a	ab

Property 5.3 *The transition function of $\mathcal{D}_f$ is defined by the recursive equation*

$$\forall a \in A\,,\ \forall p \in P_f\,,\ p \neq 1_{A^*} \qquad p \cdot a = \begin{cases} pa & \text{if} \quad pa \in P_f \\ (p)\beta \cdot a & \text{otherwise} \end{cases}$$

and the initial conditions

$$\forall a \in A \qquad 1_{A^*} \cdot a = \begin{cases} a & \text{if} \quad a \in P_f \\ 1_{A^*} & \text{otherwise} \end{cases} .$$

Proof. According to the foregoing, (p, a, q), where p is in P_f and a in A, is a transition of $\mathcal{D}_f$ if and only if q is the longest suffix of pa which is in P_f; that is, which is a prefix of f.

Let $p = 1_{A^*}$, then if $a \in P_f$, $q = a$, and if $a \notin P_f$, $q = 1_{A^*}$.

Suppose now that $|p| > 0$. If $pa \in P_f$ then $q = pa$. If $pa \notin P_f$, let q be the longest suffix of pa which is a prefix of f. Suppose that q is strictly longer than $(p)\beta a$; then $q = ta$ and t is a suffix of p, prefix of f and strictly longer than $(p)\beta$: this contradicts the definition of $(p)\beta$. Thus q is a *suffix* of $(p)\beta a$; it is the longest suffix of $(p)\beta a$ which is a prefix of f: $q = (p)\beta \cdot a$. ■

Property 5.4 *The function β can itself be computed by an analogous recursive equation*

$$\forall a \in A\,,\ \forall p \in P_f\,,\ p \neq 1_{A^*} \qquad (pa)\beta = \begin{cases} (p)\beta a & \text{if} \quad (p)\beta a \in P_f \\ ((p)\beta a)\beta & \text{otherwise} \end{cases}$$

and the initial conditions

$$\forall a \in A \qquad (a)\beta = 1_{A^*} \ .$$

[52] In some works, *a* border of u is a factor of u which is both a prefix and a suffix; *the* border of u is then the longest such (in [68]) or *maximal border* of u (in [23]).

Proof. By definition of β, $(a)\beta = 1_{A^*}$ for all a in A. We therefore must compute, for p and pa in P_f, $s = (pa)\beta$, the longest proper suffix of pa which is a prefix of f. If $(p)\beta a$ is a suffix of s, $s = ta$, then t is a proper suffix of p and a prefix of f, and t cannot be longer than $(p)\beta$, hence $t = (p)\beta$ and $s = (p)\beta a$. If not, s is a proper suffix of $(p)\beta a$, in fact the longest proper suffix of $(p)\beta a$ that is a prefix of f: $s = ((p)\beta a)\beta$ by definition. ■

We conclude from the above that we can compute all the $(p)\beta$, for p in P_f, in $O(|f|)$, then all the $p \cdot a$, for p in P_f and a in A, in $O(\|A\|\,|f|)$. Finding f in t then corresponds to reading t with the deterministic automaton $\mathcal{D}_f$ and therefore has a complexity of $O(|t|)$.

5.3.2 String matching with a sliding window

Suppose that we wish to solve the problem of finding a word in a text in a purely algorithmic manner, without needing to keep the automaton model in mind.[53] The idea that no doubt springs to mind is to imagine the text t written on a tape in the classic manner, an array whose ith square is occupied by the ith letter t_i of t; the word f is written in a window (another table), which is slid along the tape (*cf.* Figure 5.8).

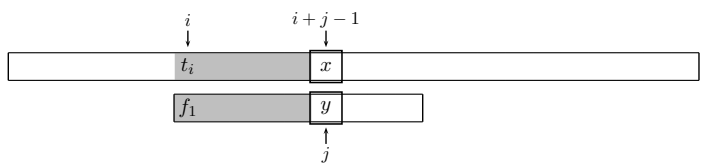


Figure 5.8: Sliding a window along a text

The first square of the window is in front of the ith square of t and we compare f_1 and t_i; if they are equal, we compare f_2 and t_{i+1}, then if they too are equal, f_3 and t_{i+2}, and so on. There are two possible outcomes. We may reach the end of the window: all the comparisons have succeeded and the word f indeed occurs in t starting at position i. Alternatively the jth comparison failed: f_j and t_{i+j-1} are different and f is not a factor of t starting at position i. In both cases, we slide the window along and start again. The question is how far to slide the window.

If we slide one position each time we have the naive algorithm; it is slow and makes $O(|f|\,|t|)$ comparisons, but it will definitely find all occurrences of f in t.

We can be cleverer and take into account what we learn about t from successful comparisons. If we have found the prefix p of f in t starting at position i, we can slide the window in such a way that we do not need to retest any of the letters of t that we have already tested. In the new position, the window should overlap the prefix p by a factor q which is the longest prefix of p that is also a suffix of p: q is the *border* of p

[53] *There are those minds whose agile train of thought is by lacunae made fragile and short* (with apologies to Nicolas Boileau).

(*cf.* Figure 5.9). This algorithm is known as the Morris–Pratt algorithm; we can show that the number of comparisons is $O(|t|)$ and that the computation of the distance to slide at each step (performed once as an initialisation step) is $O(|f|)$.

We can improve this algorithm by noticing that the first letter z of f to test after sliding the window must be different from the y that caused the slide (see Figure 5.9). We thus obtain a sliding function that is always at least as good as the previous one, which improves the efficiency of the resulting algorithm, known as the Knuth–Morris–Pratt or K M P algorithm (henceforth simply K M P). It is notable that the computation of the sliding function in this case is exactly equal to that of the transition function of the automaton $\mathcal{D}_f$: modelling it as an automaton turns this most subtle algorithm into a canonical construction, which is an important merit of modelling.[54]

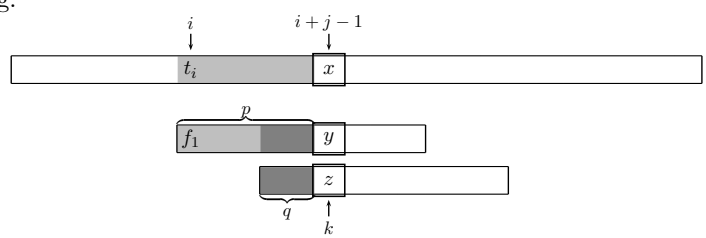


Figure 5.9: Sliding the window over the text intelligently

5.3.3 Implementation with a default successor

It remains to note that the automaton $\mathcal{D}_f$ has a grave defect compared with K M P. Its size, and therefore the number of operations required to describe it, is proportional to the number of letters in the alphabet A, which becomes prohibitive when we go from the alphabet used in our examples to the ASCII alphabet of 256 letters that is actually used by word processors and compilers.[55] (The initial phase of K M P associates with each position of the word f the distance to slide the window on a failed comparison; its computation and the amount of storage it requires is $O(|f|)$, and is independent of the size of the alphabet.)

However, the amount of essential, of 'useful', information encoded in the automaton $\mathcal{D}_f$ is indeed linear in $|f|$, independent of $\|A\|$. Imagine that we wished to recognise the word $abaab$ from Example 5.2 in a text written in ASCII; the new automaton $\mathcal{D}_{abaab}$ would be identical to the previous one except that in each state, and for each letter other than a and b, it would have a transition to state $\{0\}$. But instead we can describe a *complete* automaton $\mathcal{A}$ over A, using the idea of a 'default successor': a state s is distinguished and, for each state p, certain transitions, called *active transitions*, are given explicitly; the letters of A which do not appear as the label of one of these transitions are defined to be the label of a transition from p to s. The size of $\mathcal{A}$ is by definition equal to the number of active transitions. Figure 5.10 shows the automaton $\mathcal{D}_{abaab}$ with default successors.

[54] *Cf.* also Note II.75, p. 342.

[55] *Note added in translation:* Not to mention Unicode, which has anywhere up to 2^{32} letters.

The construction of an automaton for finding f with a default successor is due to Imre Simon. It can be shown that its size is linear in $|f|$. Furthermore, for such an automaton to read a word t requires that for each letter read the transition can be found which is labelled with that letter and leaves from the state reached at that point in the reading; that implies a certain number of comparisons. It can also be shown that the list of active transitions which leave from each state can be organised in such a way that the total number of comparisons required to read a word t by Simon's automaton is no greater than the number of comparisons made by KMP applied to the same word t.

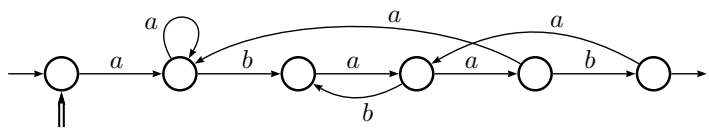


Figure 5.10: The automaton $\mathcal{D}_{a\,b\,a\,a\,b}$ (again); the default successor state is shown by a double incoming arrow.

Exercise

5.7 **Simon's automaton.** We want to show that the number of active transitions of the automaton that finds f is less than $2\,|f|$.

(a) Recall that the *period* of a word f is an integer p such that for all i, $1 \leqslant i \leqslant |f| - p$, we have $f_{i+p} = f_i$. Write $\mathsf{per}(f)$ for the *shortest period* of f and $f\beta$ for the *border* of f. Show that, for all f, $\mathsf{per}(f) + |f\beta| = |f|$. *cf. Exer. 0.3.6, p. 23*

(b) The states of $\mathcal{D}_f$ are the prefixes of f. A transition (p, a, q) in $\mathcal{D}_f$ is active if $q \neq 1_{A^*}$; an active transition is called 'forward' if $q = p\,a$ and 'backward' otherwise. Show that if (p, a, q) and (r, b, s) are two backward transitions, then $\mathsf{per}\,p\,a \neq \mathsf{per}\,q\,b$.

(c) Draw conclusions.

6 Star height

The object of this section is two-fold. First, to present a sort of refinement of Kleene's Theorem. Not only do automata and expressions correspond to each other and characterise the same class of languages, but for both of them a measure of complexity can be defined: *loop complexity* for automata and *star height* for expressions, which we shall also show to correspond to each other (Eggan's Theorem).

Next we shall tackle the problem of applying this measure to languages themselves. The determination of the star height of a language turns out to be one of the most difficult problems in automata theory. We shall confine ourselves here to its presentation and the proof, due to Dejean and Schützenberger, that this notion really corresponds to a measure of complexity; that is, that there are rational languages with an arbitrarily large star height. We will end by touching on another, even more mysterious notion, the *generalised star height*, which may well flatten the family of rational languages.

cf. Sec. II.8, p. 342

In the next chapter we will see how to calculate the star height of a particular family of languages. The solution of the general case, due to K. Hashiguchi, is left to another volume,[56] but in the next chapter we shall at least see why it is difficult to justify its being entirely satisfactory.[57]

6.1 Two heights and a degree

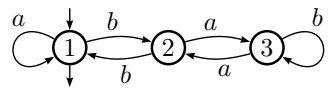
We have seen that the same language can be represented by distinct rational expressions and that, except for finite languages, we can unroll an infinity of these expressions.

We have also seen that if we choose an automaton $\mathcal{A}$ that recognises a language L – its minimal deterministic automaton, for example – the different methods of calculating, from $\mathcal{A}$, a rational expression that denotes L lead to distinct expressions. Consider for example $\mathcal{P}_2$, the 'divider by 3'. Figure 6.1 shows the result of applying the McNaughton–Yamada algorithm to the states of $\mathcal{P}_2$ using three different orders.

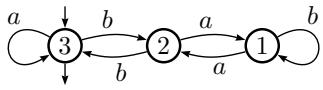
cf. Exer. 2.15, solution p. 191

The question naturally arises therefore whether, from all the expressions that denote L, we can designate one to be canonically associated with L, or failing that, if we can find common characteristics in all the expressions that can be linked to L.

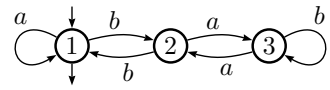
cf. Exer. 6.1



(a) $L(\mathcal{P}_2) = a^* + a^*b\,(b\,a^*b)^*b\,a^* + a^*b\,(b\,a^*b)^*a\,(b + a\,(b\,a^*b)^*a)^*a\,(b\,a^*b)^*b\,a^*$



(b) $L(\mathcal{P}_2) = (a + b\,(a\,b^*a)^*b)^*$



(c) $L(\mathcal{P}_2) = a^* + a^*b\,(a\,b^*a + b\,a^*b)^*b\,a^*$

Figure 6.1: The McNaughton–Yamada algorithm on $\mathcal{P}_2$

6.1.1 Star height of an expression

We defined rational expressions (over an alphabet A) as well-formed formulas made from atomic formulas, which are $\mathsf{0}$, $\mathsf{1}$ and the elements of A, and from the binary operators $+$ and $\cdot$ and the unary operator $*$.

The operator $*$ is the one that 'gives access to the infinite', hence the idea of measuring the complexity of expressions by counting the number of nested uses of this operator, a number which is usually called the *star height*, which we will write $\mathsf{h}[\mathsf{E}]$ (for $\mathsf{E} \in \mathsf{RatE}\,A^*$) and which is defined by induction:

[56] Of all such promises, this is without doubt the most presumptuous.

[57] *Note added in translation:* D. Kirsten has recently given (in [132] and subsequent works) a new proof of Hashiguchi's result which seems far more intelligible. My promise now seems within reach of fulfilment.

$$\text{if } \mathsf{E} = 0,\ \mathsf{E} = 1 \text{ or } \mathsf{E} = a \in A, \qquad \mathsf{h}[\mathsf{E}] = 0 \ , \tag{6.1}$$

$$\text{if } \mathsf{E} = \mathsf{E}' + \mathsf{E}'' \text{ or } \mathsf{E} = \mathsf{E}' \cdot \mathsf{E}'', \qquad \mathsf{h}[\mathsf{E}] = \mathsf{max}(\mathsf{h}[\mathsf{E}'], \mathsf{h}[\mathsf{E}'']) \ , \tag{6.2}$$

$$\text{if } \mathsf{E} = \mathsf{F}^*, \qquad \mathsf{h}[\mathsf{E}] = 1 + \mathsf{h}[\mathsf{F}] \ . \tag{6.3}$$

Examples 6.1 (i) $\mathsf{h}[(a+b)^*] = 1\,;\qquad \mathsf{h}[a^*\,(b\,a^*)^*] = 2\,.$

(ii) The heights of the three expressions in Figure 6.1 are

$\mathsf{h}[a^* + a^*b\,(b\,a^*b)^*b\,a^* + a^*b\,(b\,a^*b)^*a\,(b + a\,(b\,a^*b)^*a)^*a\,(b\,a^*b)^*b\,a^*] = 3\,,$

$\mathsf{h}[(a + b\,(a\,b^*a)^*b)^*] = 3\,;\qquad \mathsf{h}[a^* + b\,(a\,b^*a + b\,a^*b)^*b\,a^*] = 2\,.$ □

By induction on h, the distributivity of sum over product implies:

Property 6.1 *Every rational expression* E *over* A *of height less than or equal to* h *is equivalent (modulo* $\mathbf{D}$*) to a finite sum of expressions:*

$$\mathsf{E} = \sum_{i \in I} \mathsf{E}_i \ ,$$

where each $\mathsf{E}_i = \mathsf{F}$ *is a product of the form*

$$\mathsf{F} = u_0\,(\mathsf{F}_1)^*\,u_1\,(\mathsf{F}_2)^*\,u_2 \cdots u_{k-1}\,(\mathsf{F}_k)^*\,u_k \ ,$$

where each u_j, $0 \leqslant j \leqslant k$, *is a word in* A^* *and each* F_j, $1 \leqslant j \leqslant k$, *is a rational expression over* A *of height less than or equal to* $h-1$, *and of the same form.* ■

The expressions in Examples 6.1 have this form, which one could call normalised.

6.1.2 Star height of a language

The preceding examples draw attention to the fact that two equivalent expressions (that is, that denote the same language) can have different star heights.

Definition 6.1 The star height of a rational language L over A^*, written $\mathsf{h}[L]$, is the minimum of the star heights of the rational expressions that denote L:

$$\mathsf{h}[L] = \mathsf{min}\{\mathsf{h}[\mathsf{E}] \mid \mathsf{E} \in \mathsf{Rat}\mathsf{E}\,A^* \quad |\mathsf{E}| = L\} \ . \qquad \square$$

This definition brings with it two questions:

- Are there rational languages with arbitrarily large star heights?
- Can the star height of a rational language given by, for example, a finite automaton be calculated effectively?

The answer to the first question, which is affirmative, will be given a little later in a purely combinatorial fashion, and we shall return to it in the next chapter. The answer to the second, also affirmative, constitutes without doubt, as we have said,

the most complex result in automata theory. Finding an illuminating and intelligible proof remains an open problem.[58]

cf. for example the proofs of Th. 6.2 below or of Th. IV.6.2, p. 598.

Even if we do not know how to calculate the star height of a given language L, the fact of knowing that the height must be less than or equal to that of an expression that denotes L allows us to show some properties of rational languages by induction, for example starting from a statement such as Property 6.1. We will have occasion to do this several times.

6.1.3 Loop complexity of an automaton

The star height of an expression also reflects a structural property of an automaton which corresponds to this expression, or rather of the graph underlying the automaton.

We shall call the the strongly connected components of a graph $\mathcal{G} = \langle S, V \rangle$ that contain at least one edge the *ball* of $\mathcal{G}$. In other words, a strongly connected component that contains at least two nodes is a ball, and a strongly connected component reduced to a single node s is a ball if and only if s is the source (and the destination) of at least one loop. Balls are pairwise disjoint but do not form a covering (hence a partition) of S since a node may belong to no ball (*cf.* Figure 6.2).

(a) (b)

Figure 6.2: A graph and its balls

Definition 6.2 The *loop complexity* of a graph $\mathcal{G} = \langle S, V \rangle$ is an integer $\mathsf{lc}(\mathcal{G})$ defined recursively by the following equations:

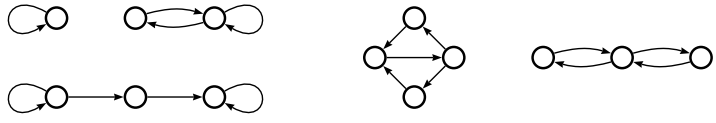
$\mathsf{lc}(\mathcal{G}) = 0$ if $\mathcal{G}$ contains no balls (in particular if $\mathcal{G}$ is empty);

$\mathsf{lc}(\mathcal{G}) = \mathsf{max}\{\mathsf{lc}(\mathcal{P}) \mid \mathcal{P} \text{ a ball in } \mathcal{G}\}$ if $\mathcal{G}$ is not strongly connected;

$\mathsf{lc}(\mathcal{G}) = 1 + \mathsf{min}\{\mathsf{lc}(\mathcal{G} \setminus \{s\}) \mid s \text{ node of } \mathcal{G}\}$ if $\mathcal{G}$ is strongly connected. □

Examples 6.2

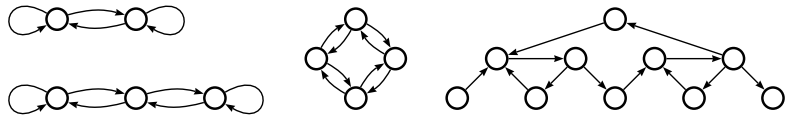
(i) Graphs with loop complexity 1:



(a) (b) (c)

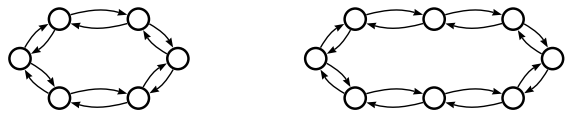
[58]See K. Hashiguchi [117] for the original proof. *Note added in translation:* Kirsten's proof (see previous footnote) may answer this problem.

(ii) Graphs with loop complexity 2:



(d) (e) (f)

(iii) Graphs with loop complexity 3:



(g) (h)

Since an automaton $\mathcal{A}$ is a graph (the first definition of this chapter), we will refer to the balls of $\mathcal{A}$ and to the loop complexity of $\mathcal{A}$, written $\mathsf{lc}(\mathcal{A})$. But as it is a property of the underlying graph, $\mathsf{lc}(\mathcal{A})$ depends on neither the transition labels nor on the choice of initial or final states. Similarly, if $\mathcal{A}_{\mathsf{nor}}$ denotes the normalised automaton of $\mathcal{A}$, we have:

Property 6.2 $\mathsf{lc}(\mathcal{A}_{\mathsf{nor}}) = \mathsf{lc}(\mathcal{A})$. ■

Exercises

• 6.1 (a) Give an algorithm which calculates the shortest expression that denotes a given rational language, or rather the expression which is, among the shortest ones, the smallest in a given lexicographic order.

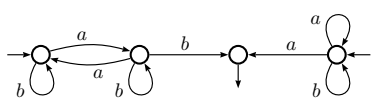
(b) Explain why this 'minimal' expression, which is therefore canonical in some sense, has no algorithmic interest.

6.2 Justify the assertions of Example 6.2: that is, verify the stated loop complexities.

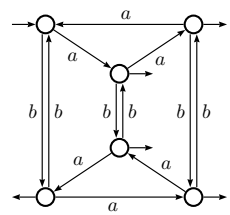
• 6.3 Calculate the loop complexity of $\mathcal{D}_n$, the 'double chain' of length n, and of $\mathcal{R}_n$, the 'double ring' of length n [Example 6.2 (c) is a double chain of length 3, and 6.2 (g) a double ring of length 6].

6.4 Calculate the loop complexity of the following automata:

(a)



(b)



• 6.5 Imitating Definition 6.1, we can define the star height of a rational subset of any monoid:

cf. Sec. II.1.2, p. 220 for the definition of rational subset in any monoid.

Definition 6.3 The star height of a rational subset P of a monoid M, written $\mathsf{h}[P]$, is the minimum star height of the rational expressions over M that denote P:

$$\mathsf{h}[P] = \min\{\mathsf{h}[\mathsf{E}] \mid \mathsf{E} \in \mathsf{Rat}\mathsf{E}\, M \quad |\mathsf{E}| = P\} \ . \qquad \square$$

Show that if M is commutative, the star height of every infinite rational subset is equal to 1.

6.2 Eggan's Theorem

What relationship can we establish between the loop complexity of an automaton and, *if not the star height of the language* which it recognises and which we have said is nearly impossible to find, then *at least the star height of the expressions* that can be calculated from the automaton? This is the aim of the following result.

Theorem 6.1 *The loop complexity of a trim automaton $\mathcal{A}$ is the lower bound of the star height of the expressions of $L(\mathcal{A})$ given by the different possible applications of the McNaughton–Yamada algorithm.*

Conversely, we can associate with every rational expression E *an automaton whose loop complexity is equal to the star height of* E *(and which recognises the language denoted by* E*).*

cf. Prop. 4.8, p. 135

Notice that the properties shown in Section 4 ensure that we can replace the McNaughton–Yamada algorithm above by the state elimination method, or BMC algorithm. In fact it is with this last that we shall perform the computations that lead to the proof.

6.2.1 From expressions to automata

The converse part of the theorem, which is the more obvious, is proved by induction on the depth of expressions, by a construction adapted from that of Subsection 2.3. The base cases, which are the empty word and each letter of the alphabet, are recognised by acyclic automata, which therefore have zero loop complexity (Equation (6.1)). The loop complexity of the union and concatenation of two normalised automata is clearly the maximum of those of the two automata (Equation (6.2)). Star is dealt with by the following construction.

cf. p. 90 et seq.

cf. Exer. 6.6

Let, as in Figure 6.3, $\mathcal{A}$ be a normalised (and trim) automaton, $\mathcal{A}'$ and $\mathcal{A}''$ be two distinct copies of $\mathcal{A}$, and $\mathcal{B}$ be the automaton obtained by concatenating $\mathcal{A}'$ and $\mathcal{A}''$ then identifying the initial state with the final state. If we specify further that the bridge state between $\mathcal{A}'$ and $\mathcal{A}''$ is also final, it is obvious that $L(\mathcal{B}) = (L(\mathcal{A}))^*$ and hence that if $\mathcal{A}$ is associated with E, we can associate $\mathcal{B}$ with E^*. It therefore follows, so finishing the proof of this direction of Theorem 6.1, that:

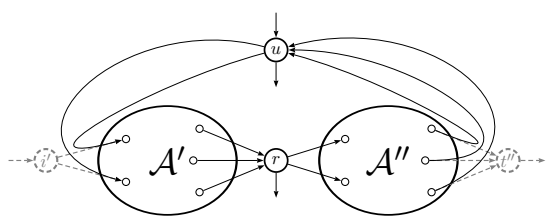


Figure 6.3: A ball that increases the loop complexity

Lemma 6.1 $\mathsf{lc}(\mathcal{B}) = \mathsf{lc}(\mathcal{A}) + 1$.

Proof. The automaton $\mathcal{B}$ is a ball; by definition, we have

$$\mathsf{lc}(\mathcal{B}) = 1 + \mathsf{min}\{\mathsf{lc}(\mathcal{B} \setminus s) \mid s \in \mathcal{B}\} \ .$$

If we set $s = r$ or $s = u$, the balls of $\mathcal{B} \setminus s$ are those of $\mathcal{A}'$ and $\mathcal{A}''$ and hence $\mathsf{lc}(\mathcal{B}) \leqslant \mathsf{lc}(\mathcal{A}) + 1$. If we take an arbitrary s in $\mathcal{A}'$ (or $\mathcal{A}''$), $\mathcal{B} \setminus s$ contains a copy of $\mathcal{A}$ with $\mathcal{A}''$ (or $\mathcal{A}'$), so $\mathsf{lc}(\mathcal{B}) \geqslant \mathsf{lc}(\mathcal{A}) + 1$. ■

6.2.2 From automata to expressions: calculating the index

To prove the forward direction of the theorem, we shall define the *index* of an automaton $\mathcal{A}$, which is both a generalisation and a refinement of loop complexity.

First, and as in Section 2.4, we will consider some *generalised automata*: that is, some automata whose transitions are labelled by rational expressions. If $\mathcal{A}$ is such an automaton, then the *index of a transition* e of $\mathcal{A}$, written $\mathsf{i}(e)$, is the star height of the label of e: *cf. p. 97*

$$\mathsf{i}(e) = \mathsf{h}[|e|] \ .$$

If $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is a 'classic' automaton over A – that is, an automaton all of whose transitions are labelled with letters – then we have

$$\forall e \in E \qquad \mathsf{i}(e) = 0 \ .$$

Next, we will define the index of an automaton $\mathcal{A}$, not absolutely, but *relative to a total order* ω on the set Q of states of $\mathcal{A}$, the same order that underlies the computation of a rational expression for $L(\mathcal{A})$, by the state elimination algorithm or the McNaughton–Yamada algorithm.

We will use the following notation and conventions. If ω is an order on Q, we write $\overline{\omega}$ for the *greatest element* of Q according to ω. If $\mathcal{S}$ is a sub-automaton of $\mathcal{A}$, we also write ω for the trace of the order ω over the set R of states of $\mathcal{S}$ and, in such a context, $\overline{\omega}$ for the greatest element of R according to ω.

Then, the *index of* $\mathcal{A}$ *relative to* ω, written $\mathsf{i}(\mathcal{A}, \omega)$, is the integer defined by the following algorithm:

- if $\mathcal{A}$ is not itself a ball then

$$\mathsf{i}(\mathcal{A}, \omega) = \mathsf{max}\big(\{\mathsf{i}(e) \mid e \text{ does not belong to a ball in } \mathcal{A}\} \cup \{\mathsf{i}(\mathcal{P}, \omega) \mid \mathcal{P} \text{ ball in } \mathcal{A}\}\big) \ ; \tag{6.4}$$

- if $\mathcal{A}$ is a ball then

$$\mathsf{i}(\mathcal{A}, \omega) = 1 + \mathsf{max}\big(\{\mathsf{i}(e) \mid e \text{ is adjacent to } \overline{\omega}\}, \mathsf{i}(\mathcal{A} \setminus \overline{\omega}, \omega)\big) \ . \tag{6.5}$$

The difference with respect to Definition 6.2 is that the state that we remove from a strongly connected automaton (to proceed to the induction) is fixed by the order ω rather than being the result of a minimisation. If $\mathcal{A}$ is a 'classic' automaton, (6.4) and (6.5) become respectively:

- if $\mathcal{A}$ is not itself a ball then

$$\mathsf{i}(\mathcal{A},\omega) = \mathsf{max}\big(\{\mathsf{i}(\mathcal{P},\omega) \mid \mathcal{P} \text{ ball in } \mathcal{A}\}\big) \ ; \tag{6.6}$$

- if $\mathcal{A}$ is a ball then

$$\mathsf{i}(\mathcal{A},\omega) = 1 + \mathsf{i}(\mathcal{A} \setminus \overline{\omega},\omega) \ . \tag{6.7}$$

To these equations must be added, for the induction's base case:

- if $\mathcal{A}$ contains no ball, or is empty then

$$\mathsf{i}(\mathcal{A},\omega) = 0 \ . \tag{6.8}$$

From this we conclude immediately:

Property 6.3 $\mathsf{lc}(\mathcal{A}) = \mathsf{min}\{\mathsf{i}(\mathcal{A},\omega) \mid \omega \textit{ is an order on } Q\ \}$. ■

(a) $\mathsf{i}(\mathcal{B}_1,\omega_1) = 2$; $\mathsf{i}(\mathcal{B}_1,\omega_2) = 3$ (b) $\mathsf{i}(\mathcal{B}_2,\omega_1) = \mathsf{i}(\mathcal{B}_2,\omega_2) = 3$

Figure 6.4: Computation of the index of two automata for the two possible orders on the states: $\omega_1 = p < q$, $\omega_2 = q < p$.

cf. Sec. 4.3, p. 134

Furthermore, we write $\mathsf{E}_{\mathrm{B\,MC}}(\mathcal{A},\omega)$ for the rational expression obtained by applying the state elimination method with order ω to $\mathcal{A}$: that is, by removing the states of $\mathcal{A}$ starting at the *smallest*. The forward direction of the theorem is a consequence of the following proposition.

Proposition 6.2 *Let ω be a total order on the states of an automaton $\mathcal{A}$. The index of $\mathcal{A}$ relative to ω is equal to the star height of the rational expression obtained by applying the state elimination method with order ω to $\mathcal{A}$; that is,*

$$\mathsf{i}(\mathcal{A},\omega) = \mathsf{h}[\mathsf{E}_{B\,MC}(\mathcal{A},\omega)] \ .$$

Proof. We shall proceed by induction on the number of states of $\mathcal{A}$. The state elimination method consists in the first place of transforming $\mathcal{A}$ into an automaton $\mathcal{B}$ by adding two states to $\mathcal{A}$, an initial state and a final state, and transitions which are all labelled by the empty word, the index of $\mathcal{B}$ being equal to that of $\mathcal{A}$. By convention, the added states are greater than all the other states of $\mathcal{A}$ in the order ω and are never removed by the state elimination method. On the other hand, the transition labels, including those of the new transitions, may be modified in the course of the state elimination method.

The base case of the induction is therefore a generalised automaton with 3 states of the form of Figure 6.5 (a) or (b).

In case (a), $\mathcal{B}$ contains no balls and we have

$$\mathsf{i}(\mathcal{B},\omega) = \mathsf{max}\big(\mathsf{h}[\mathsf{E}],\mathsf{h}[\mathsf{F}],\mathsf{h}[\mathsf{H}]\big) = \mathsf{h}[\mathsf{E}+\mathsf{F}\cdot\mathsf{H}] = \mathsf{h}[\mathsf{E}_{\mathrm{B\,MC}}(\mathcal{B},\omega)] \ .$$

In case (b), the unique state of $\mathcal{B}$ that is neither initial nor final is a ball whose index is $1+\mathsf{h}[\mathsf{G}]$, and we have

$$\mathsf{i}(\mathcal{B},\omega) = \mathsf{max}\big(\mathsf{h}[\mathsf{E}],\mathsf{h}[\mathsf{F}],\mathsf{h}[\mathsf{H}],(1+\mathsf{h}[\mathsf{G}])\big) = \mathsf{h}[\mathsf{E}+\mathsf{F}\cdot\mathsf{G}^*\cdot\mathsf{H}] = \mathsf{h}[\mathsf{E}_{\mathrm{B\,MC}}(\mathcal{B},\omega)] \ .$$

Note that this reasoning is the essential part of the induction step and that for rigour, if not for clarity, we could have taken the automaton of Figure 6.5(c) as our base case, for which the statement is even more easily verified (and which corresponds to an automaton $\mathcal{A}$ with no state).

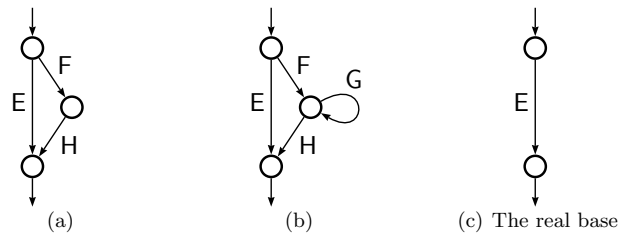


Figure 6.5: Base case of the induction

Now let $\mathcal{B}$ be an automaton of the prescribed form with $n+2$ states, q the smallest state in the order ω and $\mathcal{B}'$ the automaton which results from the first step of the state elimination method applied to $\mathcal{B}$ (which consists of the elimination of q). Because the same states (other than q) are adjacent in $\mathcal{B}$ as in $\mathcal{B}'$, and since q is the *smallest element* in the order ω, the algorithm for computing the index runs the same way in $\mathcal{B}$ and $\mathcal{B}'$: that is, the succession of balls constructed in each automaton is identical, excluding the examination of q in $\mathcal{B}$. It remains to show that the values calculated are also identical.

Let $\mathcal{P}$ be the smallest ball in $\mathcal{B}$ that strictly contains q – if no such ball exists, take $\mathcal{P}=\mathcal{B}$; and let $\mathcal{P}'$ the 'image' of $\mathcal{P}$ in $\mathcal{B}'$. There are two possible cases: either (a) q is not the source (and the destination) of a loop; or (b) it is – that is, q is a ball in $\mathcal{B}$ all by itself – and the label of this loop is an expression G.

The transitions of $\mathcal{P}'$ are either identical to those of $\mathcal{P}$, or, in case (a), labelled by products $\mathsf{F}\cdot\mathsf{H}$, where F and H are labels of transitions of $\mathcal{P}$, or, in case (b), labelled by products $\mathsf{F}\cdot\mathsf{G}^*\cdot\mathsf{H}$. It therefore follows that, in case (a)

$$\begin{aligned}
\mathsf{i}(\mathcal{P}',\omega) &= \mathsf{max}\big(\mathsf{max}\{\mathsf{i}(e) \mid e \text{ does not belong to a ball in } \mathcal{P}'\}, \\
&\qquad\qquad \mathsf{max}\{\mathsf{i}(\mathcal{Q},\omega) \mid \mathcal{Q} \text{ ball in } \mathcal{P}'\}\big) \\
&= \mathsf{max}\big(\mathsf{max}\{\mathsf{i}(e) \mid e \text{ does not belong to a ball in } \mathcal{P}\}, \\
&\qquad\qquad \mathsf{max}\{\mathsf{i}(\mathcal{Q},\omega) \mid \mathcal{Q} \text{ ball in } \mathcal{P}\}\big) \\
&= \mathsf{i}(\mathcal{P},\omega) \ . \qquad (6.9)
\end{aligned}$$

In case (b), since $\mathsf{i}(\{q\},\omega) = 1 + \mathsf{h}[\mathsf{G}]$,

$$\begin{aligned}
\mathsf{i}(\mathcal{P}',\omega) &= \mathsf{max}\big(\ \mathsf{max}\{\mathsf{i}(e) \mid e \text{ does not belong to a ball in } \mathcal{P}'\}, \\
&\qquad\qquad\qquad\qquad\qquad\qquad \mathsf{max}\{\mathsf{i}(\mathcal{Q},\omega) \mid \mathcal{Q} \text{ ball in } \mathcal{P}'\}\ \big) \\
&= \mathsf{max}\big(\ \mathsf{max}\{\mathsf{i}(e) \mid e \text{ does not belong to a ball in } \mathcal{P}\},\ (1+\mathsf{h}[\mathsf{G}]), \\
&\qquad\qquad\qquad\qquad\qquad\qquad \mathsf{max}\{\mathsf{i}(\mathcal{Q},\omega) \mid \mathcal{Q} \text{ ball in } \mathcal{P}'\}\ \big) \\
&= \mathsf{max}\big(\ \mathsf{max}\{\mathsf{i}(e) \mid e \text{ does not belong to a ball in } \mathcal{P}\},\ \mathsf{i}(\{q\},\omega), \\
&\qquad\qquad\qquad\qquad \mathsf{max}\{\mathsf{i}(\mathcal{Q},\omega) \mid \mathcal{Q} \text{ ball in } \mathcal{P}, \text{ different from } \{q\}\}\ \big) \\
&= \mathsf{i}(\mathcal{P},\omega)\ . \qquad\qquad (6.9)'
\end{aligned}$$

If $\mathcal{P} = \mathcal{B}$ (and $\mathcal{P}' = \mathcal{B}'$), the equalities (6.9) and (6.9)′ become

$$\mathsf{i}(\mathcal{B}',\omega) = \mathsf{i}(\mathcal{B},\omega) \qquad (6.10)$$

which proves the induction and hence the proposition. If not, and without starting an induction on the number of overlapping balls that contain $\{q\}$, we can get from (6.9) to (6.10) by noting that the transitions of $\mathcal{B}'$ are either identical to those of $\mathcal{B}$, or correspond to transitions that are adjacent to q.

In case (a), the labels of these transitions (those corresponding to transitions adjacent to q) are products of transitions of $\mathcal{B}$: their index is obtained by taking a maximum, and (6.10) is the result of the identity $\mathsf{max}(a,b,c) = \mathsf{max}(a, \mathsf{max}(b,c))$.

In case (b), the labels of the same transitions are, as before, of the form $\mathsf{F}\cdot\mathsf{G}^*\cdot\mathsf{H}$, with index $\mathsf{max}(\mathsf{h}[\mathsf{F}], \mathsf{h}[\mathsf{H}], 1+\mathsf{h}[\mathsf{G}])$. The corresponding transition in $\mathcal{B}$ has label F (or H); it is inspected in the algorithm for computing the index when the indices of the transition with label H (or F) and that of the *ball* $\{q\}$, with index $1+\mathsf{h}[\mathsf{G}]$, have already been taken into account. The result, which is (6.10), follows for the same reason as above. ■

6.2.3 Not so fast

After Theorem 6.1 showed that the correspondence between automata and expressions extends to a correspondence between loop complexity and star height, we might have thought that there would be an *expression of minimal star height* corresponding to the minimal automaton. Clearly, there is nothing of the sort: the problem of star height would not have been described as formidable if its solution could be obtained by a simple reduction. The following example, to which we shall return, is one of the simplest where the minimal automaton is not the automaton with minimal loop complexity.

cf. Sec. II.8, p. 342

Example 6.3 Let E_2, E_3 and E_6 be languages over $A^* = \{a,b\}^*$ formed from words whose number of a's is not congruent to the number of b's modulo 2, 3 and 6 respectively:

$$E_2 = \{f \mid |f|_a \not\equiv |f|_b \bmod 2\}\ , \quad E_3 = \{f \mid |f|_a \not\equiv |f|_b \bmod 3\}$$
$$\text{and} \qquad E_6 = \{f \mid |f|_a \not\equiv |f|_b \bmod 6\}\ .$$

We obviously have $\qquad E_6 = E_2 \cup E_3$.

The minimal automaton of E_6 is the double ring of length 6, whose loop complexity is 3. The minimal automata of E_2 and E_3 have loop complexity 1 and 2, thus the star height of E_6 is at most 2 (*cf.* Figure 6.6). □ *cf. Exam. 6.2 (f)*

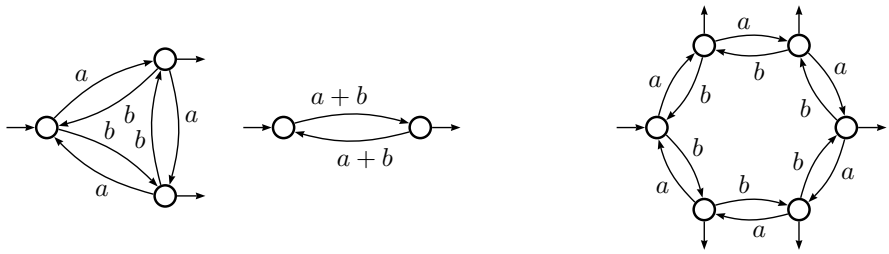
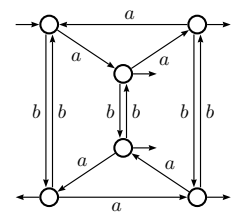


Figure 6.6: An automaton with minimal loop complexity (left) which is not the minimal automaton (right) for E_6

Exercises

• 6.6 Show, using one of Examples 6.2, that the standard construction of an automaton $\mathcal{B}$ that recognises the star of a language recognised by an automaton $\mathcal{A}$ does not necessarily lead to an increase in the loop complexity.

6.7 Verify, in imitation of Example 6.3, that the star height of the language recognised by the automaton in the adjacent figure is not equal to the automaton's loop complexity.



6.3 An infinite hierarchy

The answer to the question whether there are languages with arbitrarily large star height is positive:

Theorem 6.2 *The language W_q over $\{a,b\}^*$ consisting of words whose number of a's is congruent to the number of b's modulo 2^q has star height q.* *cf. also Cor. II.8.2, p. 342*

Proof. First we must give an expression of star height q which denotes W_q, which is easy. The language W_q is recognised by the 'ring automaton' $\mathcal{W}_q$ of length 2^q. We can check that (judicious) use of the state elimination method on $\mathcal{W}_q$ gives an expression of height q for W_q (*cf.* Figure 6.7). First set *cf. also Exer. 6.3 and Th. 6.1*

$$X_1 = a^2 \ , \quad Y_1 = b^2 \qquad \text{and} \qquad Z_1 = a\,b + b\,a \ ;$$

then, for every integer n,

$$X_{n+1} = X_n\,Z_n^*\,X_n \ , \quad Y_{n+1} = Y_n\,Z_n^*\,Y_n \ , \quad \text{and} \quad Z_{n+1} = Z_n + X_n\,Z_n^*Y_n + Y_n\,Z_n^*X_n \ .$$

It therefore follows that

$$W_q = (X_q + Y_q + Z_q)^* \ .$$

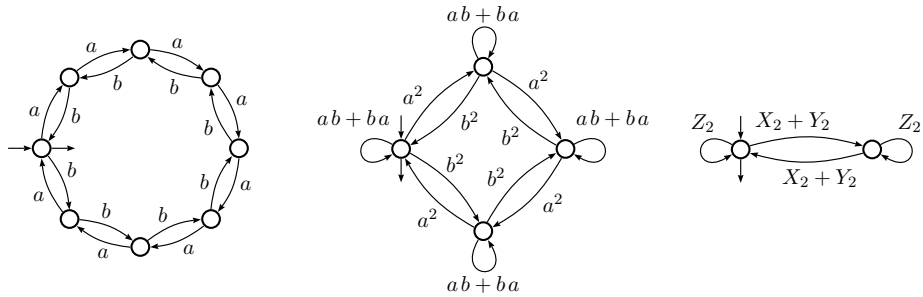


Figure 6.7: The (judicious) application of the state elimination method to $\mathcal{W}_3$

The proof that W_q is of height at least q is far more subtle. For each integer n, let $w_{0,n}$, $w_{1,n}$, ..., $w_{q-1,n}$ be the sequence of q words in W_q recursively defined as follows:

$$w_{0,n} = a\,b \ , \quad w_{1,n} = a^2\,(a\,b)^n\,b^2\,(a\,b)^n \ , \ \ldots \ ,$$
$$w_{k,n} = a^{2^k}\,(w_{k-1,n})^n\,b^{2^k}\,(w_{k-1,n})^n \ , \ \ldots \ , \ w_{q-1,n} = a^{2^{q-1}}\,(w_{q-2,n})^n\,b^{2^{q-1}}\,(w_{q-2,n})^n \ .$$

Note an elementary combinatorial property of these words $w_{k,n}$ which we will use later:

Property 6.4 *Any left factor u and right factor v of $(w_{k,n})^n$ satisfy the equations*

$$0 \leqslant |u|_a - |u|_b \leqslant 2^{k+1} - 1 \ , \qquad 0 \leqslant |v|_b - |v|_a \leqslant 2^{k+1} - 1 \ . \tag{6.11}$$

From these 'witness words' we define for each k, $k = 0, 1, \ldots, q-1$, the property (P_k) of languages over A^*:

L satisfies (P_k) if there exists an infinite number of values of n such that $(w_{k,n})^n$ is a factor of at least one word in L.

Note that if L satisfies (P_k), it also satisfies $(P_{k'})$ for $k' \leqslant k$ since $(w_{k-1,n})^n$ is a factor of $(w_{k,n})^n$. Then let $\mathfrak{W}_k$ denote,[59] for each k, the family of (rational) languages L over A^* which satisfy the following conditions:

(i) $L \subseteq W_q$;

(ii) L satisfies (P_k) ;

(iii) L has the minimum star height of any language satisfying (i) and (ii).

If h_k is the common value of the star height of the elements of $\mathfrak{W}_k$, we have

$$0 < h_0 \leqslant h_1 \leqslant \cdots \leqslant h_{q-1} \leqslant q \ . \tag{6.12}$$

The (strict) inequality on the left comes from the fact that a language in $\mathfrak{W}_0$ is necessarily infinite, the inner inequalities because (P_k) implies (P_{k-1}), and the rightmost inequality because W_q itself satisfies (i) and (P_{q-1}), and because we have found

[59] Read 'W sub k': $\mathfrak{W}$ is a gothic 'W'.

an expression of star height q for W_q. To prove the theorem it is therefore enough to show that $h_{k-1} < h_k$ for each k, $k = 0, 1, \ldots, q-1$.

Let L be in $\mathfrak{W}_k$. Since L's star height is h_k, it can be written as a finite union of languages F_j of the form

$$F_j = f_0 H_1^* f_1 \cdots H_m^* f_m \quad , \tag{6.13}$$

where the H_i are of star height at least $h_k - 1$. Since there is only a finite number of the F_j, at least one of them satisfies (P_k) and we can assume that L itself is of the form (6.13). We then note, by elementary reasoning about the number of occurrences of a and b, that each of the H_i must be contained in W_q. Since L is of minimal star height, none of the H_i satisfies (P_k). Our aim is to show that there exists at least one that satisfies (P_{k-1}), to prove the theorem. In effect, we will have $h_{k-1} \leqslant h_k - 1$.

Since there exist words $(w_{k,n})^n$ which are factors of L, with n arbitrarily large, we can apply the Block Star Lemma and obtain the result that there is an infinity of n's such that for each one there is an infinity of l's such that $(w_{k,n})^l$ is a factor of L. As the integer m in (6.13) is fixed and finite, we deduce that there exists at least one index i such that H_i^* satisfies (P_k); that is, that for infinitely many n's, $(w_{k,n})^n$ is a factor of a word r_n of H_i^*. We shall show that for these n's, $(w_{k-1,n})^n$ is a factor of a word in H_i, thus that H_i satisfies (P_{k-1}).

Write $r_n = g_0 g_1 \cdots g_l$. If a factor $(w_{k,n})$ (in $(w_{k,n})^n$) is a factor of some g_j, the condition is satisfied (since $(w_{k-1,n})^n$ is a factor of $(w_{k,n})$). Otherwise, let us see how a factor $(w_{k,n})^2$, which we write explicitly, is covered by the factorisation $g_0 g_1 \cdots g_l$. It is here that we can use the particular form of the $w_{k,n}$ and Property 6.4:

$$a^{2^k} (w_{k-1,n})^n b^{2^k} (w_{k-1,n})^n a^{2^k} (w_{k-1,n})^n b^{2^k} (w_{k-1,n})^n$$

The 'central' factor b^{2^k} is covered, at least partially, by a factor g_j. There are two possible cases:

(i) b^{2^k} is a factor of g_j;

(ii) a left factor of b^{2^k} is a right factor of g_j.

In case (i), $g_j = v\, b^{2^k} u$. If v (resp. u) completely covers the factor $(w_{k-1,n})^n$ to the left (resp. to the right) of b^{2^k}, the condition is satisfied. If not, v is a right factor of $(w_{k-1,n})^n$ and u a left factor. Set

$$x = |g_j|_b - |g_j|_a \quad , \quad y = |u|_a - |u|_b \qquad \text{and} \qquad z = |v|_b - |v|_a \quad .$$

Hence

$$x = 2^k - (y - z) \quad .$$

By Property 6.4, $1 - 2^k \leqslant y - z \leqslant 2^k - 1$, so

$$0 < x < 2^{k+1} \leqslant 2^q \quad ,$$

which contradicts the fact that g_j is in H_i and thus in W_q.

In case (ii), $g_j = v\, b^r$ where $r > 0$. If v completely covers the factor $(w_{k-1,n})^n$ to the left of b^{2^k}, the condition is satisfied. If not, v is a right factor of $(w_{k-1,n})^n$ and we have produced the same contradiction as above. ■

6.4 Generalised star height

As the rational languages over A^* are closed under complement (and hence under intersection and difference) it is tempting, when one defines them with expressions, to allow oneself, in addition to the union, product and star operators, the complement operator, and at the same time those of intersection and difference. The expressions thus obtained are called *generalised rational expressions*[60] and we write the set of such expressions $\mathsf{GRatE}\, A^*$. In this context, to make formulas more readable, we write the complement operator with an overbar. As for classic expressions, the language denoted by an expression E, written $|\mathsf{E}|$, is defined in an inductive manner by Equations (4.2) and (4.3) and by the equality:

$$|\overline{\mathsf{E}}| = \complement_{A^*}\, |\mathsf{E}| \tag{6.14}$$

We have, for example:

$$A^* = |\overline{\emptyset}| \qquad \text{and}$$
$$(a\,b)^* = |(a\,A^* \cap A^*b) \setminus (A^*a\,a\,A^* \cup A^*b\,b\,A^*) \cup \{1_{A^*}\}| \ .$$

We can then define the *generalised star height* of a generalised rational expression, parallel to the formulas of § 6.1.3, by induction on the complexity of expressions:

$$\text{if } \mathsf{E} = \mathsf{0},\ \mathsf{E} = \mathsf{1} \text{ or } \mathsf{E} = a \in A, \qquad \mathsf{gsh}[\mathsf{E}] = 0 \ , \tag{6.15}$$

$$\text{if } \mathsf{E} = \mathsf{E}' + \mathsf{E}'' \text{ or } \mathsf{E} = \mathsf{E}' \cdot \mathsf{E}'', \qquad \mathsf{gsh}[\mathsf{E}] = \mathsf{max}(\mathsf{gsh}[\mathsf{E}'], \mathsf{gsh}[\mathsf{E}'']) \ , \tag{6.16}$$

$$\text{if } \mathsf{E} = \mathsf{F}^*, \qquad \mathsf{gsh}[\mathsf{E}] = 1 + \mathsf{gsh}[\mathsf{F}] \tag{6.17}$$

$$\text{if } \mathsf{E} = \overline{\mathsf{F}}, \qquad \mathsf{gsh}[\mathsf{E}] = \mathsf{gsh}[\mathsf{F}] \ . \tag{6.18}$$

We can verify immediately:

Property 6.5 *If* $\mathsf{E} = \mathsf{E}' \cap \mathsf{E}''$ *or* $\mathsf{E} = \mathsf{E}' \setminus \mathsf{E}''$ *, then* $\mathsf{gsh}[\mathsf{E}] = \mathsf{max}\big(\mathsf{gsh}[\mathsf{E}'], \mathsf{gsh}[\mathsf{E}'']\big)$. ■

Still in step with Section 6.1, we define the generalised star height of a rational language:

Definition 6.4 The generalised star height of a rational language L over A^*, written $\mathsf{gsh}[L]$, is the minimum of the generalised star height of the generalised rational expressions that denote L:

$$\mathsf{gsh}[L] = \mathsf{min}\{\mathsf{gsh}[\mathsf{E}] \mid \mathsf{E} \in \mathsf{GRatE}\, A^* \quad |\mathsf{E}| = L\} \ . \qquad \square$$

While the determination of the star height of a rational language is a difficult problem, but one which has been solved, that of generalised star height remains utterly mysterious and the two questions which are raised directly by Definition 6.4,

[60] Beware: the 'generalised' in 'generalised rational expressions' is not the same sort of generalisation as in 'generalised automata'.

namely the existence of an infinite hierarchy and the computation of (generalised) star height, remain open.

Not only do we not know whether there are rational languages with arbitrarily large generalised star height, but we know of no language whose generalised star height can be proved to be greater than 1. On the other hand, we can characterise the languages whose generalised star height is equal to 0, a result that we shall mention in the next chapter.

Exercise

• 6.8 Give a generalised rational expression with generalised star height equal to 1 for the following languages:

(a) $G_2 = (a\,a \cup b\,(a\,a)^*b)^*$; (b) $G_3 = (a\,b^*a \cup b\,(a\,b^*a)^*b)^*$;

(c) $G_4 = (a\,b^*a \cup b\,a^*b\,(a\,b^*a)^*b\,a^*b)^*$.

7 A field of automata

We first defined automata as labelled graphs, then adopted a functional point of view. In both cases we abstracted from real machines, and our abstractions referred neither to any physical realisation of these machines, nor to their mode of operation. The advantage and interest of this final abstraction is to extract the intrinsic properties of the objects studied, to strip the proofs of details incidental to the matter at hand and, as we said when we started, to allow later generalisation of notions already defined.

Nevertheless we still have in our terminology *input* alphabet, *final state*, *reading from left to right*, traces if not of the description of physical computation devices then at least of their modelling in terms less removed from reality. We now turn to these older models, partly to assure ourselves that we can indeed comprehend them in our models, but mainly to see what problems, properties and notions their peculiarities suggest that have disappeared from our purer model.

7.1 The Rabin–Scott model

The first model of finite automata from which the distinction between deterministic and non-deterministic automata was drawn is a Turing machine with limited operational capacities. It is due to M. Rabin and D. Scott.

In this model a finite automaton then consists, as in Figure 7.1, of a *finite control* equipped with a finite memory represented by a finite number of *states* in which the finite control can be, and an arbitrarily long *input tape*, which is divided into cells in which are written symbols from the input alphabet.

The finite control is connected to a *read head* which reads the symbol written in the square over which it is placed. At each step of the computation the finite control, considering the state p in which it is and the symbol a read by the read head, changes to a state q and moves the read head one square to the right. If the state q is determined uniquely by the state p and the letter a, the automaton is *deterministic*;

if, for some state p and some letter a, the finite control can change to more than one state, the automaton is *non-deterministic*.

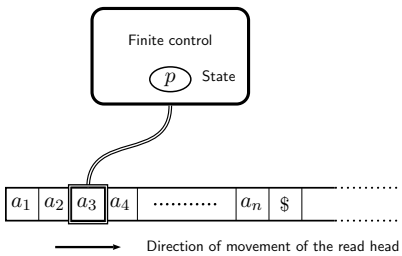


Fig. 7.1: The Rabin–Scott model

Because there is *one* input tape and the read head always moves in the same direction, this type of machine is called a *one-way one-tape automaton*.

At the start of a computation, a word is written on the input tape, the finite control is in a distinguished state, called the *initial state*, and the read head is placed on the first square of the input tape. A computation terminates after a series of steps like that described above when the read head reaches the square in which is written an *endmarker*,[61] represented in Figure 7.1, according to common usage, by a $.

The computation is termed *successful* if the state in which the finite control finds itself after reading the endmarker is a special state, called the *final state*. A word is accepted by the automaton if it can be read by a successful computation.

The equivalence of this model to the functional and thence to the labelled graph model is so obvious that it is unnecessary to develop it further. The interest of this model resides in the questions that arise for different generalisations: the operation of the read head in the next subsection, and the presence of several input tapes in Chapters IV and V.

Exercises

7.1 Explain how to deal with the endmarker when translating from the Rabin–Scott model to the labelled graph model, and *vice versa*.

7.2 Modify the Rabin–Scott model to allow a direct translation of automata with spontaneous transitions (that is, without needing to manufacture a classical automaton beforehand).

7.2 Two-way automaton

The previous model was obtained from a Turing machine by imposing two restrictions: first, the head that moves over the tape can only read and not write; and second, it moves in only one direction. It should be fairly obvious that if we retain the restriction that the head can only move in one direction, then the model's power is not increased by allowing it to write to the tape, and that a machine equipped with one finite control, one tape and one read-write head which moves only from left to right accepts a recognisable language. The question that then arises is this: what is the power of a machine equipped with one finite control, one input tape and a read head that can move in both directions? It is called a *two-way (one-tape) automaton*.[62]

[61] A slightly confusing term, as it could also be placed at the beginning. See below and Section V.1.4.

[62] In French, the more evocative term 'boustrophedonic automaton' is often used, after Schützenberger [219].

At each step of a computation by a two-way automaton, the finite control, according to its current state p and the symbol a read by the read head, changes to a state q and moves the read head one square either left or right. If the state q and the direction of movement of the head are determined uniquely by the state p and the letter a, for all p and a, the automaton is *deterministic*; if not, it is *non-deterministic*.

At the start of a computation, a word f is written on the input tape, the finite control is in the initial state, and the read head is placed is placed on the first square of the input tape (which contains the left \$). We can see that there are three possibilities for the automaton: it can stop by leaving f to the left, compute indefinitely without leaving f, or stop by leaving f to the right.[63] A computation is *successful* if the automaton leaves to the right and is at that moment in the final state. A word is accepted by the automaton if it can be read by a successful computation.

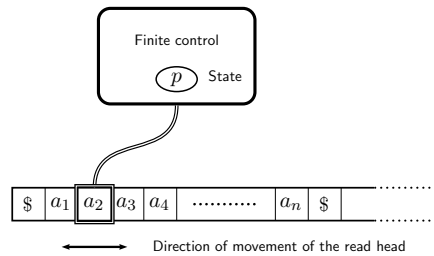


Fig. 7.2: Two-way automaton

Such a machine seems *a priori* more complicated and thus more powerful than a normal finite automaton, hence the interest of the following result:

Theorem 7.1 *The language accepted by a two-way automaton is rational.*

To prove this result we can no longer be content with an informal description of the model and its mode of operation. A two-way automaton is a quintuple $\mathcal{A} = \langle Q, A, \beta, i, t \rangle$ where Q is the set of states, A the input alphabet, i the initial state, t the final state and β the transition function of $\mathcal{A}$:

$$\beta\colon Q \times A \quad \longrightarrow \quad \mathfrak{P}(Q \times \{-1, +1\}) \ .$$

A *configuration*[64] of $\mathcal{A}$ is an element $f\,p\,g$ of $A^*Q\,A^*$; it describes the situation of the automaton where the word $f\,g$ is written on the input tape, the finite control is in state p and the read head is placed over the square that contains the first letter of g. Each step of a computation of $\mathcal{A}$ corresponds to the succession of two snapshots:

$$\forall p \in Q\,,\ \forall a \in A\,,\ \forall f, g \in A^* \qquad \begin{array}{lll} f\,a\,p\,b\,g \underset{\mathcal{A}}{\vdash\!\!\!-\!\!\!-} f\,a\,b\,q\,g & \text{if} & (q, +1) \in (p, b)\beta \\ f\,a\,p\,b\,g \underset{\mathcal{A}}{\vdash\!\!\!-\!\!\!-} f\,q\,a\,b\,g & \text{if} & (q, -1) \in (p, b)\beta \end{array}$$

and the language recognised by $\mathcal{A}$ is defined by: $\quad L(\mathcal{A}) = \{f \in A^* \mid i\,f \overset{\star}{\underset{\mathcal{A}}{\vdash\!\!\!-\!\!\!-}} f\,t\}$.

Proof of Theorem 7.1. We will show that $L(\mathcal{A})$ has only a finite number of (left) quotients, and thus is recognisable, by Proposition 3.10. p. 114

Let x and y be two elements not in Q. We define a map τ from A^+ to *the set of maps* from $Q \cup \mathsf{x}$ to $Q \cup \mathsf{y}$ (a finite set) in the following manner: for all f in A^*, all a

[63] The automaton's computation could also 'block' (in the sense that there is no defined transition); it is not difficult then to imagine an equivalent automaton: that is, one that accepts the same words and never blocks.

[64] Or *snapshot* or *instantaneous description*.

in A and all p in Q, we set[65]

$$\begin{array}{lll}
(p)[(fa)\tau] = q & \text{if} & f\,p\,a \underset{\mathcal{A}}{\overset{\star}{\vdash\!\!\!-}} f\,a\,q \\
(p)[(fa)\tau] = \mathsf{y} & \text{if} & f\,p\,a \quad \text{never leaves} \\
(\mathsf{x})[(fa)\tau] = q & \text{if} & i\,f\,a \underset{\mathcal{A}}{\overset{\star}{\vdash\!\!\!-}} f\,a\,q \\
(\mathsf{x})[(fa)\tau] = \mathsf{y} & \text{if} & i\,f\,a \quad \text{never leaves} \ ,
\end{array}$$

and we will show that if $(fa)\tau = (gb)\tau$, then $(fa)^{-1}L(\mathcal{A}) = (gb)^{-1}L(\mathcal{A})$; that is:

$$\forall h \in A^* \qquad f\,a\,h \in L(\mathcal{A}) \quad \Longleftrightarrow \quad g\,b\,h \in L(\mathcal{A}) \ . \tag{7.1}$$

Verifying (7.1) is elementary. Figure 7.3 shows the mechanism of the proof. There is no relationship between q_1 and r_1; on the contrary, since $(fa)\tau = (gb)\tau$, we have

$$q_2 = (\mathsf{x})[(fa)\tau] = (\mathsf{x})[(gb)\tau] = r_2 \ .$$

The parts of the computation between q_2 and q_4 on one hand, and between r_2 and r_4 on the other, are identical since they are entirely determined by h, hence $q_4 = r_4$ (the computation step from the configurations $f\,a\,q_3\,h$ and $g\,b\,r_3\,h$ depends entirely on the first letter of h, and not on a on one side and b on the other as the figure could lead one to think).

We can say nothing about computations 'inside' the factors $f\,a$ and $g\,b$ respectively but if they leave to the right then they leave *in the same state*, again since $(fa)\tau = (gb)\tau$: $q_5 = (q_4)[(fa)\tau] = (r_4)[(gb)\tau] = r_5$. Thus we know how to make each successful computation that accepts $f\,a\,h$ correspond to a successful computation that accepts $g\,b\,h$ and *vice versa*. ■

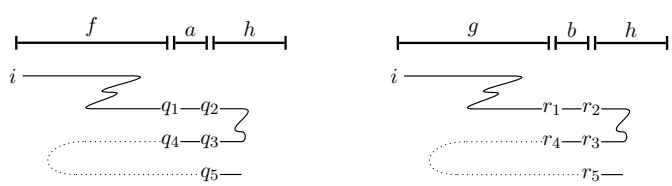


Figure 7.3: Two computations of the two-way automaton

7.3 Moore and Mealy machines

The model of automata defined by Moore is almost exactly the same as the functional version of deterministic automata. Moore's aim was to define the notion of *indistinguishable states* in a machine, which we have called equivalent states according to Nerode's equivalence, and from there to define and compute the *minimal machine* equivalent to a given finite machine.

[65]The notation is a little cumbersome, but it is explicit: $(fa)\tau$ is a map from $Q \cup \mathsf{x}$ to $Q \cup \mathsf{y}$, and $(p)[(fa)\tau]$ is the image of p under this map.

A Moore machine is a deterministic automaton such that every state is associated with a value, called its *output*. Each word over the input alphabet is considered as the sequence of letters that constitutes it. We imagine that these letters are read successively by the machine which changes state as it goes along following its transition function, which is supposed to be unknown, and we observe not the states (which are said to be inaccessible) but the succession of outputs that they cause.

More formally, a Moore machine is a sextuple $\mathcal{A} = \langle Q, A, S, \delta, i, T \rangle$ where Q, A, δ and i are as in a deterministic automaton, S is the alphabet of outputs and T is a *function* from Q to S. If $S = \{0, 1\}$, we can consider a state q to be final if $(q)T = 1$ and not final otherwise (that is, T is the characteristic function of the set of final states) and we have returned exactly to our definition of deterministic automata.

All the observable properties of a state q of $\mathcal{A}$ are properties of the function $\omega_q \colon A^* \to S^*$ defined by induction on the length of words:

$$(1_{A^*})\omega_q = (q)T \qquad \text{and} \qquad (f\,a)\omega_q = (f)\omega_q\big((q, f\,a)\delta\big)T$$

and two states p and q are indistinguishable if $\omega_p = \omega_q$.

The model defined by Mealy is identical to Moore's except that outputs, rather than being associated with the state reached after reading a letter, are associated with a pair (state, letter read). The equivalence of these two models is easy to prove. This model of automata in fact corresponds to that which we will call a *letter-to-letter transducer*, and is studied in Chapter IV. *cf. Sec. IV.6, p. 595*

8 A crop of properties

We are better persuaded, in general,
by reasons we have found ourselves,
than by those given to us by others.
Blaise Pascal

The definitions and results we have seen in this chapter (and particularly those of Section 3) enable us to prove a great number of properties of rational languages. We give here a glimpse of five of them, in the form of meaty exercises.

These properties are better expressed and interpreted with the notions that we will put in place in the following chapters, and will be revisited when the time is ripe. They are nonetheless presented here and proved with the means we now have at our disposal with the double, if contradictory, aim of making them accessible to readers put off by the algebraic tour on which we are about to embark, and to illustrate, by contrast, the power of those same algebraic methods which will allow us to prove them with incomparable elegance.

• 8.1 **Roots of a recognisable language.** The aim of this problem is to show that the set of words whose square, or any fixed power k, is in a recognisable language, is itself an effectively computable recognisable language.

For every language K over A^*, write $\mathcal{R}_2(K)$ for the set of words whose square[66] is in K, and $\mathcal{R}_k(K)$ for the set of words whose kth power is in K:

$$\mathcal{R}_2(K) = \{f \in A^* \mid f\,f \in K\} \ , \qquad \mathcal{R}_k(K) = \{f \in A^* \mid f^k \in K\} \ .$$

(a) *First method.* Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ be a deterministic automaton which recognises K.

(i) Show that $\mathcal{R}_2(K)$ is recognisable by expressing it in terms of $L_p = \{f \mid p \cdot f \in T\}$ and $L'_q = \{g \mid i \cdot g = q\}$, for all p and q in Q.

(ii) Generalisation. Show that $\mathcal{R}_k(K)$ is recognisable for all k.

(b) *Left regular automaton.* Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ be a deterministic automaton; by definition, $\mathcal{A}$ is *right regular* in the sense that we have

$$\forall f, g \in A^* \qquad i \cdot f = i \cdot g \quad \Longrightarrow \quad [\,\forall v \in A^* \quad i \cdot f\,v = i \cdot g\,v\,] \ . \tag{8.1}$$

We say that $\mathcal{A}$ is *left regular* if we have

$$\forall f, g \in A^* \qquad i \cdot f = i \cdot g \quad \Longrightarrow \quad [\,\forall u \in A^* \quad i \cdot u\,f = i \cdot u\,g\,] \ . \tag{8.2}$$

By a construction similar to that of determinisation by the subset method, we will show:

Property 8.1 *Every complete deterministic automaton is equivalent to a complete deterministic left regular automaton.*

Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ be a finite complete deterministic automaton. We shall construct an automaton whose set of states is (a subset of) $\mathfrak{T}(Q)$, the set of all the maps from Q to itself. Each element r of $\mathfrak{T}(Q)$ is identified with a row vector of dimension Q whose pth entry r_p is the image under r of $p \in Q$. The identity map is written ε: $\varepsilon_p = p$ for all p in Q. We therefore define $\mathcal{A}_\mathcal{G} = \langle \mathfrak{T}(Q), A, \eta, \varepsilon, U \rangle$ by

cf. Exam. 0.2.1, p. 15

$$\forall r \in \mathfrak{T}(Q)\,,\ \forall a \in A \qquad (r, a)\eta = s \qquad \text{with} \qquad \forall p \in Q \quad s_p = (r_p, a)\delta \ ,$$

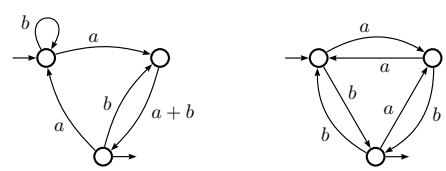
and by

$$U = \{r \in \mathfrak{T}(Q) \mid r_i \in T\} \ ;$$

that is, the set U of terminal states of $\mathcal{A}_\mathcal{G}$ is the set of maps which send i, the initial state of $\mathcal{A}$, to T. In fact, we are only interested in the accessible subset of $\mathcal{A}_\mathcal{G}$, which we also write $\mathcal{A}_\mathcal{G}$.

(i) Show that the two automata $\mathcal{S}$ and $\mathcal{D}$ shown in Figure 8.1 are *not* left regular. Give the graph representation of $\mathcal{S}_\mathcal{G}$ and of $\mathcal{D}_\mathcal{G}$.

(ii) Verify that $\mathcal{A}_\mathcal{G}$ is a complete deterministic left regular automaton and is equivalent to $\mathcal{A}$.



(a) The automaton $\mathcal{S}$ (b) The automaton $\mathcal{D}$

Fig. 8.1: Two non-left-regular automata

(c) *Second method for the computation of the roots of a language.* Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ be a deterministic automaton which recognises K. Write V_2 for the subset of Q defined by

$$V_2 = \{p \in Q \mid \exists f \in A^* \quad i \cdot f = p \quad \text{and} \quad i \cdot f\,f \in T\} \ ,$$

and let $\mathcal{A}_{\mathcal{R}2} = \langle Q, A, \delta, i, V_2 \rangle$ be the automaton obtained from $\mathcal{A}$ by taking V_2 as the set of terminal states.

[66] The notation $\mathcal{R}_2$ is obviously intended to remind one of the square root operation.

(i) Show that if $\mathcal{A}$ is left regular, we have

$$L(\mathcal{A}_{\mathcal{R}2}) = \mathcal{R}_2\,(L(\mathcal{A})) \quad . \tag{8.3}$$

(ii) Show that the set V_2 is effectively computable.

(iii) Taking $\mathcal{B} = \mathcal{S}_{\mathcal{G}}$, give the set of terminal states of $\mathcal{B}_{\mathcal{R}2}$. Verify, using the automaton $\mathcal{S}$ from Fig. 8.1 for example, that (8.3) does not necessarily hold for a deterministic automaton.

(iv) *Generalisation.* Give the effective construction for all integers $k \geqslant 3$ of an automaton $\mathcal{A}_{\mathcal{R}k}$ that recognises $\mathcal{R}_k(K)$.

• 8.2 **Iterating factors of a recognisable language.** In this problem, which follows on from the previous one, we want to prove:

Proposition 8.1 *The set of iterating factors of a recognisable language is recognisable.*

Recall that, for every language K over A^*, a word v is an *iterating factor* of K if there exist two words u and w such that $u\,v^*\,w \subseteq K$. We write $\mathcal{I}(K)$ for the set of iterating factors of K. In the following, $\mathcal{A} = \langle\, Q, A, E, I, T \,\rangle$ is a finite automaton and $K = L(\mathcal{A})$. *cf.* § *1.3.3, p. 70*

(a) For each state p in Q, let $\mathcal{A}_p = \langle\, Q, A, \delta, \{p\}, \{p\} \,\rangle$ be the automaton obtained by replacing in $\mathcal{A}$ both the set I of initial states and the set T of terminal states by p.

Show that, for every useful state p, $L(\mathcal{A}_p)$ is included in $\mathcal{I}(L(\mathcal{A}))$.

(b) Show that, for every iterating factor v of $L(\mathcal{A})$, there exists an integer k and a state p in Q such that v^k belongs to $L(\mathcal{A}_p)$.

(c) Deduce (it is not immediate) that $\mathcal{I}(K)$ is recognised by a finite automaton effectively computable from A.

• 8.3 **Slices of rationals.** For each word g in A^*, the *first half* of g, written $\mathsf{FH}(g)$, is the prefix f of g whose length is half that of g (which assumes that g is of even length). By additivity, the first half of a language L, written $\mathsf{FH}(L)$, is the set of first halves of words of L:

$$\mathsf{FH}(L) = \{f \in A^* \mid \exists h \in A^* \qquad |h| = |f| \quad f\,h \in L\} \ .$$

In the same way, we define the first and second *thirds* of L, written $\mathsf{FT}(L)$ and $\mathsf{ST}(L)$:

$$\mathsf{FT}(L) = \{f \in A^* \mid \exists h \in A^* \qquad |h| = 2\,|f| \quad f\,h \in L\} \ ,$$
$$\mathsf{ST}(L) = \{g \in A^* \mid \exists f, h \in A^* \qquad |f| = |g| = |h| \quad f\,g\,h \in L\} \ .$$

(a) Show that if L is recognisable, $\mathsf{FH}(L)$ is too, effectively; that is, one can compute an automaton which recognises $\mathsf{FH}(L)$ from one which recognises L.

(b) Do the same for $\mathsf{FT}(L)$. (c) Do the same for $\mathsf{ST}(L)$.

[Hint: given a deterministic automaton $\mathcal{A} = \langle\, Q, A, \delta, i, T \,\rangle$ which recognises L, it is helpful to consider the automaton $\langle\, \mathfrak{P}(Q), A, \beta, \{T\}, U \,\rangle$ whose transition function is defined by

$$(p, a)\beta = \{q \in Q \mid \exists x \in A \quad (q, x)\delta = p\} \ ,$$

and then extended additively.]

• 8.4 **Thin languages.** We make the following definitions. (i) A *ray* is a set of words of the form $u\,v^*\,w$, where u, v and w are words in A^*; if $v = 1_{A^*}$, the ray is reduced to a single word.

(ii) A language is *thin* if it is a finite union of rays.

(iii) A language L is *of bounded growth* if the number of words in L of length n is bounded by a constant that is independent of n. We will now show:

Proposition 8.2 *A language is thin if and only if it is recognisable and of bounded growth.*

(a) Show that if a language L is thin, it is of bounded growth.

(b) In the following, L is a language recognised by an automaton $\mathcal{A}$ which we assume to be deterministic and trim.

Let $\mathcal{C}$ be a strongly connected component of $\mathcal{A}$. Show that if $\mathcal{C}$ is not reduced to a single loop, L is not of bounded growth.

(c) Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two strongly connected components of $\mathcal{A}$. Show that if there exists a path in $\mathcal{A}$ that leads from a state of $\mathcal{C}_1$ to a state of $\mathcal{C}_2$, L is not of bounded growth.

(d) Deduce that if L is a language of bounded growth, it is thin.

• 8.5 **Number of words of a given length in a recognisable language.**

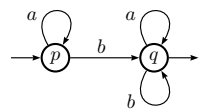
In this problem we will give a method for computing the number of words of length n in a recognisable language.

(a) *Example 1.* Characterise the language recognised by the automaton $\mathcal{B}$ of Figure 8.2 (a).

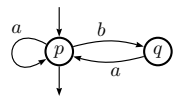
Deduce the number of words of length n recognised by $\mathcal{B}$, for every integer n.

(b) *Example 2.* Characterise the language recognised by the automaton $\mathcal{F}$ of Figure 8.2 (b).

For every integer n, we write $f(n)$ for the number of words of length n recognised by $\mathcal{F}$. Express $f(n+2)$ as a function of $f(n+1)$ and $f(n)$. Compute $f(8)$. Express $f(n)$ as a function of n.



(a) The automaton $\mathcal{B}$



(b) The automaton $\mathcal{F}$

Fig. 8.2: Two automata

(c) *Generalisation.* Let L be the language recognise by a deterministic automaton $\mathcal{A} = \langle Q, A, E, i, T \rangle$ and as usual, write $L_p = \{f \in A^* \mid \exists t \in T \quad p \xrightarrow{f} t\}$ for all $p \in Q$. Write $l(n)$ for the number of words in L of length n and $l_p(n)$ for the number of words in L_p of length n, for all $n \in \mathbb{N}$. [In particular, $L = L_i$ and $l(n) = l_i(n)$.]

(i) Show that, for all positive n and for all $p \in Q$, $l_p(n)$ is a linear combination with non-negative integer coefficients of the $l_q(n-1)$, for q in Q.

(ii) Deduce that the function $l(n)$ satisfies a linear recurrence relation of the form

$$\forall n \in \mathbb{N} \qquad l(n+k) = z_1\, l(n+k-1) + z_2\, l(n+k-2) + \cdots + z_k\, l(n) \ ,$$

where k is the number of states of $\mathcal{A}$ and the z_j are integers. [Hint: it could be helpful to write the result of question (i) above in vector form.]

(iii) Explain how to compute the initial conditions; that is, $l(0)$, $l(1)$, ..., $l(k-1)$.

(d) *Application.* Apply the method of the previous question to the computation of the number of words of length n recognised by the automaton $\mathcal{B}$ of question (a).

SOLUTIONS TO THE EXERCISES

There is no royal road to science,
and only those who do not fear
the fatiguing ascent of its steep paths
may attain its luminous summits.
Karl Marx

1. What is an 'automaton'?

1.1 *Give the matrix form of the automaton $\mathcal{P}_2$.*

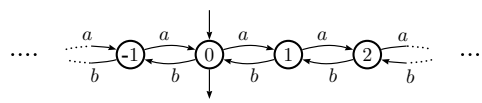
$$\mathcal{P}_2 = \left\langle \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} a & b & 0 \\ b & 0 & a \\ 0 & a & b \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\rangle$$

1.2 *How many automata over $\{a, b\}$ have three states?*

Solution: Write $\mathcal{A} = \langle \{1,2,3\}, \{a,b\}, E, I, T \rangle$. Each of the nine elements of E can take four possible values, and there are 2^3 possibilities for I and for T. This gives a total of $2^3 \times 2^3 \times 4^9 = 2^{24} = 4\ 194\ 304$ automata with three states. This count includes automata that are equal except for the names of their states. However, dividing it by $3! = 6$ (which gives about 700 000) does not take into account that some automata are invariant under permutation of their states. The precise enumeration of distinct automata is hard.

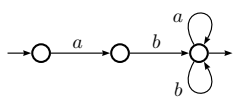
1.3 *Show that the language recognised by $L(\mathcal{Z}_1)$ is*

$$L(\mathcal{Z}_1) = \{f \in \{a,b\}^* \mid |f|_a = |f|_b\}\,.$$

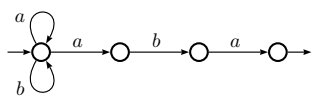


Solution: By induction on $|f|$, we verify that for each f in $\{a,b\}^*$ there exists one, and only one, computation with source 0 and label f, and that this computation's destination is the state $z = |f|_a - |f|_b$. The successful computations there are therefore labelled with words f such that $|f|_a - |f|_b = 0$.

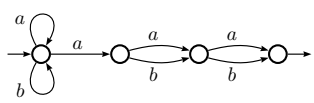
1.4 *Describe the languages recognised by the following automata.*

(a) 

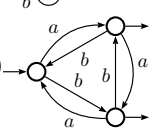
recognises words which start with $a\,b$.

(b) 

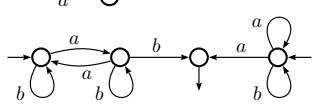
recognises words that end in $a\,b\,a$.

(c) 

recognises words whose antepenultimate[67] letter is a.

(d) 

recognises letters whose number of a's and b's is different modulo 3.

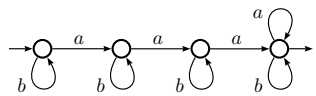
(e) 

recognises words that either end in a or end in b and contain an odd number of a's.

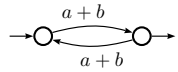
[67] Last but two.

1.5 *Construct finite automata that accept that following languages ($A = \{a, b\}$).*

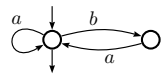
(a) $L_3 = \{f \in A^* \mid |f|_a \geqslant 3\}$ is accepted by the automaton in the adjacent figure.



(b) $L_4 = \{f \in A^* \mid |f|_a \not\equiv |f|_b \bmod 2\}$ is the set of words of odd length, hence the automaton in the adjacent figure.



(c) L_5, the set of words in A^* such that each occurrence of b is followed by an occurrence of a, is accepted by the automaton in the adjacent figure.



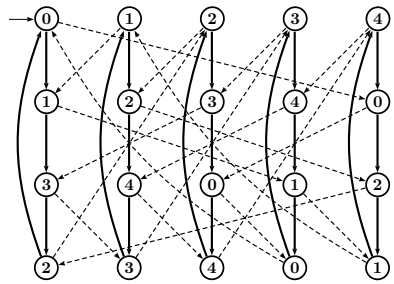
1.6 *Give an automaton that computes the remainder after division by 5 of numbers written in binary, reading the numbers (a) from right to left; (b) from left to right.*

Solution: (a) We have

$$2^0 \equiv 1\ [5]\ ,\quad 2^1 \equiv 2\ [5]\ ,\quad 2^2 \equiv 4\ [5]\ ,$$
$$2^3 \equiv 3\ [5]\ ,\quad 2^4 \equiv 1\ [5]\quad \ldots$$

hence the table

...	4	3	2	1	0
...	1	3	4	2	1

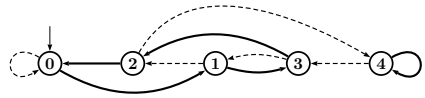


and the automaton in the adjacent figure. We have not included the second component of the states, which corresponds to the row number. In order not to over-burden the picture, the labels of each arrow are not shown either; instead, a bold arrow is used for a label 1 and a dashed arrow for label 0.

(b) We calculate thus that

if the remainder of the number written	f	is:	0 ,	1 ,	2 ,	3 ,	or 4 ,
then the remainder of the number written	$f\,0$	is:	0 ,	2 ,	4 ,	1 ,	or 3 ,
and the remainder of the number written	$f\,1$	is:	1 ,	3 ,	0 ,	2 ,	or 4 ,

hence the automaton in the adjacent figure (with the same conventions as above).



1.9 *Finish the proof that*

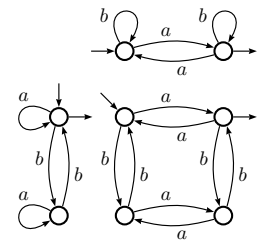
(a) $L(\mathcal{A}' \cup \mathcal{A}'') = L(\mathcal{A}') \cup L(\mathcal{A}'')$; (b) $L(\mathcal{A}' \times \mathcal{A}'') = L(\mathcal{A}') \cap L(\mathcal{A}'')$.

Solution: (a) It is not enough to say that 'a transition is in $\mathcal{A}$ if and only if it is either a transition of $\mathcal{A}'$, or a transition of $\mathcal{A}''$'; it must be added that, as Q' and Q'' are disjoint, a transition from $\mathcal{A}'$ and a transition from $\mathcal{A}''$ *cannot be adjacent* (the destination of one cannot be the source of the other) and hence a computation in $\mathcal{A}$ has all its transitions either in $\mathcal{A}'$, or in $\mathcal{A}''$. Then, and then only, it is legitimate to write that 'it follows that a (successful) computation of $\mathcal{A}$ is a (successful) computation of either $\mathcal{A}'$, or $\mathcal{A}''$'.

(b) On p. 61 the proof of the converse is missing: that is, that if f is both the label of a computation of $\mathcal{A}'$ and of a computation of $\mathcal{A}''$, then it is indeed the label of a computation of $\mathcal{C} = \mathcal{A}' \times \mathcal{A}''$, and that this last is successful if and only if the two first are. All of this is fairly obvious, and easily verified from the definition of $\mathcal{C}$ by induction on the length of f.

1.10 *Construct an automaton that recognises the set of words which contain an odd number of a's and an even number of b's.*

Look again at the automaton in Example 1.5 which recognises the set L_2 of words that have an odd number of a's. By swapping the a's and b's, and changing the final state, we obtain an automaton that recognises the set of words which have an even number of b's. The product of these two automata, shown in the adjacent figure, recognises the set of words which have both an odd number of a's and an even number of b's.



1.11 *Let L be a language of A^*. Which languages are defined by $[A^*]^{-1}L$, $L\,[A^*]^{-1}$ and $[A^*]^{-1}L\,[A^*]^{-1}$?*

Solution: $[A^*]^{-1}L = \mathsf{Suf}(L)$, $L\,[A^*]^{-1} = \mathsf{Pre}(L)$, and $[A^*]^{-1}L\,[A^*]^{-1} = \mathsf{Fac}(L)$. *cf. Prop. 1.10, p. 68*

1.12 *Let $K_1 = \{a^n b^m \mid n, m \in \mathbb{N}\}$. Describe the languages $[a^*]^{-1}K_1$, $K_1\,[b^*]^{-1}$ and $[a^*]^{-1}K_1\,[b^*]^{-1}$. Do the same for $P_1 = Z_1 \cap K_1 = \{a^n b^n \mid n \in \mathbb{N}\}$.* *cf. Exam. 1.4, p. 57*

Solution: We have $[a^*]^{-1}K_1 = K_1\,[b^*]^{-1} = [a^*]^{-1}K_1\,[b^*]^{-1} = K_1$. On the other hand,

$$[a^*]^{-1}P_1 = \{a^n b^m \mid n, m \in \mathbb{N} \quad n \leqslant m\} \ , \quad P_1\,[b^*]^{-1} = \{a^n b^m \mid n, m \in \mathbb{N} \quad n \geqslant m\} \ ,$$
$$\text{and} \qquad [a^*]^{-1}P_1\,[b^*]^{-1} = K_1 \ .$$

1.13 *Prove:* $\forall f, g, h \in A^* \qquad (f\,g)^{-1}h = g^{-1}(f^{-1}h)$.

Solution: The associativity of product $(f\,g)\,k = f\,(g\,k)$ implies the series of equivalences: *cf. Propty 3.7, p. 112*

$$(f\,g)^{-1}h = k \ \Leftrightarrow \ h = f\,g\,k \ \Leftrightarrow \ f^{-1}h = g\,k \ \Leftrightarrow \ g^{-1}(f^{-1}h) = k$$

1.14 *Prove:* $\forall f, g, h \in A^* \qquad (f^{-1}g)^{-1}h = g^{-1}(f\,h) \quad$ *if* $\quad |f| \leqslant |g|$.

Solution: If $f^{-1}g \neq \emptyset$, then $g = f\,k$ and $g^{-1}(f\,h) = k^{-1}(f^{-1}(f\,h)) = k^{-1}h = (f^{-1}g)^{-1}h$. If $f^{-1}g = \emptyset$ and $|f| \leqslant |g|$, then $f = k\,a\,u$ and $g = k\,b\,v$ where $a \neq b$, whence $g^{-1}(f\,h) = \emptyset$.

The condition on the lengths is necessary, since if g is a proper prefix of f then $f = g\,k$, and we have $f^{-1}g = \emptyset = (f^{-1}g)^{-1}h$ and $g^{-1}(f\,h) = k\,h$.

1.15 *Show that a non-empty language L over A^* is prefix if and only if one of the following conditions is satisfied:*[68]

(a) $f^{-1}L = 1_{A^*}$ *for all f in L;* (b) $L^{-1}L = 1_{A^*}$; (c) $L \cap L\,A^+ = \emptyset$.

Solution: (c) is almost the definition of a prefix set. We deduce both (a) and (b) from the following equivalences

$$\forall f \in L \quad f^{-1}L = 1_{A^*} \quad \Leftrightarrow \quad \forall f, g \in L \quad f \text{ is not a prefix of } g \quad \Leftrightarrow \quad L \text{ is prefix.}$$

1.16 ***Shuffle product.*** (a) *Give some (simple) examples of recognisable languages which are shuffles of two languages taken from previous examples.*

(b) *Verify that shuffle makes $\mathfrak{P}(A^*)$ a commutative monoid with a zero.*

(c) *Prove that the shuffle of two recognise languages is recognisable (Prop. 1.7).*

Solution: (a) The simplest example is without doubt $\{a, b\}^* = a^* \between b^*$. Another example, the language of words which have both an odd number of a's and an even number of b's is the shuffle of the set of words of the form a^{2n+1} with the set of those of the form b^{2m}. *cf. Exer. 1.10 above*

[68] Borrowed from [80, Prop. IV.2.1, and Exer. IV.2.1 and IV.2.2].

(c) Let $\mathcal{A}' = \langle Q', A, E', I', T' \rangle$ and $\mathcal{A}'' = \langle Q'', A, E'', I'', T'' \rangle$ be two automata over A. The *shuffle* of $\mathcal{A}'$ and $\mathcal{A}''$ is the automaton defined by

$$\mathcal{A}' \between \mathcal{A}'' = \langle Q' \times Q'', A, G, I' \times I'', T' \times T'' \rangle \ ,$$

where the set of transitions G is given by

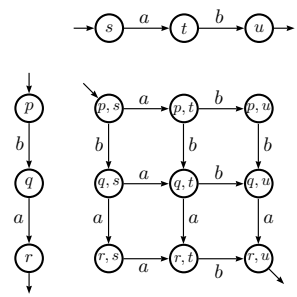
$$G = \{((p', p''), a, (q', p'')) \mid (p', a, q') \in E' \quad \text{and} \quad p'' \in Q''\} \cup \\ \{((p', p''), a, (p', q'')) \mid p' \in Q' \quad \text{and} \quad (p'', a, q'') \in E''\} \ .$$

Thus, a transition is in $\mathcal{A}' \between \mathcal{A}''$ if and only if either the projection on the first component of states is a transition of $\mathcal{A}'$ and the states have the same second component, or the projection on the second component of states is a transitions of $\mathcal{A}''$ and the states have the same first component.

It follows that in a (successful) computation c of $\mathcal{A}' \between \mathcal{A}''$ we can find two disjoint sub-computations whose union covers c such that the first projects on its first component to a (successful) computation of $\mathcal{A}'$ and the second on its second component to a (successful) computation of $\mathcal{A}''$. Hence

$$L(\mathcal{A}' \between \mathcal{A}'') = L(\mathcal{A}') \between L(\mathcal{A}'') \ .$$

As $\mathcal{A}' \between \mathcal{A}''$ is finite if $\mathcal{A}'$ *and* $\mathcal{A}''$ are, Proposition 1.7 is proved. The adjacent figure shows the construction for $a\,b \between b\,a$.



1.17 ***Infiltration product.***

(a) *Give a definition of shuffle using the notation of Definition 1.8.*

(b) *Verify that infiltration makes $\mathfrak{P}(A^*)$ a commutative monoid with a zero.*

(c) *Prove that the infiltration of two recognisable languages is recognisable (Prop. 1.8).*

Solution [partial]: (a) If $f \uparrow g = \{h \in A^* \mid \exists I, J \subseteq [|h|],\ I \cup J = [|h|],\ h_I = f \ \text{ and } \ h_J = g\}$, then we can write

$$f \between g = \{h \in A^* \mid \exists I, J \subseteq [|h|] \quad I \cup J = [|h|] \ , \quad I \cap J = \emptyset \ , \quad h_I = f \quad \text{and} \quad h_J = g\}$$

(c) Using the notation from the last exercise, the *infiltration* of $\mathcal{A}'$ and $\mathcal{A}''$ is the automaton defined by

$$\mathcal{A}' \uparrow \mathcal{A}'' = \langle Q' \times Q'', A, H, I' \times I'', T' \times T'' \rangle \ ,$$

where the set H of transitions is described by

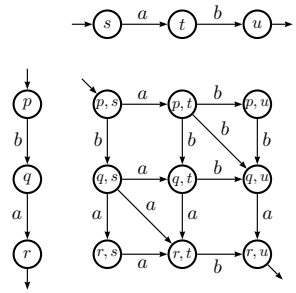
$$H = \{((p', p''), a, (q', p'')) \mid (p', a, q') \in E' \quad \text{and} \quad p'' \in Q''\} \cup \\ \{((p', p''), a, (p', q'')) \mid p' \in Q' \quad \text{and} \quad (p'', a, q'') \in E''\} \cup \\ \{((p', p''), a, (q', q'')) \mid (p', a, q') \in E' \quad \text{and} \quad (p'', a, q'') \in E''\} \ .$$

In other words, the transitions of $\mathcal{A}' \uparrow \mathcal{A}''$ are the *union* of the transitions of $\mathcal{A}' \between \mathcal{A}''$ and $\mathcal{A}' \times \mathcal{A}''$. A transition is therefore in $\mathcal{A}' \uparrow \mathcal{A}''$ if and only if either the projection on the first component of states is a transition of $\mathcal{A}'$ and the states have the same second component, or the projection on the second component of states is a transition of $\mathcal{A}''$ and the states have the same first component, or the projection on the first component of states is a transition of $\mathcal{A}'$ *and* the projection on the second computation of states is a transition of $\mathcal{A}''$ *and hence the two projections have the same label.*

It follows that we can find in a (successful) computation c of $\mathcal{A}' \uparrow \mathcal{A}''$ two sub-computations (not necessarily disjoint) whose union covers c such that the first projects on its first component to a (successful) computation of $\mathcal{A}'$ and the second on its second component to a (successful) computation of $\mathcal{A}''$. Hence

$$L(\mathcal{A}' \uparrow \mathcal{A}'') = L(\mathcal{A}') \uparrow L(\mathcal{A}'') \ .$$

As $\mathcal{A}' \uparrow \mathcal{A}''$ is finite if $\mathcal{A}'$ *and* $\mathcal{A}''$ are, Proposition 1.8 is proved. The adjacent figure shows the construction for $a\,b \uparrow b\,a$.



1.18 *Show that we can decide whether $L(\mathcal{A})$ is empty, finite or infinite, and compute $\mathcal{A}_{\mathsf{m}}$ in complexity $O(k+m)$, where k is the number of initial states and m is the number of transitions of $\mathcal{A}$ (Proposition 1.11).*

Solution : The corresponding algorithm traverses the underlying graph of $\mathcal{A}$. To decide whether $L(\mathcal{A})$ is empty, or to compute $\mathcal{A}_{\mathsf{m}}$, any graph traversal algorithm will do; to decide whether $L(\mathcal{A})$ is infinite: that is if $\mathcal{A}$ contains a circuit, the standard method consists of using and augmenting a depth-first search.[69]

To evaluate the complexity of these algorithms, the data structures involved must be made explicit, as must the computations themselves. To this end, we use a sort of pseudo-code inspired by the language Pascal, which should be comprehensible to every reader who has studied the rudiments of any programming language.

```
procedure traverse ( SUCC, MARK )
{ SUCC is the array of successor lists }
{ MARK is an auxiliary array, initialised to 0 }
  procedure dfsearch1 ( x )
  MARK[x] := 1 ;
  for all y in SUCC[x]
   do if MARK[y] = 0 then dfsearch1 ( y ) od
  return
for all p in SUCC[0]
 do if MARK[p] = 0 then dfsearch1 ( p ) od
return ( MARK )
```

(a) simple traversal

```
procedure traverse+loops ( SUCC, MARK, LOOP )
{ SUCC is the array of successor lists }
{ MARK, LIVE, LOOP are initialised to 0 }
  procedure dfsearch2 ( x )
  COUNT := COUNT + 1 ;
  MARK[x] := COUNT ; LIVE[x] := 1 ;
  for all y in SUCC[x]
   do if MARK[y] = 0 then
        then dfsearch2 ( y )
        else  do if MARK[y] <= MARK[x] then
                  if LIVE[y] = 1 then LOOP[y] := 1
              od
   od
  LIVE[x] := 0 ;
  return
COUNT := 0 ;
for all p in SUCC[0]
 do if MARK[p] = 0 then dfsearch2 ( p ) od
return ( MARK, LOOP )
```

(b) traversal and finding circuits

Figure SE.1.1: Traversal of an automaton

The graph underlying $\mathcal{A}$ is described by the array $SUCC$ of size $n+2$, where $n = \|Q\|$, each element $SUCC[p]$ of which is the list of successors of p. By convention, $SUCC[0]$ is the list of initial states and $SUCC[p]$ contains $n+1$ if and only if p is a final state.

The procedure **traverse** (*cf.* Figure SE.1.1 (a)) uses an auxiliary array $MARK$ of size $n+2$, initialised to 0, and calls the recursive procedure **dfsearch1** defined inside **traverse**, which is assumed to have access to all the variables and parameters of its parent. When the `for`

[69] This algorithm is described and studied in all algorithmic and graph-theoretic texts, such as [5, Sec. 6.5] and [23, Ch. 4].

statement in **traverse** terminates, $MARK[p] = 1$ if and only if p is accessible, and $L(\mathcal{A})$ is non-empty if and only if $MARK[n+1] = 1$. Each initial state is examined in the `for` statement in **traverse** and then the successor lists of the accessible states are each examined once in the `for` statement in **dfsearch1**: the complexity of the program is $O(k+m)$.

The computation of useful states, and hence of $\mathcal{A}_{\mathsf{m}}$, is performed by applying the same procedure **traverse** to $\mathcal{A}^{\mathsf{t}}$ in which the initial states are taken to be the *accessible final states* of $\mathcal{A}$ (whose number is clearly less than $k+m$).

The procedure **traverse+loops** (Figure SE.1.1 (b)), or rather the recursive procedure **dfsearch2** which it calls, is more subtle and uses a classic algorithmic result on graphs (see the references given in the notes). At the end of **traverse+loops**, every elementary circuit of $\mathcal{A}$ contains a state p such that $LOOP[p] = 1$. The complexity of **traverse+loops** is obviously the same as that of **traverse**.

1.19 ***Roy–Warshall algorithm.*** *An algorithm for calculating the* transitive closure *of a graph* $\mathcal{G} = \langle Q, G \rangle$.

We set $Q = [n]$ *and for each integer* k, $0 \leqslant k \leqslant n$, *we write* $\mathcal{F}^{(k)} = \langle Q, F^{(k)} \rangle$ *for the graph such that there exists an* edge *from* i *to* j *in* $\mathcal{F}^{(k)}$ *if and only if there exists a* path *in* $\mathcal{G}$ *from* i *to* j all of whose intermediate vertices *are less than or equal to* k. *Clearly,* $\mathcal{F}^{(0)} = \mathcal{G}$ *and* $\mathcal{F}^{(n)}$ *is the transitive closure of* $\mathcal{G}$.

Give an algorithm that computes $\mathcal{F}^{(n)}$ *with complexity* $O(n^3)$.

Solution: If there exists in $\mathcal{G}$ a path from i to j all of whose intermediate vertices are less than or equal to k, then there exists an *elementary path* with the same property. If no vertex in the path is equal to k, then all the vertices are less than or equal to $k-1$; otherwise this path is the concatenation of a path from i to k and a path from k to j, all of whose intermediate vertices are less than or equal to $k-1$. That is, writing $F^{(k)}$ for the incidence matrix $\mathcal{F}^{(k)}$,

$$F^{(k)}_{i,j} = F^{(k-1)}_{i,j} \vee [F^{(k-1)}_{i,k} \wedge F^{(k-1)}_{k,j}] \qquad \text{(SE.1.1)}$$

is the computation performed by the procedure `Roy-Warshall-1` whose complexity is clearly $O(n^3)$ (Figure SE.1.2 (a)).

```
procedure Roy--Warshall-1 ( G, F )
{ G is the incidence matrix of 𝒢 }
{ F is that of the transitive closure of 𝒢 }
{ H is an auxiliary matrix for the computation }
F := G ;
for i from 1 to n do F[i,i] := 1 od
H := F ;
for k from 1 to n
 do for i from 1 to n
    do for j from 1 to n
       H[i,j] := F[i,j] or (F[i,k] and F[k,j])   {1}
    od
    F := H
 od
return ( F )
```

(a) direct implementation

```
procedure Roy--Warshall-2 ( G, F )
{ G is the incidence matrix of 𝒢 }
{ F is that of the transitive closure of 𝒢 }
F := G ;
for i from 1 to n do F[i,i] := 1 od
for k from 1 to n
 do for i from 1 to n
    do for j from 1 to n
       F[i,j] := F[i,j] or (F[i,k] and F[k,j])   {1}
    od
 od
return ( F )
```

(b) efficient implementation

Figure SE.1.2: Roy–Warshall algorithm

The procedure `Roy-Warshall-2` (Figure SE.1.2 (b)) is a classic variant. It is more efficient both in space, since the auxiliary matrix H is no longer required, and in time, since it avoids making $n+1$ copies of an $n \times n$ matrix.

The difference between the operation of the two procedures comes from the execution of statement $\{\mathbf{1}\}$: in the second procedure the value of $\mathcal{F}^{(k)}_{i,j}$, for $i < k < j$ for example, is obtained by the formula

$$F^{(k)}_{i,j} = F^{(k-1)}_{i,j} \vee [F^{(k)}_{i,k} \wedge F^{(k-1)}_{k,j}] \ , \qquad \text{(SE.1.2)}$$

instead of that given for (SE.1.1). The result, however, in this case as in the others, is identical, since we observe that if (i,k) (resp. (k,j)) is an edge in $\mathcal{F}^{(k)}$, it is already in $\mathcal{F}^{(k-1)}$.[70]

1.20 ***Transitive closure revisited.*** *Once again, let $\mathcal{G} = \langle Q, G \rangle$ be a graph. Considering the edges rather than the nodes of $\mathcal{G}$, give an algorithm that computes the transitive closure $\mathcal{F}$ of $\mathcal{G}$ with complexity $O(d\,m')$, where m' is the number of edges in $\mathcal{F}$ and d is the largest number of edges that leaves any node of $\mathcal{G}$ (maximum Out-degree).*

Solution: The graph $\mathcal{G}$ is described both by its incidence matrix G and by the array $SUCC$ of size $n = \|Q\|$, each of whose elements $SUCC[p]$ is the list of successors of p.

The procedure **closure** (Figure SE.1.3) uses the auxiliary variable $EDGE$, a *queue* which contains edges, or pairs (i,j) of vertices of $\mathcal{G}$, and which is initialised with all the edges of $\mathcal{G}$. The matrix F is initialised to G.

```
procedure closure ( SUCC, F )
{ SUCC is the array of lists of successors in G }
{ F is an incidence matrix, initialised to G }
{ EDGE is a queue of edges, initialised with the edges of G }
while NonEmpty(EDGE)
 do (i,j) := Front(EDGE) ;
    for all s ∈ SUCC[j]
     do if F[i,s] = 0 then
           do F[i,s] := 1 ;
              Enter((i,s), EDGE)
           od
     od
 od
return ( F )
```

Figure SE.1.3: Another transitive closure algorithm

Showing that the procedure **closure** indeed calculates the transitive closure of $\mathcal{G}$ is almost direct. We imagine that each edge in the queue has an index: all the edges of $\mathcal{G}$ are assigned the index 1 during initialisation, and in the course of the computation and edge created while considering an edge with index k is given the index $k+1$. We verify then, by induction on k, that the index of an edge (i,j) in $\mathcal{F}$ is exactly the minimum length of a path from i to j in $\mathcal{G}$ and that conversely if such a path from i to j exists in $\mathcal{G}$, the edge (i,j) is added to $EDGE$ with the correct index. When the procedure terminates, F is the incidence matrix of $\mathcal{F}$. Each edge of $\mathcal{F}$ occurs once in $EDGE$. The examination of each edge (i,j) consists of examining all the successors of j, from which we obtain the stated complexity.[71]

1.22 *Show that the following languages over $A = \{a,b\}$ are not recognisable:*

(a) $Z'_2 = \{a^n b^m \mid n,m \in \mathbb{N} \quad n \leqslant m\}$; (b) $Z'_3 = \{a^n b^m \mid n,m \in \mathbb{N} \quad n \neq m\}$;

(c) $Z_1 = \{f \in A^* \mid |f|_a = |f|_b\}$;

(d) *the language of palindromes:* $\mathsf{Pal} = \{f \in A^* \mid f = f^{\mathsf{t}}\}$;

(e) *the language of squares:* $\mathsf{Car} = \{f \mid \exists g \in A^* \quad f = g^2\}$.

Solution: (a) We choose $a^N b^N$, where N is given by Lemma 1.13 and find in b^N an iterating factor which, when it is deleted, makes us leave Z'_2.

(b) Two methods. Direct: consider the word $a^N b^{N+N!}$; find in a^N an iterating factor $v = a^k$ where $k < N$; there exists p such that $p\,k = N!$; by iterating v p times we obtain the word $a^{N+N!} b^{N+N!}$ which makes us leave Z'_3.

[70] The computation of the transitive closure of a graph is also (that is, along with graph traversal) a classic part of the algorithmics of graphs, of which the Roy–Warshall algorithm is only the very beginning; see [245, Sec. 1.4] for a complete overview of this subject.

[71] This algorithm is taken from [25].

Indirect: $P_1 = A^* \setminus [Z_3' \cup A^*baA^*]$, the recognisable languages are closed under complement (but we do not know that yet, see Proposition 3.4) and P_1 is not recognisable.

cf. Rem. 1.10, p. 72

(c) $P_1 = Z_1 \cap K_1$ (where $K_1 = \{a^n b^m \mid n, m \in \mathbb{N}\}$), the recognisable languages are closed under intersection, and P_1 is not recognisable.

cf. Exam. 1.4, p. 57

(d) $\mathsf{Pal} \cap ba^*bba^*b = \{ba^n bba^n b \mid n \in \mathbb{N}\}$ and we proceed as for P_1.

(e) Same as (d).

1.23 (a) *Give an example of a non-recognisable language X such that $\mathsf{Pre}(X)$ and $\mathsf{Suf}(X)$ are recognisable.*

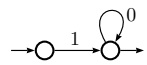
(b) *Give an example of a language X such that $\mathsf{Pre}(X)$ and $\mathsf{Suf}(X)$ are non-recognisable and that $\mathsf{Fac}(X)$ is recognisable.*

Solution: (a) Z_1 is not recognisable and $\mathsf{Pre}(Z_1) = \mathsf{Suf}(Z_1) = A^*$.

(b) $\mathsf{Pre}(P_1) = Z_2'$, $\mathsf{Suf}(P_1) = [Z_2']^{\mathsf{t}}$ and $\mathsf{Fac}(P_1) = K_1$.

1.25 ***k-recognisable sets of numbers.*** *Show that the set of powers of 3, $T = \{3^n \mid n \in \mathbb{N}\}$ is 3-recognisable but not 2-recognisable.*

Solution: The set T_3 of ternary representations of the elements of T, $T_3 = \{10^n \mid n \in \mathbb{N}\}$ is 3-recognisable (see figure opposite).



Suppose that the set T_2 of binary representations of elements of T is 2-recognisable, and let $f_1 = uvw$ be a word in T_2 whose factor v is iterating. Then, for all n, $f_n = uv^n w$ is in T_2 and we can easily verify that

$$\lim_{n\to\infty} \frac{\overline{f_{n+1}}}{\overline{f_n}} = 2^{|v|} \quad ,$$

where $\overline{f}$ is the integer whose binary representation is f. We conclude that we would have to find an infinite series of integers $\{k_i\}_{i\in\mathbb{N}}$ such that

$$\lim_{i\to\infty} 3^{k_i - k_{i-1}} = 2^{|v|} \quad ,$$

which is clearly impossible, so T_2 is not 2-recognisable.[72]

1.26 *What can be said about the union and the product*

(a) *of two trim automata?* (b) *of two unambiguous automata?*

(c) *of two complete automata?*

Solution: (a) The union of two trim automata is obviously trim. The product of two trim automata is not necessarily trim.

(b) Let $\mathcal{A}' = \langle Q', A, E', I', T' \rangle$ and $\mathcal{A}'' = \langle Q'', A, E'', I'', T'' \rangle$ be two unambiguous automata. The automaton $\mathcal{B} = \mathcal{A}' \cup \mathcal{A}''$ satisfies condition (i) of Definition 1.11. It satisfies condition (ii) if and only if $L(\mathcal{A}') \cap L(\mathcal{A}'') = \emptyset$: that is, if the union of the languages recognised is unambiguous.

cf. p. 74

The product of two unambiguous automata is unambiguous.

(c) The union of two complete automata is a complete automaton; the same applies to the product.

[72] This result is a most elementary case of a very general result due to Cobham [62]: *if k and l are two multiplicatively independent integers: that is, there are no integers p and q such that $k^p = l^q$ (alternatively, $\log k / \log l$ is irrational) a set of numbers is both k and l-recognisable if and only if it is ultimately periodic.* This fundamental and superb result has been the object of much work both to simplify its proof and to generalise it; a clear exposition may be found in the article by V. Bruyère, G. Hansel, C. Michaux and R. Villemaire [39] and other references in [159, Ch. 7]. I hope to deal with it in a future volume.

1.28 *Verify that a trim automaton $\mathcal{A}$ is unambiguous if and only if for all f in $L(\mathcal{A})$, there exists a* unique *successful computation whose label is f.*

Solution: If $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is trim there exists, for all p in Q, u_p and v_p in A^*, i_p in I and t_p in T such that $i_p \xrightarrow{u_p} p \xrightarrow{v_p} t_p$ is a computation of $\mathcal{A}$. There exists a pair of states (p, q) such that there exist at least two distinct computations (with source p and destination q) with label g if and only if the word $u_p\, g\, v_q$ in $L(\mathcal{A})$ is the label of two distinct successful computations with the same source and destination. Condition (ii) of Definition 1.11 ensures the equivalence for successful computations with the same label but which do not necessarily have the same source or destination.

1.29 *What is the complexity of the procedure for deciding whether a trim automaton is unambiguous?*

Solution: If $\mathcal{A}$ has n states and m transitions, $\mathcal{A}\times\mathcal{A}$ has at most m^2 transitions and the procedure for computing the trim part of $\mathcal{A}\times\mathcal{A}$ is linear in the number of transitions of the automaton: that is, $O(m^2)$. The comparison with the diagonal of $\mathcal{A}\times\mathcal{A}$, which is $O(n^2)$, does not increase the total complexity.

1.30 *Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be a finite automaton and (p, q) and (r, s) two pairs of states (not necessarily distinct) of $\mathcal{A}$. Show directly (that is, without using the automaton $\mathcal{A}\times\mathcal{A}$ explicitly) that if there exists a word f in A^* which is the label of* two distinct computations *in $\mathcal{A}$, one from p to q and the other from r to s, then there exists at least one such word of length less than $\|Q\|^2 + \|Q\|$.*

Draw appropriate conclusions and comment.

Solution: We proceed by imitating the proof of the Star Lemma, using the Pigeonhole Principle on pairs of states. Let f be a word of minimal length such that there exist two distinct computations from p to q with label f. Suppose that $|f|$ is greater than or equal to $\|Q\|^2+\|Q\|$. Set $m = \|Q\|^2$, and write

$$f = g = a_1 a_2 \cdots a_m a_{m+1} \cdots a_n \ ;$$

then the two distinct computations are

$$\begin{aligned} c' \ &:= \ p = q'_0 \xrightarrow{a_1} q'_1 \xrightarrow{a_2} \cdots \xrightarrow{a_m} q'_m \xrightarrow{a_{m+1}} \cdots \xrightarrow{a_n} q'_n = q \\ c'' \ &:= \ r = q''_0 \xrightarrow{a_1} q''_1 \xrightarrow{a_2} \cdots \xrightarrow{a_m} q''_m \xrightarrow{a_{m+1}} \cdots \xrightarrow{a_n} q''_n = s \end{aligned}$$

As m is equal to $\|Q\|^2$, there exist two *distinct* indices i and j, $0 \leqslant i < j \leqslant m$, such that

$$(q'_i, q''_i) = (q'_j, q''_j) \ .$$

The two computations

$$\begin{aligned} d' \ &:= \ p = q'_0 \xrightarrow{a_1} q'_1 \xrightarrow{a_2} \cdots \xrightarrow{a_i} q'_i = q'_j \xrightarrow{a_{j+1}} q'_{j+1} \xrightarrow{a_{j+2}} \cdots \xrightarrow{a_m} q'_m \xrightarrow{a_{m+1}} \cdots \xrightarrow{a_n} q'_n = q \\ d'' \ &:= \ r = q''_0 \xrightarrow{a_1} q''_1 \xrightarrow{a_2} \cdots \xrightarrow{a_i} q''_i = q''_j \xrightarrow{a_{j+1}} q''_{j+1} \xrightarrow{a_{j+2}} \cdots \xrightarrow{a_m} q''_m \xrightarrow{a_{m+1}} \cdots \xrightarrow{a_n} q''_n = s \end{aligned}$$

share the label $g = a_1 a_2 \cdots a_i a_{j+1} \cdots a_m a_{m+1} \cdots a_n$ which is strictly shorter than f since i and j are distinct. If these two computations are distinct, we have a contradiction with the hypothesis that f is minimal.

However, it could be that the two computations c' and c'' differ only in the portion from q'_i to q'_j on one hand and q''_i and q''_j on the other and that the two comps d' and d'' are equal. Hence, for all h, $0 \leqslant h \leqslant i$ we have $j + 1 \leqslant h \leqslant n$, and $q'_h = q''_h$ – so in particular $p = r$ and $q = s$ – and there can be no more than $\|Q\|$ distinct pairs (q'_h, q''_h). Since $n - m \geqslant \|Q\|$, there exist two *distinct* indices k and l, $m \leqslant k < l \leqslant n$, such that $q'_k = q'_j$ $[= q''_k = q''_l]$. The two computations

$$\begin{aligned} p &= q'_0 \xrightarrow{a_1} q'_1 \xrightarrow{a_2} \cdots \xrightarrow{a_m} q'_m \xrightarrow{a_{m+1}} \cdots \xrightarrow{a_k} q'_k = q'_l \xrightarrow{a_{l+1}} \cdots \xrightarrow{a_n} q'_n = q \\ p &= q''_0 \xrightarrow{a_1} q''_1 \xrightarrow{a_2} \cdots \xrightarrow{a_m} q''_m \xrightarrow{a_{m+1}} \cdots \xrightarrow{a_k} q''_k = q''_l \xrightarrow{a_{l+1}} \cdots \xrightarrow{a_n} q''_n = q \end{aligned}$$

are distinct (since the two initial computations are) and have as their common label

$$g = a_1 a_2 \cdots a_m a_{m+1} \cdots a_l a_{k+1} \cdots a_n \quad ,$$

which is strictly shorter than f since k and l are distinct. We have a new contradiction: the length of f must be less than $\|Q\|^2 + \|Q\|$.

This result also allows us to conclude in favour of the decidability of the ambiguity of a finite automaton, although the complexity of the procedure that we can deduce directly is much greater than that which we deduced from Lemma 1.16.

cf. Rem. 1.11, p. 72

1.31 *Prove that an* accessible automaton $\mathcal{A}$ is deterministic if and only if the accessible subset of the cartesian square $\mathcal{A}\times\mathcal{A}$ is equal to its diagonal $\mathcal{D}$ *(Property 1.7).*

Solution: (i) The square $\mathcal{A}\times\mathcal{A}$ has a single initial state if and only if all the initial states of $\mathcal{A}\times\mathcal{A}$ are in $\mathcal{D}$ and hence if and only if $\mathcal{A}$ has a single initial state.

(ii) By induction on the length of computations used to reach an accessible state of $\mathcal{A}\times\mathcal{A}$. Let (p,p) be an accessible state of $\mathcal{A}\times\mathcal{A}$ (in $\mathcal{D}$). If $\mathcal{A}$ is deterministic, then for each letter a there exists at most one transition (p,a,q) in $\mathcal{A}$ and thus at most one transition $((p,p),a,(q,q))$ in $\mathcal{A}\times\mathcal{A}$ and the new state reached is in $\mathcal{D}$. If $\mathcal{A}$ is not deterministic, there exists at least one accessible state p in $\mathcal{A}$ and hence an accessible state (p,p) in $\mathcal{A}\times\mathcal{A}$, such that there exists a letter a such that there exist *two* transitions, (p,a,q) and (p,a,r), in $\mathcal{A}$. The automaton $\mathcal{A}\times\mathcal{A}$ contains the transition $((p,p),a,(q,r))$, and the state (q,r), which is not in $\mathcal{D}$, is accessible.

1.32 *Show that* the product of two deterministic automata is a deterministic automaton *(Property 1.10).*

Solution: Let $\mathcal{A}' = \langle Q', A, E', i', T' \rangle$ and $\mathcal{A}'' = \langle Q'', A, E'', i'', T'' \rangle$ be two deterministic automata over A. For each state (p', p'') in $Q'\times Q''$ and each letter a in A, there is at most a single transition (p', a, q') in E' and a single transition (p'', a, q'') in E''. There is therefore at most a single transition $((p', p''), a, (q', q''))$ in $\mathcal{A} = \mathcal{A}'\times\mathcal{A}''$: $\mathcal{A}$, whose unique initial state, $i'\times i''$, is deterministic.

1.33 *Let $\mathcal{A}'$ and $\mathcal{A}''$ be two deterministic automata. Construct a deterministic automaton that recognises $L(\mathcal{A}') \cup L(\mathcal{A}'')$.*

Solution: Assume that $\mathcal{A}'$ and $\mathcal{A}''$ are complete (if they are not, complete them, which is a determinism-preserving operation). Then form the product $\mathcal{A}'\times\mathcal{A}''$ but instead of taking $T'\times T''$ as the set of final states take $T'\times Q'' \cup Q'\times T''$, obtaining an automaton which fulfils the requirement, and which is deterministic according to the last exercise.

1.37 *Give a direct proof that the forward and backward closures of an automaton with spontaneous transitions are equivalent (independent from the proof that both are equivalent to the original automaton).*

Solution: The backward closure of $\mathcal{A}$ is the transpose of the forward closure of the transpose of $\mathcal{A}$.

2. Rational languages

2.1 *Verify that* $\forall X \subseteq A^* \quad X^* = (X \setminus 1_{A^*})^*$ *(Property 2.1).*

Solution: We have $(X \setminus 1_{A^*})^* \subseteq X^*$ since, $\forall X, Y \in A^* \quad Y \subseteq X \Rightarrow Y^* \subseteq X^*$.

For the converse, note first that the inverse inclusion holds if $X = \emptyset$ or $X = \{1_{A^*}\}$. Assume that X is not reduced to the empty word and let g be in X^*: that is, $g = f_1 f_2 \cdots f_n$ with all the f_i in X. If all the f_i are equal to 1_{A^*}, $g = 1_{A^*} \in (X \setminus 1_{A^*})^*$. If not, we can assume that none of the f_i is equal to 1_{A^*} and $g \in (X \setminus 1_{A^*})^*$.

2.2 (a) *Show*[73] *that a non-empty language L over A^* is prefix if and only for all disjoint languages H and K over A^*, $L\,H$ and $L\,K$ are disjoint.*

(b) *Show that if K is a prefix language of A^*, then (Property 2.2)*

$$\forall f, f' \in K\,,\ \forall g, g' \in A^* \qquad f\,g \leqslant f'\,g' \quad\Longrightarrow\quad f = f' \qquad \text{and} \qquad g \leqslant g'\ .$$

(c) *Show that a non-empty language L over A^* is prefix if and only if $K\,L$ is prefix for all non-empty prefix languages K over A^*.*

Solution: (a) Let f_1 and f_2 be in L, h in H and k in K such that $f_1 h = f_2 k$. Since h cannot be equal to k, this equality implies that of f_1 and f_2 one is a prefix of the other and L is not prefix.

Conversely, if L is not prefix it contains two words f and $g = f\,h$. Take $K = 1_{A^*}$ and $H = \{h\}$, then $L\,H \cap L\,K$ contains at least g.

(b) If $f\,g$ is a prefix of $f'\,g'$, then f is a prefix of f' or f' is a prefix of f. Since K is prefix, in both cases $f = f'$. Thus $f\,g \leqslant f'\,g'$ implies $g \leqslant g'$.

(c) It follows from (b) that if L is prefix, $K\,L$ is prefix. Conversely, L is prefix, if $K\,L$ is prefix for some non-empty K.

2.3 *Show that the product of two prefix matrices is prefix (Property 2.3).*

see p. 85 for the definition of a prefix matrix.

Solution: Let m be a $P{\times}Q$ matrix with entries in $\mathfrak{P}(A^*)$. For each p of P, write $m_{p\bullet}$ for the union of entries of row p of m:

$$m_{p\bullet} = \bigcup_{q\in Q} m_{p,q}\ .$$

Let m and m' be two prefix matrices of dimension $P{\times}Q$ and $Q{\times}R$ respectively. Consider any row p of $m\,m'$ and suppose that there exist r and s in R such that

$$\exists u \in (m\,m')_{p,r}\,,\ \exists v \in (m\,m')_{p,s} \qquad u \leqslant v\ .$$

By definition of the matrix product, there exist j and k in Q such that

$$\exists f \in m_{p,j}\,,\ \exists f' \in m'_{j,r} \quad u = f\,f' \qquad \text{and} \qquad \exists g \in m_{p,k}\,,\ \exists g' \in m'_{k,r} \quad v = g\,g'\ .$$

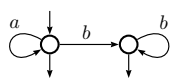
Since f and g belong to the prefix set $m_{p\bullet}$, we have (Property 2.2)

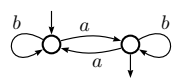
$$f\,g \leqslant f'\,g' \quad\Longrightarrow\quad f = f' \qquad \text{and} \qquad g \leqslant g'\ .$$

Since the $\{m_{p,q}\}_{q\in Q}$ are pairwise disjoint, $f = g$ implies $j = k$. Thus, f' and g' belong to the same row j of the prefix matrix m' and $f' \leqslant g'$ implies $f' = g'$ and $r = s$. Hence $u = v$ and row p of $m\,m'$ indeed forms a prefix family.

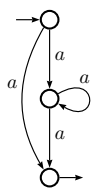
2.4 *Normalise the following automata.*

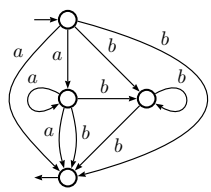
(a)

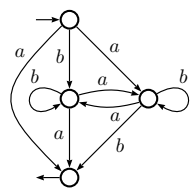
(b) 

(c) 

Solution:

(a) 

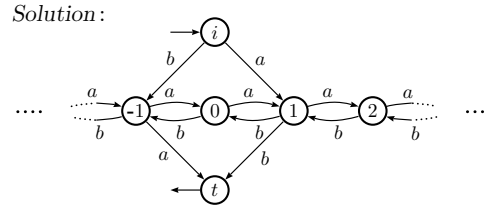
(b) 

(c) 

[73] Exercise adapted from [80].

2.5 *Normalise the automaton $\mathcal{Z}_1$ (see Example 1.3, p. 52).*

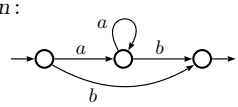
Solution:



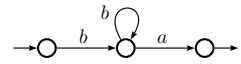
2.6 *Construct a normalised automaton with three states that recognises:*
(a) a^*b ; (b) bb^*a .

Solution:

(a)



(b)



2.7 *Give a procedure for the construction of $\mathcal{A}_{\mathsf{nor}}$ from $\mathcal{A}$ using spontaneous transitions and the construction used in the proof of Lemma 2.3.*

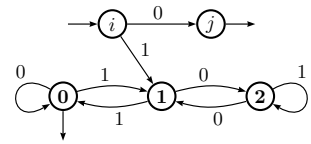
Solution: We first add to $\mathcal{A}$ a state i (which is not in Q) and a spontaneous transition from i to each initial state of $\mathcal{A}$. We then remove these spontaneous transitions by forward closure, then ensure, if need be, that i is not final.

cf. Def. 1.14, p. 80

We then carry out the dual construction: that is, we add to $\mathcal{A}$ a state t (which is not in Q) and a spontaneous transition from each final state of $\mathcal{A}$ to t. We then remove these spontaneous transitions by backward closure and ensure, if need be, that t is not initial.

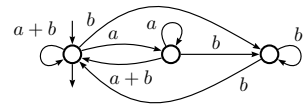
2.8 *Construct a finite automaton that accepts normalised binary representations of numbers divisible by 3.*

Solution: See opposite. The normalised representation of 0 (the number) is 0 (the word) and not the empty word.

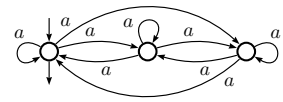


2.9 *Construct an automaton that recognises:*

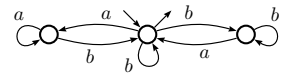
(a) $(a^*b^*)^*$ *from the normalised automaton computed in Exercise 2.4 (b).*



(b) $((a^+)^+)^*$ *from the automaton computed in Example 2.4.*



(c) $(a^*b + bb^*a)^*$ *from the normalised automata computed in Exercise 2.6.*



See also Exer. II.4.6, p. 279 and Exer. IV.1.17, p. 538

2.12 ***Conjugate of a language.*** *Let* $\mathsf{Conj}\colon A^* \to A^*$ *be the relation which relates a word f to the set of its* conjugates*:*

$$\mathsf{Conj}(f) = \{v\,u \mid u, v \in A^* \quad u\,v = f\} \ .$$

The relation extends by additivity: $\mathsf{Conj}(L) = \bigcup_{f \in L} \mathsf{Conj}(f)$. *Show that* $\mathsf{Conj}(L)$ *is recognisable when L is.*

Solution: Let $\mathcal{A} = \langle Q, A, E, i, t \rangle$ be a normalised automaton which recognises $L \setminus 1_{A^*}$, and $\mathcal{A}' = \langle Q', A, E', i', t' \rangle$ a (disjoint) copy of $\mathcal{A}$. As in the proof of Proposition 2.4, construct the (normalised) automaton $\mathcal{C}$ which recognises L^2 by identifying t and i'. For each q in Q, let $\mathcal{C}_q$ be the (not necessarily normalised) automaton obtained from $\mathcal{C}$ by taking q as the initial state and q' as the final state. (*cf.* Figure SE.2.1).

There is a bijection between the successful computations of $\mathcal{C}_q$, written $q \xrightarrow[\mathcal{C}_q]{f_2} t \xrightarrow[\mathcal{C}_q]{f_1} q'$, and the successful computations of $\mathcal{A}$ that go through q, written $i \xrightarrow[\mathcal{A}]{f_1} q \xrightarrow[\mathcal{A}]{f_2} t$. We conclude in particular that, for all q in Q, $L(\mathcal{C}_q) \subseteq \mathsf{Conj}(L)$.

For all f in L and every factorisation $f = f_1 f_2$ such that neither f_1 nor f_2 is equal to the empty word, there exists some q in Q, distinct from i and t, such that $i \xrightarrow[\mathcal{A}]{f_1} q \xrightarrow[\mathcal{A}]{f_2} t$ from which we deduce that $g = f_2 f_1$ is accepted by $\mathcal{C}_q$: that is, that every conjugate of f is either in L or in one of the $L(\mathcal{C}_q)$. Taking into account the previous inclusion, we have

$$\mathsf{Conj}(L) = \bigcup_{q \in Q \setminus \{i,t\}} L(\mathcal{C}_q) \quad \cup \quad L \ ,$$

which proves the property.

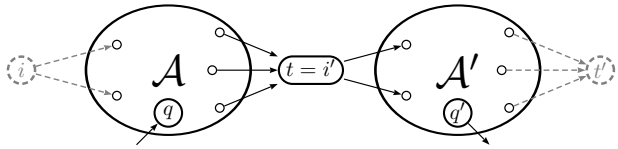
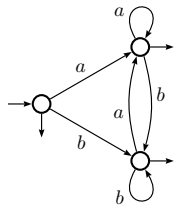


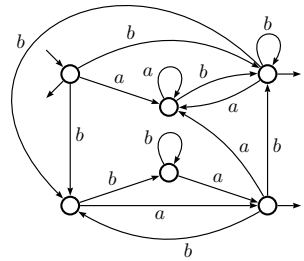
Figure SE.2.1: The automaton $\mathcal{C}_q$

2.13 *Give the standard automata which correspond to the following rational expressions:* (a) $(a+b)^*$; (b) $(a^*b^*)^*$; (c) $(a^*b + bb^*a)^*$.

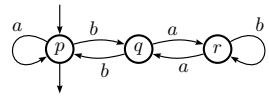
(a) *and* (b) give



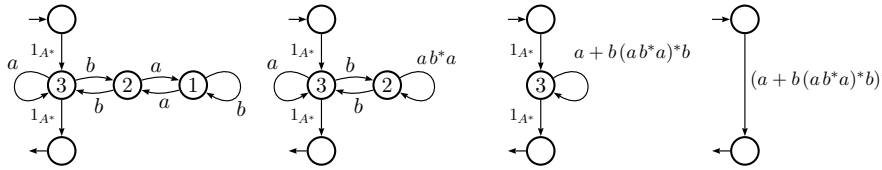
(c) gives



2.15 *Compute the language recognised by the automaton $\mathcal{P}_2$.*

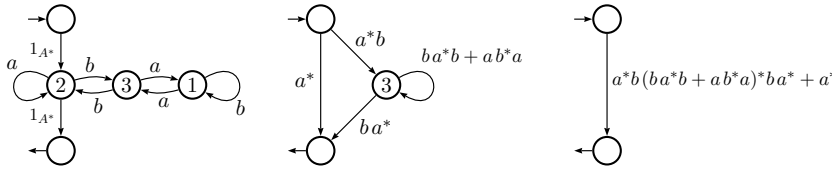


Solution: (a) Using the state elimination method we obtain in four steps, with the order $r < q < p$,



(a) (b) (c) (d)

For the order $r < p < q$, we obtain (the second step, identical to (b) above, is not shown):



(e) (f) (g)

(b) With the McNaughton–Yamada algorithm, and the order $p < q < r$:

$$R^{(0)} = \begin{pmatrix} a & b & 0 \\ b & 0 & a \\ 0 & a & b \end{pmatrix} \qquad R^{(1)} = \begin{pmatrix} a + a\,a^*a & b + a\,a^*b & 0 \\ b + b\,a^*a & 0 + b\,a^*b & a \\ 0 & a & b \end{pmatrix} = \begin{pmatrix} a^+ & a^*b & 0 \\ b\,a^* & b\,a^*b & a \\ 0 & a & b \end{pmatrix}$$

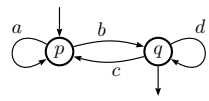
$$R^{(2)} = \begin{pmatrix} a^+ + a^*b\,(b\,a^*b)^*b\,a^* & a^*b + a^*b\,(b\,a^*b)^*b\,a^*b & a^*b\,(b\,a^*b)^*a \\ b\,a^* + b\,a^*b\,(b\,a^*b)^*b\,a^* & b\,a^*b + b\,a^*b\,(b\,a^*b)^*b\,a^*b & a + b\,a^*b\,(b\,a^*b)^*a \\ a\,(b\,a^*b)^*b\,a^* & a + a\,(b\,a^*b)^*b\,a^*b & b + a\,(b\,a^*b)^*a \end{pmatrix}$$

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

$$R^{(3)} = \begin{pmatrix} a^+ + a^*b\,(b\,a^*b)^*b\,a^* + a^*b\,(b\,a^*b)^*a\,[b + a\,(b\,a^*b)^*a]^*a\,(b\,a^*b)^*b\,a^* & & \cdots & -- & -- \\ -- & & \cdots & -- & -- \\ -- & & \cdots & -- & [b + a\,(b\,a^*b)^*a]^* \end{pmatrix}$$

whence $L(\mathcal{P}_2) = a^* + a^*b\,(b\,a^*b)^*b\,a^* + a^*b\,(b\,a^*b)^*a\,[b + a\,(b\,a^*b)^*a]^*a\,(b\,a^*b)^*b\,a^*$.

2.16 *Compute the language recognised by the automaton in the adjacent figure.*



Solution: With MNY, and $p < q$, we obtain:

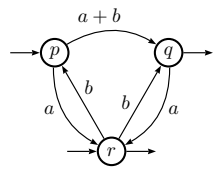
$$R^{(0)} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad , \qquad R^{(1)} = \begin{pmatrix} a + a\,(a)^*a & b + a\,(a)^*b \\ c + c\,(a)^*a & d + c\,(a)^*b \end{pmatrix} = \begin{pmatrix} a^+ & a^*b \\ c\,a^* & d + c\,a^*b \end{pmatrix}$$

$$R^{(2)} = \begin{pmatrix} a^+ + a^*b\,(d + c\,a^*b)^*c\,a^* & a^*b + a^*b\,(d + c\,a^*b)^*(d + c\,a^*b) \\ c\,a^* + (d + c\,a^*b)(d + c\,a^*b)^*c\,a^* & d + c\,a^*b + (d + c\,a^*b)\,(d + c\,a^*b)^*(d + c\,a^*b) \end{pmatrix}$$

$$R = R^{(2)} + I = \begin{pmatrix} a^* + a^*b\,(d + c\,a^*b)^*c\,a^* & a^*b\,(d + c\,a^*b)^* \\ (d + c\,a^*b)^*c\,a^* & (d + c\,a^*b)^* \end{pmatrix}$$

and the language we seek is $R_{p,q} = a^*b\,(d + c\,a^*b)^*$.

2.17 *Compute the language recognised by the automaton in the adjacent figure.*



Solution: (a) With MNY, and $p < q < r$, we obtain

$$R^{(0)} = \begin{pmatrix} 0 & a + b & a \\ 0 & 0 & a \\ b & b & 0 \end{pmatrix}$$

and, setting $A = a + b$ and $B = a + b + 1$,

$$R^{(1)} = \begin{pmatrix} 0 & A & B\,a \\ 0 & 0 & a \\ b & b\,B & b\,a \end{pmatrix} \quad , \qquad R^{(2)} = \begin{pmatrix} 0 & A & a \\ 0 & 0 & a \\ b & b\,B & b\,B\,a \end{pmatrix}$$

and finally $\quad R^{(3)}_{1,2} = A + B\,a\,(b\,B\,a)^*b\,B \qquad R^{(3)}_{1,3} = B\,a + B\,a\,(b\,B\,a)^*b\,B\,a = B\,a\,(b\,B\,a)^*$

$$R^{(3)}_{3,2} = b\,B + (b\,B\,a)^+b\,B = (b\,B\,a)^*b\,B \qquad R^{(3)}_{3,3} = b\,B\,a + (b\,B\,a)^+b\,B\,a \quad = (b\,B\,a)^+$$

from which:

$$\begin{aligned} L(\mathcal{A}) &= R^{(3)}_{1,2} + R^{(3)}_{1,3} + R^{(3)}_{3,2} + R^{(3)}_{3,3} + 1 \\ &= A + B\,a\,(b\,B\,a)^*b\,B + B\,a\,(b\,B\,a)^* + (b\,B\,a)^*b\,B + (b\,B\,a)^* \quad . \end{aligned}$$

(b) With the system of equations method, we have on one hand $L = L_p + L_r$ and on the other the system of equations

$$L_p = (a+b)\,L_q + a\,L_r \qquad \text{(SE.2.1)}$$
$$L_q = a\,L_r + 1 \qquad \text{(SE.2.2)}$$
$$L_r = bL_p + bL_q + 1 \qquad \text{(SE.2.3)}$$

Substituting (SE.2.2) into (SE.2.1), we get

$$L_p = (a+b)\,(a\,L_r + 1) + a\,L_r = (a+b+1)\,a\,L_r + a + b \ .$$

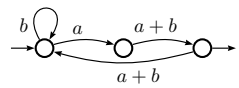
Set as above $A = a+b$ and $B = a+b+1$ and continue with elimination:

$$\begin{aligned} L_p &= B\,a\,L_r + A \\ L_r &= b\,(B\,a\,L_r + A) + b\,(a\,L_r + 1) + 1 = b\,B\,a\,L_r + b\,a\,L_r + b\,A + b + 1 \\ &= b\,B\,a\,L_r + b\,B + 1 \end{aligned}$$

since $B = A + 1$ and $B = B + 1$ (idempotence of union). Hence

$$\begin{aligned} L_r &= [b\,B\,a]^*[b\,B + 1] \\ L_p &= B\,a\,[b\,B\,a]^*[b\,B + 1] + A \\ L &= [B\,a + 1]\,[b\,B\,a]^*[b\,B + 1] + A \\ &= [(a+b+1)\,a + 1]\,[b\,(a+b+1)\,a]^*[b\,(a+b+1) + 1] + a + b \ . \end{aligned}$$

2.19 *Compute the language recognised by the automaton in the adjacent figure.*



Solution: The system of equations method gives

$$(a\,(a+b)^2 + b)^* a\,(a+b)\ .$$

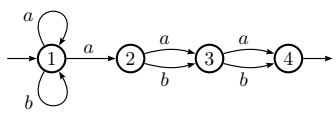
2.20 *Give the general form of solutions of* $X = K\,X + L$ *when* $1_{A^*} \in K$.

Solution: P is a solution if and only if $P = K^*R$ where $L \subseteq R$.

3. The functional point of view

3.1 *Give the transition table of the following automata:*

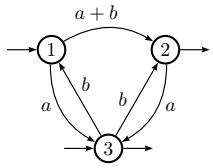
(a)



Solution:

	${}^I\{1\}$	$\{2\}^T$	$\{3\}$	$\{4\}^T$
a	$\{1,2\}$	$\{3\}$	$\{4\}$	$\emptyset$
b	$\{1\}$	$\{3\}$	$\{4\}$	$\emptyset$

(b)



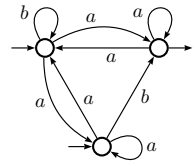
Solution:

	${}^I\{1\}$	$\{2\}^T$	${}^I\{3\}^T$
a	$\{2,3\}$	$\{3\}$	$\emptyset$
b	$\{2\}$	$\emptyset$	$\{1,2\}$

3.2 *Give the graph representation of the automaton defined by the following transition table:*

	${}^Ip$	q^T	${}^Ir$
a	$\{q,r\}$	$\{p,q\}$	$\{p,r\}$
b	p	$\emptyset$	q

Solution:



3.3 *Say what it means in terms of the transition function for an automaton to be complete.*

Solution: $\mathcal{A} = \langle\, Q, A, \delta, I, T \,\rangle$ is complete if δ is defined on the whole of $Q \times A$.

3.4 *The function* $\delta\colon \mathfrak{P}(Q)\times A^* \to \mathfrak{P}(Q)$ *is defined from* E *with the relations:*

$$\forall P \subseteq Q \qquad (P, 1_{A^*})\delta = P \tag{3.3}$$

$$\forall P \subseteq Q\,,\ \forall f \in A^*\,,\ \forall a \in A \qquad (P, fa)\delta = ((P, f)\delta, a)\delta \tag{3.4}$$

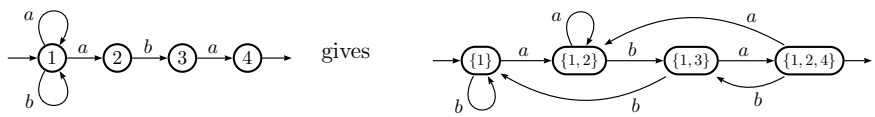
Prove Property 3.1: $\forall P \subseteq Q\,,\ \forall f, g \in A^* \qquad (P, fg)\delta = ((P, f)\delta, g)\delta\,.$

Solution: By induction on $|g|$. The property is true for $|g| = 0$ by (3.3), and for $|g| = 1$ by (3.4). For all P in $\mathfrak{P}(Q)$, for all a in A, and for all f and g in A^*, we have

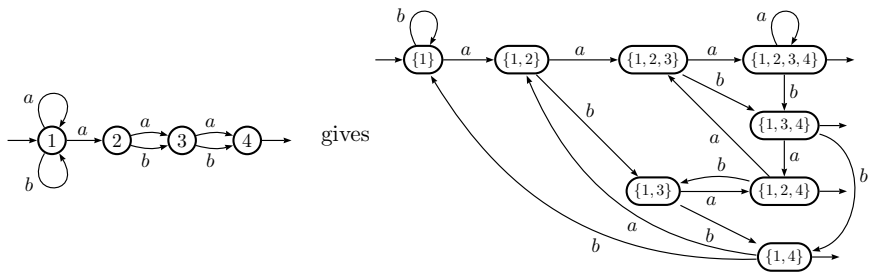
$$\begin{aligned}
(P, fag)\delta &= ((P, fa)\delta, g)\delta && \text{by the induction hypothesis,}\\
&= (((P, f)\delta, a)\delta, g)\delta && \text{by (3.4),}\\
&= ((P, f)\delta, ag)\delta && \text{by the induction hypothesis.}
\end{aligned}$$

3.5 *Determinise the automata below.*

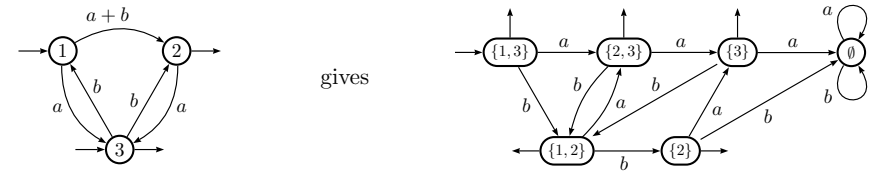
(a)



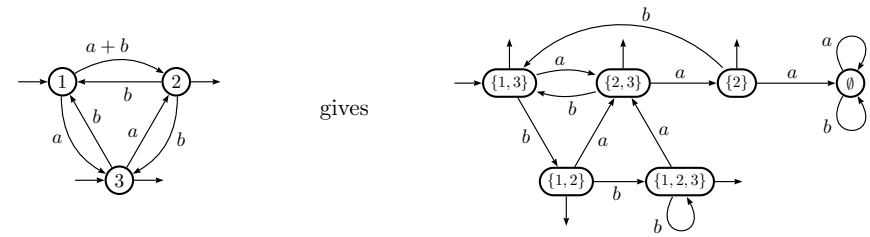
(b)



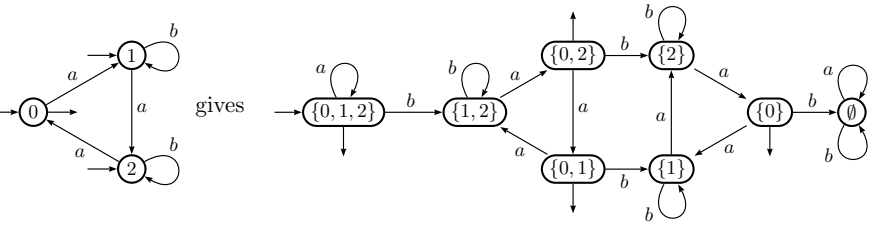
(c)



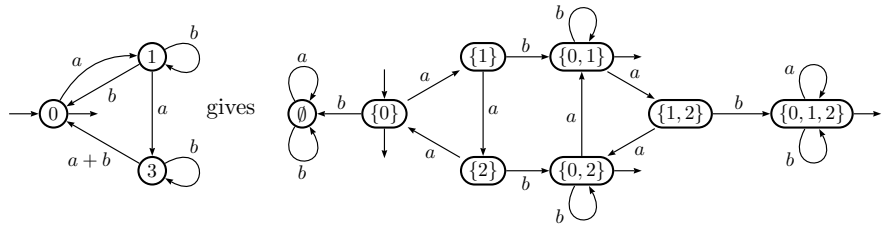
(d)



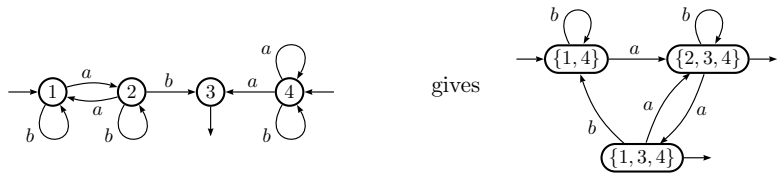
(f)



(g)



(h)



3.6 *Which of the automata in the previous exercise suggest the definition of families of automata (indexed by n) that have n states and whose determinisation has exactly 2^n states?* *cf. Fact 3.1, p. 107*

Solution: Set $Q_n = \mathbb{Z}/n\mathbb{Z}$ and $\mathcal{C}_n = \langle\, Q_n, A, \delta, Q_n, \{0\}\,\rangle$ with $i \cdot a = i + 1$ for all i in Q_n and $i \cdot b = i$ for $i \neq 0$; $(0, b)\delta$ is undefined ($\mathcal{C}_3$ is shown in (f) above).

By induction on $k = n - \|P\|$, we show that every subset P of Q_n is accessible in the determinisation of $\mathcal{C}_n$. The property is true for $k = 0$ since Q_n is the set of initial states of $\mathcal{C}_n$. Let P be of cardinal $n - k - 1$. If $0 \notin P$, $R = P \cup \{0\}$ is of cardinal $n - k$, is accessible by the induction hypothesis, and $P = R \cdot b$. If $0 \in P$, let $j \notin P$ and $0 \notin P' = P \cdot a^{n-j}$; then $R = P' \cup \{0\}$ is accessible and $P = R \cdot b\, a^j$. Thus, the determinisation of $\mathcal{C}_n$ has 2^n states.

In a similar way, we define $\mathcal{D}_n = \langle\, Q_n, A, \eta, \{0\}, \{0\}\,\rangle$ with $i \cdot a = i + 1$ for all i in Q_n and $i \cdot b = \{0, i\}$ for $i \neq 0$ ($(0, b)\eta$ is undefined; $\mathcal{D}_3$ is shown in (g) above) and we show that the determinisation of $\mathcal{D}_n$ has 2^n states.[74]

Note that the automaton $\mathcal{E}_n$ which recognises words of length greater than or equal to n whose nth but last letter is an a – we considered $\mathcal{E}_3$ in (b) above – has $n + 1$ states and its determinisation has 2^n: that is already a combinatorial explosion, not as great as possible, but simpler to show.

3.7 *Find the determinisation of the transpose of $\mathcal{P}_1$.*

Solution: We find $\mathcal{P}_2$!

3.10 ***Discriminating length I.*** (a) *Show (using the methods of this section) that two complete deterministic automata $\mathcal{A}$ and $\mathcal{B}$, with n and m states respectively, are equivalent if and only if they accept the same words of length less than $k = n\,m$.*

(b) *Deduce a bound for the discriminating length of any two automata with n and m states respectively.*

Solution: (a) The condition is obviously necessary. The automaton which recognises $\overline{L(\mathcal{A})} \cap L(\mathcal{B})$, like that which recognises $L(\mathcal{A}) \cap \overline{L(\mathcal{B})}$, is constructed from $\mathcal{A} \times \mathcal{B}$, which has $n\,m$ states, by modifying the initial and final states. Such an automaton does not recognise any word unless it recognises some word of length less than $k = n\,m$.

(b) A direct consequence of Proposition 3.2 is that two automata $\mathcal{C}$ and $\mathcal{D}$, with r and s states respectively, are equivalent if and only if they accept the same words of length less than $l = 2^{r+s}$.[75]

---

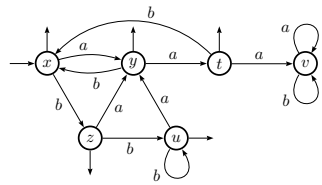
[74] Example $\mathcal{C}_n$ is adapted from an example in [251]; Example $\mathcal{D}_n$ seems to belong to folklore and was passed on to me by A. Arnold.

[75] We will see in Chapter III that the bound k can be reduced to $m + n$ and that it is then optimal, and that the bound l is then smaller than $2^r + 2^s$ (*cf.* Exer. III.4.13 and Rem. III.SE.4.1).

3.13 *Minimise the deterministic automata calculated in Exercise 3.5.*

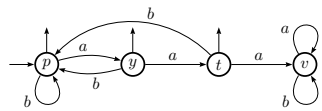
Solution: In (a), (b), (f) and (h) the initial automata are co-deterministic (and co-accessible); their determinisations are therefore minimal. We verify that in (c) and (g) the determinisations are also minimal.

Moore's algorithm on the determinisation computed in (d) (whose states have been renamed; see opposite) gives:



$$\begin{aligned}
&\mathcal{P}_0 = x, y, z, t, u/v \\
&\mathcal{P}_1 = x, y, z, u/t/v && \text{since } t \cdot a \not\equiv u \cdot a \bmod \mathcal{P}_0 \\
&\mathcal{P}_2 = x, z, u/y/t/v && \text{since } y \cdot a \not\equiv u \cdot a \bmod \mathcal{P}_1 \ .
\end{aligned}$$

Hence the minimal automaton $\mathcal{X}_5$ in the adjacent figure (we have called p the subset $\{x, z, u\}$).



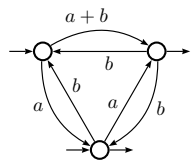
3.14 *Which of the automata in Exercise 3.5 suggest the definition of families of automata (indexed by n) which are evidence for Fact 3.2; that is, automata with n states that recognise languages whose minimal automaton has 2^n states?*

cf. pp. 107 and 114

cf. p. 195

Solution: As in Exercise 3.6, the examples of Exercise 3.5(f) and (g) satisfy the property; the automaton (f) is co-deterministic, so its determinisation is therefore minimal; checking that (g) also satisfies the property is more laborious.

3.15 *Find the language recognised by the adjacent automaton.*



Solution: We would obviously do well to use the minimal automaton $\mathcal{X}_5$ computed above. We have $L(\mathcal{X}_5) = L_p$ and obtain the system of equations:

$$\begin{aligned}
&L_v = \emptyset \ , \qquad L_t = b\,L_p + 1 \\
&L_y = b\,L_p + a\,L_t + 1 = b\,L_p + a\,b\,L_p + a + 1 \\
&L_p = b\,L_p + a\,L_y + 1 = b\,L_p + a\,b\,L_p + a^2 b\,L_p + a^2 + a + 1
\end{aligned}$$

which can be immediately solved:

$$L_p = (a^2 + a + 1)\,b\,L_p + a^2 + a + 1 = ((a^2 + a + 1)\,b)^*(a^2 + a + 1) \ .$$

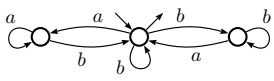
The language we want is the set of words that does not contain the factor a^3, which can be seen equally well for $\mathcal{X}_5$ and for the expression computed for L_p.

3.16 *Find the minimal automaton of the language recognised by $\mathcal{P}_1$.*

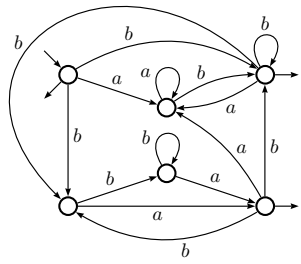
Solution: We find $\mathcal{P}_2$, obviously!

3.17 *Determinise and minimise the following automata:*

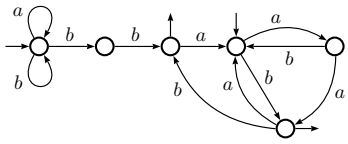
(a)



(b)

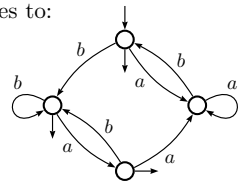


(c)

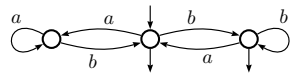


Solution:

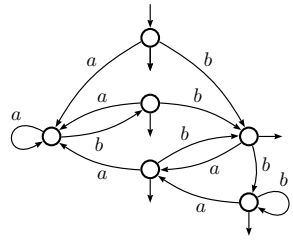
(a) determinises to:



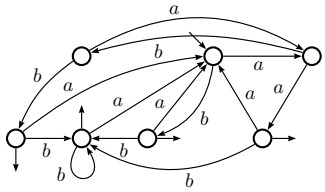
then minimises to:



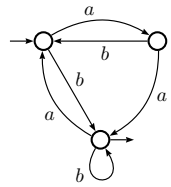
(b) determinises to the automaton in the adjacent figure. The minimal automaton is the same as in (a) since the language recognised is $(a^*b + bb^*a)^*$ in both cases (*cf.* Exercise 2.9 (c) p. 190 and Exercise 2.13 p. 191). The states which are conflated by Moore's algorithm are aligned vertically.



(c) determinises to:

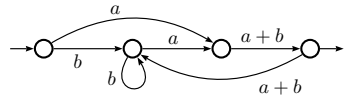


then minimises to:

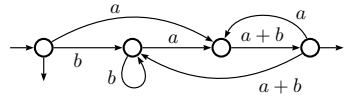


3.18 *Give the minimal automaton of the language* $\big((a\,(a+b)^2+b)^*a\,(a+b)\big)^*$.

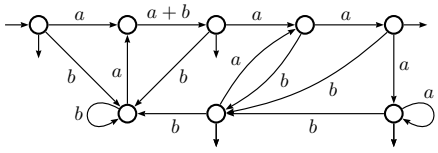
Solution: We saw in Exercise 2.19 p. 193 that the language $(a\,(a+b)^2+b)^*a\,(a+b)$ is recognised by an automaton whose standard form is shown in the adjacent figure.



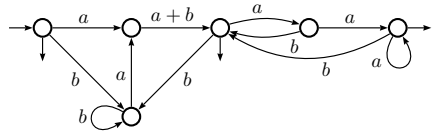
From this we obtain an automaton for the star:



which determinises to:



then minimises to:



3.20 *What is the effect of Moore's algorithm on a non-deterministic automaton?*

Solution: It produces an automaton which recognises the same language and which is equally minimal for a certain property, which holds for the automaton rather than the language. We do not yet have the words to describe it; we will in Section II.3.2 (p. 261).

3.21 ***State complexity of operations on languages.*** *Some authors*[76] *call the number of states of the minimal automaton of a recognisable language L its* state complexity. *We will write it $\mathsf{Sc}(L)$. We can measure the effect that different operations on languages have on this complexity. Show that:*

(a) $\mathsf{Sc}(L \cup K) \leqslant \mathsf{Sc}(L)\,\mathsf{Sc}(K)$; (b) $\mathsf{Sc}(L \cap K) \leqslant \mathsf{Sc}(L)\,\mathsf{Sc}(K)$;

(c) $\mathsf{Sc}(L\,K) \leqslant (2\,\mathsf{Sc}(L) - 1)\,2^{\mathsf{Sc}(K)-1}$; (d) $\mathsf{Sc}(L^*) \leqslant 3\;2^{\mathsf{Sc}(L)-2}$;

(e) $\mathsf{Sc}(L^{\mathsf{t}}) \leqslant 2^{\mathsf{Sc}(L)}$.

and that all these bounds are tight.

Solution: Cases (a) and (b) follow from the fact that the union and intersection of two languages are recognised by the cartesian product of the automata that recognise the languages (with the exception of the definition of final states for the union); *cf.* Proposition 1.4 and Exercise 1.33. It suffices to take the languages of words whose length is divisible by l and k (l and k co-prime) to see that the bounds are tight.

cf. p. 61 and p. 188

(d) The bound is reached by the automata $\mathcal{Y}_n = \langle\, Q_n, A, \theta, \{0\}, \{n-1\}\,\rangle$ where, as in Exercise 3.6, $Q_n = \mathbb{Z}/n\,\mathbb{Z}$, and whose transition functions are defined by $i \cdot a = i + 1$ for all i in Q_n, with $i \cdot b = i + 1$ for $i \neq 0$ with $i \cdot 0 = 0$.

The language recognised by $\mathcal{Y}_3$ was computed in Exercise 2.19, and the minimal automaton of $(L(\mathcal{Y}_3))^*$ in Exercise 3.18: it indeed has six states.

(e) The inequality is obviously true since if L is recognised by a (complete) deterministic automaton with n states, L^{t} is recognised by a probably non-deterministic automaton, but with n states, whose determinisation has up to 2^n states.

The bound is tight for the $\mathcal{C}_n$ considered in the solution to Exercise 3.6: $\mathcal{C}_n^{\mathsf{t}}$ is deterministic and the minimal automaton of the language recognised by $\mathcal{C}_n = (\mathcal{C}_n^{\mathsf{t}})^{\mathsf{t}}$ has 2^n states.

3.23 ***Another characteristic iteration lemma.***[77] *Imitating the families $\mathfrak{E}'_h$ and $\mathfrak{E}_h$, we define $\mathfrak{J}'_h(A^*)$ by*

$$L \in \mathfrak{J}'_h(A^*) \iff \forall f \in A^h \quad \exists \text{ a factorisation } f = u\,v\,w\ ,\ \ |v| \geqslant 1$$
$$\forall t \in A^*,\ \forall n \in \mathbb{N} \qquad f\,t \in L \iff u\,(v)^n\,w\,t \in L\ .$$

and $\mathfrak{J}_h(A^*)$ by

$$L \in \mathfrak{J}_h(A^*) \iff \forall f \in A^h \quad \exists \text{ a factorisation } f = u\,v\,w\ ,\ \ |v| \geqslant 1$$
$$\forall t \in A^*,\ f\,t \in L \iff u\,w\,t \in L\ .$$

(a) *Show:*

Proposition 3.15 *Let L be a language of A^*. The following three conditions are equivalent:*

(i) *L is recognisable;*

(ii) *there exists an integer h such that L satisfies $\mathfrak{J}'_h$;*

(ii) *there exists an integer h such that L satisfies $\mathfrak{J}_h$.*

(b) *Explain why Proposition 3.15 cannot be considered a combinatorial property of the words of L in the same way as Theorem 3.3.*

(c) *Show that there exist, for every h, two languages over A^*, J_h and Y_h, which both satisfy $\mathfrak{J}_h$ and not $\mathfrak{J}_{h-1}$, and such that the minimal automaton of J_h has h states and that of Y_h has $\|A\|^h - 1$ states.*

---

[76] This exercise is adapted from S. Yu [251, Sec. 5]; the proofs and solutions we omit will be found there with a bibliography relevant to this topic. *Added in translation:* an early reference for state complexity of operations is A. Maslov [164]. The state complexity of the conjugacy operation (*cf.* Exer. 2.12, p. 92) has been determined recently by G. Jirásková and A. Okhotin [130].

[77] Adapted from J. Jaffe, [129]

Solution: (a) If L is recognised by a complete deterministic automaton with h states, it satisfies $\mathfrak{J}'_h$. Clearly, $\mathfrak{J}'_h$ implies $\mathfrak{J}_h$.

Let L satisfy $\mathfrak{J}_h$; we shall show, by induction on the length of words, that for all g in A^* there exists a k of length less than h such that $g^{-1}L = k^{-1}L$. This entails L having only a finite number of left quotients, so it must be recognisable.

The property holds for words of length less than h. Suppose it holds for all words of length less than M, $M \geqslant h$; let g be a word of length M and f its prefix of length h: $g = f\,g'$. Since $L \in \mathfrak{J}_h(A^*)$, there exists a factorisation $f = u\,v\,w$, $|v| \geqslant 1$, such that for all t in A^*, $u\,v\,w\,g'\,t \in L \Leftrightarrow u\,w\,g'\,t \in L$: that is $g^{-1}L = (u\,w\,g')^{-1}L$, and the length of $u\,w\,g'$ is strictly less than that of g.

(b) The conditions that define $\mathfrak{J}'_h$ and $\mathfrak{J}_h$ apply not only to the words of A^* and their factorisations, but also contain a universal quantifier on the right contexts of these words.

(c) The automata of Exercise 3.22 recognise the languages Y_h. *cf. p. 122*

3.24 ***An application***[78] ***of Ramsey's Theorem to free monoids.*** (a) *Give a proof of the following proposition based on Ramsey's Theorem and comment.*

Proposition 3.16 *For every integer m, and every integer r, there exists an integer $N = S(m,r)$ such that, for every partition $\mathcal{P}$ of A^+ into m classes, every word of length greater than N contains r consecutive factors in the same class of $\mathcal{P}$.*

(b) *Show directly that $S(m,r) \leqslant r^m$, and that this bound is tight if $\|A\| = m$.*

Solution: (a) We have $S(m,r) \leqslant R(2,m,r+1)$. Let $f = a_1\,a_2 \cdots a_n$ be a word of length n in A^* and E the set $\{0,1,\ldots,n\}$. With each pair of indices (i,j), $0 \leqslant i < j \leqslant n$ we associate the factor $f_{i,j} = a_{i+1}\,a_{i+2} \cdots a_j$. The partition $\mathcal{P}$ of A^+ induces a partition into m classes of $\mathfrak{P}_{(2)}(E)$.

If $n \geqslant R(2,m,r+1)$, there exists a subset $F = \{i_0, i_1, \ldots, i_r\}$ of E such that all the pairs (i_l, i_h), and in particular all the r consecutive factors $h_j = f_{i_{j-1},i_j}$, for $1 \leqslant j \leqslant r$, are in the same class of A^+ modulo $\mathcal{P}$.

This reasoning itself implies that it is not only the r factors h_j that are in the same class, but also that every factor $h_j h_{j+1} \cdots h_{j+l}$ belongs to that class. This is obviously a much stronger property than the one stated in the proposition, but somehow resorting to Ramsey's Theorem does not allow us to do less.

(b) Let $\{X_1, X_2 \ldots X_m\}$ be a partition of A^+ into m classes and $\lambda\colon A^* \to \mathbb{N}^m$ the map defined by

$$f\lambda = (n_1, n_2, \ldots, n_m) \qquad \text{where} \quad (f\lambda)_i = n_i = \mathsf{max}\{k \mid f \in A^*\,[X_i]^k\}\ .$$

If g is a proper prefix of f – that is, $f = g\,h$ – then h belongs to a certain class X_i and necessarily $(g\lambda)_i < (f\lambda)_i$ which implies $g\lambda \neq f\lambda$. It follows that if $|f| \geqslant r^m$ the $r^m + 1$ prefixes of f are sent to distincts elements of $\mathbb{N}^m$ and then at least one prefix g of f is such that $g\lambda$ has at least one coordinate i which is greater than or equal to r: $g \in A^*\,[X_i]^r$ and hence $f \in A^*\,[X_i]^r A^*$ which is exactly the property stated by the proposition.

If $A = \{a_1, a_2 \ldots a_m\}$, let $A_i = \{a_1, a_2 \ldots a_i\}$. Then $X_i = A_i^* \setminus X_{i-1}$, for $1 \leqslant i \leqslant m$ is a partition of A^+. Let g_1, g_2,..., g_m be sequence of words defined by $g_1 = a_1^{r-1}$ and $g_i = (g_{i-1}a_i)^{r-1}g_{i-1}$, for $2 \leqslant i \leqslant m$: g_i is of length $r^m - 1$ and does not contains any mth power modulo the partition $\{X_i\}$.

[78] Exercise taken from M. P. Schützenberger, *Quelques propriétés combinatoires de la théorie des automates* [219]. These mimeographed lecture notes (the ancestor of Lothaire's first volume [158]) contains the proof of (b) for $m = 2$; the generalisation to any m comes from [158].

4. Rational expressions

4.1 (a) *Give a definition of rational expressions using a postfix notation for the operators $+$, $\cdot$ and $*$. Give the expressions that correspond to Examples 4.1 with this convention.*

(b) *Do the same using prefix notation.*

(c) *Draw conclusions.*

Solution: (a) A *suffix rational expression* is a formula obtained inductively from the letters of A and the five function symbols $\{\mathsf{0}, \mathsf{1}, +, \cdot, *\}$ in the following manner:

(i) $\mathsf{0}$, $\mathsf{1}$, and a, for all a in A, are suffix rational expressions;

(ii) if E and F are two suffix rational expressions, then $\mathsf{E}\,\mathsf{F}+$, $\mathsf{E}\,\mathsf{F}\cdot$, and $\mathsf{E}*$ are suffix rational expressions.

The expressions (4.1) are written

$$a\,b\cdot b\,a\cdot +\ ,\quad a\,b+*a\,b\cdot\cdot a\,b+*\cdot \qquad \text{and} \qquad a\,b\,a\,b*\cdot a\cdot *b\cdot\cdot +*\ .$$

(b) A *prefix rational expression* is a formula obtained inductively from the letters of A and the five function symbols $\{\mathsf{0}, \mathsf{1}, +, \cdot, *\}$ in the following manner:

(i) $\mathsf{0}$, $\mathsf{1}$, and a, for all a in A, are prefix rational expressions;

(ii) if E and F are two prefix rational expressions, then $+\mathsf{E}\mathsf{F}$, $\cdot\mathsf{E}\mathsf{F}$, and $*\mathsf{E}$ are prefix rational expressions.

The expressions (4.1) are written:

$$+\cdot a\,b\cdot b\,a\ ,\quad \cdots *+a\,b\,a\,b*+a\,b \qquad \text{and} \qquad *+a\cdot b\cdot *\cdot\cdot a*b\,a\,b\ .$$

(c) Infix notation and parentheses are a most excellent invention!

4.2 *Give a rational expression for the language of words that do not contain a factor $b\,a$.*

Solution: a^*b^* .

4.3 *Prove the following identities:* (a) $(a+b)^* = a^* + a^*b\,(a+b)^*$;
(b) $(a+b)^* = a^*(b^+a^+)^*b^*$; (c) $(a+b)^+ = a^+ + a^*(b\,a^*)^+$.

Solution:

$$\begin{aligned}
(a+b)^* &\equiv a^*(b\,a^*)^* &&\equiv a^*(1+(b\,a^*)^+) \qquad \text{(a)}\\
&\equiv a^* + a^*(b\,a^*)^+ &&\equiv a^* + a^*b\,a^*(b\,a^*)^*\\
&\equiv a^* + a^*b\,(a+b)^*
\end{aligned}$$

$$\begin{aligned}
(a+b)^* &\equiv (b^*a)^*b^* &&\equiv ((1+b^+)\,a)^*b^* \qquad \text{(b)}\\
&\equiv (a+b^+a)^*b^* &&\equiv a^*(b^+a\,a^*)^*b^*\\
&\equiv a^*(b^+a^+)^*b^*
\end{aligned}$$

$$\begin{aligned}
(a+b)^+ &\equiv (a+b)\,(a+b)^* &&\equiv (a+b)\,a^*(b\,a^*)^* \qquad \text{(c)}\\
&\equiv a^+(b\,a^*)^* + b\,a^*(b\,a^*)^* &&\equiv a^+(1+(b\,a^*)^+) + (b\,a^*)^+\\
&\equiv a^+ + a^+(b\,a^*)^+ + (b\,a^*)^+ &&\equiv a^+ + a^*(b\,a^*)^+
\end{aligned}$$

4.4 *Prove the identities:*

(a) $(a+b+c)^* = [a+b+c\,(c+b)^*a]^*\,[1+c(c+b)^*]$;

(b) $(x+y)^* = [x+y\,(y^+x)^*x]^*\,[1+y\,(y^+x)^*y^*]$.

Solution: (a) Set $X=(a+b+c)^*$, the set of all words (over $\{a,b,c\}$), $Y=(a+b)^*$, the set of words which do not contain c, and $Z=(c+b)^*$, the set of words which do not contain a. We have

$$X = Y + Y\,c\,X$$

and

$$X = Z + Z\,a\,X$$

from which

$$X = Y + Y\,c\,Z + Y\,c\,Z\,a\,X \ ,$$

which is solved by Arden's Lemma:

$$X = (Y\,c\,Z\,a)^*(Y+Y\,c\,Z) = (Y\,c\,Z\,a)^*Y\,(1+c\,Z) \ ,$$

that is,

$$(a+b+c)^* = ((a+b)^*c\,(c+b)^*a)^*\,(a+b)^*\,(1+c\,(c+b)^*) \ ,$$

which is transformed into the required identity using

$$(e+f)^* = (e^*f)^*e^* \qquad \text{with} \quad e=a+b \quad \text{and} \quad f=c\,(c+b)^*a \ .$$

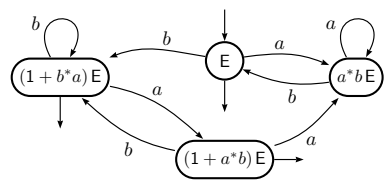
(b) We therefore have the series of equalities:

$$\begin{aligned}(x+y)^* = (y^*x)^*y^* \quad &= \quad (x+y\,x+y\,y^+x)^*y^* \\ &= [x+y\,x+y\,y^+x\,(y\,x+y\,y^+x)^*x]^*\,[1+y\,y^+x\,(y\,x+y\,y^+x)^*]y^* \\ &= [x+y\,x+y\,(y^+x)^+x]^*\,[1+y\,(y^+x)^+]y^* \\ &= [x+y\,(y^+x)^*x]^*\,[1+y\,y^*+y\,(y^+x)^+y^*] \\ &= [x+y\,(y^+x)^*x]^*\,[1+y\,(y^+x)^*y^*] \ .\end{aligned}$$

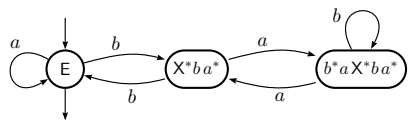
4.9 *Calculate the derivative automaton of the following expressions [we write* EFG *for* $\mathsf{E}\,(\mathsf{FG})$ *]:*

Solution:

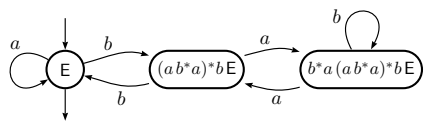
(a) $\mathsf{E} = (a^*b+b\,b^*a)^*$. We obtain:



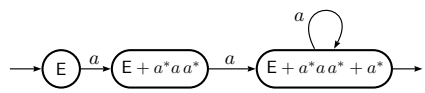
(b) $\mathsf{E} = a^*b\,(a\,b^*a+b\,a^*b)^*b\,a^*+a^*$.
Set $\mathsf{X} = a\,b^*a+b\,a^*b$. We obtain:



(c) $\mathsf{E} = (a+b\,(a\,b^*a)^*b)^*$. We obtain:



(d) $\mathsf{E} = a^*a\,a^*a\,a^*$. We obtain:



4.10 ***Differentiation and associativity of product.***[79]

(a) *Compute the derivative automaton of* $\mathsf{E}'_1 = \Big(\big((a+b)^* \cdot a\big) \cdot b\Big) \cdot (a+b)^*$.

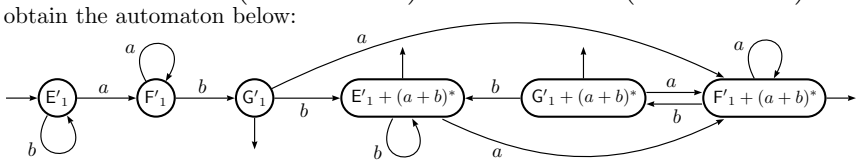
(b) *Do the same for* $\mathsf{E}'_2 = \Big(\big((a^* \cdot a) \cdot a^*\big) \cdot a\Big) \cdot a^*$.

(c) *And again for* $\mathsf{E}_3 = a \cdot \Big(b \cdot \big(c \cdot (a \cdot b)\big)^*\Big)$ *and* $\mathsf{E}'_3 = (a \cdot b) \cdot \big(c \cdot (a \cdot b)\big)^*$.

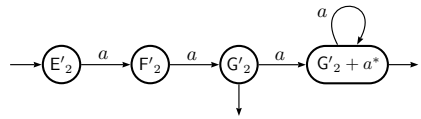
(d) *Comment and draw conclusion.*

Solution: (a) Set: $\mathsf{F}'_1 = \Big(\big((a+b)^* \cdot a + 1\big) \cdot b\Big) \cdot (a+b)^*$ and $\mathsf{G}'_1 = \Big(\big((a+b)^* \cdot a\big) \cdot b + 1\Big) \cdot (a+b)^*$.

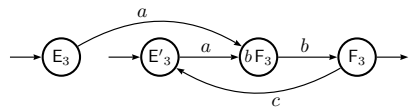
cf. Fig. 4.2, p. 144 We obtain the automaton below:



(b) Set: $\mathsf{F}'_2 = \Big(\big((a^* \cdot a + 1) \cdot a^*\big) \cdot a\Big) \cdot a^*$ and $\mathsf{G}'_2 = \Big(\big((a^* \cdot a + 1) \cdot a^* + a^*\big) \cdot a + 1\Big) \cdot a^*$. We obtain the automaton in the adjacent figure.



(c) Set: $\mathsf{F}_3 = \big(c \cdot (a \cdot b)\big)^*$. The figure opposite shows the derivative automaton for both E_3 and E'_3: for E_3 when E'_3 is not initial, and for E'_3 when E_3 is not initial (and neither accessible).

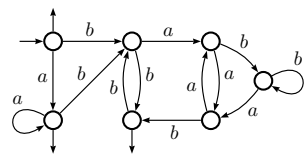


(d) With the examples of E'_1 compared to $\mathsf{E}_1 = (a+b)^* \cdot \Big(a \cdot \big(b \cdot (a+b)^*\big)\Big)$, and of E'_2 compared to $\mathsf{E}_2 = a^* \cdot \Big(a \cdot \big(a^* \cdot (a \cdot a^*)\big)\Big)$, we have $\mathbf{nd}(\mathsf{E}_1) < \mathbf{nd}(\mathsf{E}'_1)$ and $\mathbf{nd}(\mathsf{E}_2) < \mathbf{nd}(\mathsf{E}'_2)$: that is, two cases where $\mathbf{nd}(\mathsf{E} \cdot (\mathsf{F} \cdot \mathsf{G})) \leqslant \mathbf{nd}((\mathsf{E} \cdot \mathsf{F}) \cdot \mathsf{G})$. With the example of E_3 and E'_3 we have the inequality $\mathbf{nd}(\mathsf{E}_3) > \mathbf{nd}(\mathsf{E}'_3)$: that is, a case where $\mathbf{nd}(\mathsf{E} \cdot (\mathsf{F} \cdot \mathsf{G})) \geqslant \mathbf{nd}((\mathsf{E} \cdot \mathsf{F}) \cdot \mathsf{G})$. In conclusion, product cannot considered as associative when a rational expression is to be derived; but nothing more can be said independently of the expression itself.

5. From expressions to automata

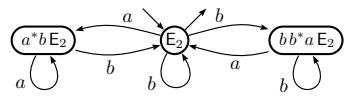
5.1 *Compute the standard automaton of the expression* $(a + b\,(a\,b^*a)^*b)^*$.

We obtain:



5.4 *Compute the derived term automaton of the expression* $\mathsf{E}_2 = (a^*b + b\,b^*a)^*$.

We obtain:



5.5 *Do the same for the expression*

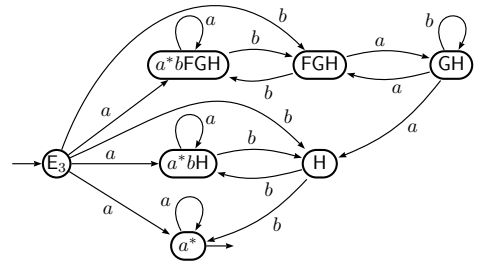
$$\mathsf{E}_3 = a^* + a^*b\,(b\,a^*b)^*b\,a^* + a^*b(b\,a^*b)^*a\,(b + a\,(b\,a^*b)^*a)^*a\,(b\,a^*b)^*b\,a^* .$$

[79] *Note added in translation:* I am grateful to my students Phan Duong Hieu for drawing my attention to this difficulty, and Pierre-Yves Angrand for correcting the erroneous formulation I gave in the French edition.

Solution: If we set

$$\begin{aligned} \mathsf{F} &= (b\,a^*b)^*a \ , \\ \mathsf{G} &= (b + a\,(b\,a^*b)^*a)^*a \qquad \text{and} \\ \mathsf{H} &= (b\,a^*b)^*b\,a^* \ , \quad \text{we have} \\ \mathsf{E}_3 &= a^* + a^*b\,\mathsf{H} + a^*b\,\mathsf{F}\,\mathsf{G}\,\mathsf{H} \ , \end{aligned}$$

and we obtain the automaton shown opposite.



5.6 *Verify from the expressions* $\mathsf{F}_1 = a\,(a+b)^* + b\,(b+a)^*$ *and* $\mathsf{F}_2 = (a\,a + a\,a\,a)^*$ *that the determinisation of the derived term automaton may be larger, or smaller, than the derivative automaton.*

Solution: Immediate computations yield on the one hand that $(\mathcal{A}_{\mathsf{F}_1})_{\mathsf{det}} = \mathcal{A}_{\mathsf{F}_1}$ and $\mathcal{B}_{\mathsf{F}_1}$ have 3 and 2 states respectively, and on the other hand that $(\mathcal{A}_{\mathsf{F}_2})_{\mathsf{det}}$ and $\mathcal{B}_{\mathsf{F}_2}$ have 4 and 6 states respectively.[80]

6. Star height

6.1 (a) *Give an algorithm which calculates the shortest expression that denotes a given rational language, or rather the expression which is, among the shortest ones, the smallest in a given lexicographic order.*

(b) *Explain why this 'minimal' expression, which is therefore canonical in some sense, has no algorithmic interest.*

Solution: (a) The language L over A^* is given (by an expression E). We choose a (total) order on $B = A \cup \{+, \cdot, *\}$. We then enumerate the expressions $\mathsf{F}_1, \mathsf{F}_2, \ldots$ by length, and for each length in the lexicographic order deduced from the previous order (we could use a postfix or prefix notation, to avoid parentheses), combining an algorithm that enumerates the words of B^* in this order with one that tests whether a given word is really a rational expression. *cf. Exer. 4.1, p. 128*

We test, in the order, whether F_i is equivalent to E. The first F_i which gives a positive result is the 'shortest' expression that denotes L. This algorithm terminates at the latest when E appears in the enumeration of expressions.

(b) The computation does not rely on any 'structural' property of L; the result equally reflects none. This 'minimal' expression is of no algorithmic interest: to decide for example whether two expressions E and E' are equivalent, it would take far longer to first compute the minimal expressions F and F' and to test their equality than to apply the equivalence algorithm directly to E and E'.

6.3 *Calculate the loop complexity of* $\mathcal{D}_n$*, the 'double chain' of length* n*, and of* $\mathcal{R}_n$*, the 'double ring' of length* n *[Example 6.2(c) is a double chain of length* 3*, and 6.2(g) a double ring of length* 6*].*

Solution: The definition of loop complexity implies directly *Def. 6.2, p. 160*

$$\mathsf{lc}(\mathcal{R}_n) = \mathsf{lc}(\mathcal{D}_{n-1}) + 1 \ , \qquad \mathsf{lc}(\mathcal{D}_{n+1}) = 1 + \min_{p+q=n} [\mathsf{max}(\mathsf{lc}(\mathcal{D}_p), \mathsf{lc}(\mathcal{D}_q))] \ ,$$

and $p < q \Rightarrow \mathsf{lc}(\mathcal{D}_p) \leqslant \mathsf{lc}(\mathcal{D}_q)$ from which we obtain

$$\mathsf{lc}(\mathcal{D}_{2n-1}) = 1 + \mathsf{lc}(\mathcal{D}_{n-1}) \qquad \text{and} \qquad \mathsf{lc}(\mathcal{R}_{2n}) = 1 + \mathsf{lc}(\mathcal{R}_n)$$

[80] *Note added in translation:* I am grateful to Alexandre Duret-Lutz who drew my attention to this further error in the French edition.

then, as $\mathsf{lc}(\mathcal{R}_2) = 1$, we have $\mathsf{lc}(\mathcal{R}_n) = \lceil \log_2 n \rceil$ and $\mathsf{lc}(\mathcal{D}_n) = \lfloor \log_2 n \rfloor$.

6.5 *Show that the star height of an infinite rational subset of a commutative monoid is equal to* 1.

Solution: Commutativity applied to Property 6.1 implies that an expression of height 1 can be written as (that is, is equivalent to) a finite union $\mathsf{E} = \bigcup_1^k \mathsf{F}_i$ of expressions of the form

cf. Prop. II.7.15, p. 334

$$\mathsf{F}_i = u_i\,(\mathsf{K}_{i,1})^*\,(\mathsf{K}_{i,2})^* \cdots (\mathsf{K}_{i,k})^* = u_i\,(\mathsf{H}_i)^*$$

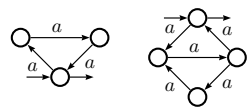
where the $\mathsf{K}_{i,j}$ are finite and $\mathsf{H}_i = \bigcup_{j=1}^{j=k} \mathsf{F}_{i,j}$. It holds $(u\,\mathsf{H}^*)^* = \{1_M\} \cup u\,(u \cup \mathsf{H})^*$ which generalises into

$$\mathsf{E}^* = \{1_M\} \cup \bigcup_{J \subseteq I} \left[\left(\prod_{j \in J} u_j \right) \left(\bigcup_{j \in J} (\{u_i\} \cup H_i) \right)^* \right] ,$$

an expression of height 1.

6.6 *Show, using one of the examples, that the standard construction of an automaton $\mathcal{B}$ that recognises the star of a language recognised by an automaton $\mathcal{A}$ does not necessarily lead to an increase in the loop complexity.*

Solution: The automaton $\mathcal{A}$ that recognises $(a^3)^*$ is the triangle opposite, with loop complexity 1. The standard construction for the star of $L(\mathcal{A})$ gives the square opposite, also of loop complexity 1 (*cf.* Example 6.2 (b)).



6.8 *Give a generalised rational expression with generalised star height equal to* 1 *for the following languages:* (a) $G_2 = (a\,a \cup b\,(a\,a)^*b)^*$; (b) $G_3 = (a\,b^*a \cup b\,(a\,b^*a)^*b)^*$; (c) $G_4 = (a\,b^*a \cup b\,a^*b\,(a\,b^*a)^*b\,a^*b)^*$.

Solution: (a) $G_2 = \overline{\mathsf{E}}$ with:[81]

$$\mathsf{E} = (a\,a \cup b\,a^*b)^* \cdot (a\,b\,(a \cup b)^* \cup b\,(a\,a)^*(1_{A^*} \cup a \cup a\,b\,(a \cup b)^*)) \ .$$

(The equalities $a^* = \overline{\overline{\emptyset}\,b\,\overline{\emptyset}}$ and $(a \cup b)^* = \overline{\emptyset}$ ensure that $\mathsf{gsh}[G_2] = 1$).

(b) $G_3 = (\mathsf{E} \cap \mathsf{F}) \cup \big((b \cup a\,b^*a)\,\mathsf{E} \cap (a^*b\,a^*b\,\mathsf{F})\big)$ with:[82]

$$\mathsf{E} = \big((b \cup a\,b^*a)^2\big)^* \qquad \text{and} \qquad \mathsf{F} = \big((b \cup b^*a\,b^*a\,b^*a\,b^*a)\big)^*$$

(c) Example suggested by Denis Thérien (see [188]); solution due to Mike Robson.

8. A crop of properties

8.1 ***Roots of a recognisable language.*** *The aim of this problem is to show that the set of words whose square, or any fixed power k, is in a recognisable language is itself an effectively computable recognisable language.*

For every language K over A^*, write $\mathcal{R}_2(K)$ for the set of words whose square is in K, and $\mathcal{R}_k(K)$ for the set of words whose kth power is in K:

$$\mathcal{R}_2(K) = \{f \in A^* \mid f\,f \in K\} \ , \qquad \mathcal{R}_k(K) = \{f \in A^* \mid f^k \in K\} \ .$$

[81] Example and solution due to Kosaburo Hashiguchi (private communication (1982)).

[82] Solution due to Henneman (cited in [188]).

(a) First method. *Let* $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ *be a deterministic automaton which recognises* K.

(i) *Show that* $\mathcal{R}_2(K)$ *is recognisable by expressing it in terms of* $L_p = \{f \mid p \cdot f \in T\}$ *and* $L'_q = \{g \mid i \cdot g = q\}$ *, for all* p *and* q *in* Q.

(ii) Generalisation. *Show that* $\mathcal{R}_k(K)$ *is recognisable for all* k.

(b) Preparation for the second method: left regular automata. *A deterministic automaton* $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ *is* left regular *if we have*

$$\forall f, g \in A^* \qquad i \cdot f = i \cdot g \quad \Longrightarrow \quad [\, \forall u \in A^* \quad i \cdot u\, f = i \cdot u\, g \,] \quad . \tag{SE.8.1}$$

Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ *be a finite complete deterministic automaton and let* $\mathcal{A}_{\mathcal{G}} = \langle \mathfrak{T}(Q), A, \eta, \varepsilon, U \rangle$ *be (the accessible subset of) the automaton defined by*

$$\forall r \in \mathfrak{T}(Q)\,,\ \forall a \in A \qquad (r, a)\eta = s \qquad \text{with} \qquad s_p = (r_p, a) \quad \forall p \in Q\delta \quad , \tag{SE.8.2}$$

and by

$$U = \{r \in \mathfrak{T}(Q) \mid r_i \in T\} \quad . \tag{SE.8.3}$$

(i) *Show that the two automata* $\mathcal{S}$ *and* $\mathcal{D}$ *shown in Figure SE.8.1* are not *left regular. Give the graph representation of* $\mathcal{S}_{\mathcal{G}}$ *and of* $\mathcal{D}_{\mathcal{G}}$.

(ii) *Verify that* $\mathcal{A}_{\mathcal{G}}$ *is a complete deterministic left regular automaton and is equivalent to* $\mathcal{A}$.

(c) Second method for the computation of the roots of a language. *Let* $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ *be a deterministic automaton which recognises* K. *Write* V_2 *for the subset of* Q *defined by*

$$V_2 = \{p \in Q \mid \exists f \in A^* \quad i \cdot f = p \quad \text{and} \quad i \cdot f\, f \in T\} \quad ,$$

and let $\mathcal{A}_{\mathcal{R}2} = \langle Q, A, \delta, i, V_2 \rangle$ *be the automaton obtained from* $\mathcal{A}$ *by taking* V_2 *as the set of terminal states.*

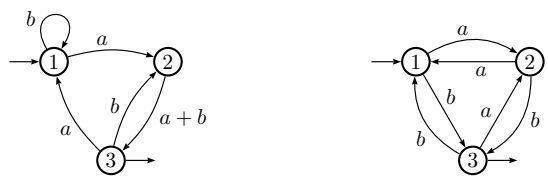
(i) *Show that if* $\mathcal{A}$ *is left regular, we have*

$$L(\mathcal{A}_{\mathcal{R}2}) = \mathcal{R}_2\,(L(\mathcal{A})) \quad . \tag{SE.8.4}$$

(ii) *Show that the set* V_2 *is effectively computable.*

(iii) *Taking* $\mathcal{B} = \mathcal{S}_{\mathcal{G}}$ *, give the set of terminal states of* $\mathcal{B}_{\mathcal{R}2}$. *Verify, using the automaton* $\mathcal{S}$ *from Figure SE.8.1 as an example, that Equation (SE.8.4) does not necessarily hold for a deterministic automaton.*

(iv) Generalisation. *Give the effective construction for all integers* $k \geqslant 3$ *of an automaton* $\mathcal{A}_{\mathcal{R}k}$ *that recognises* $\mathcal{R}_k(K)$.



(h) The automaton $\mathcal{S}$ (i) The automaton $\mathcal{D}$

Figure SE.8.1: Two non-left-regular deterministic automata

Solution: (a) *First method.* For each $p \in Q$, $L_p = \{f \mid p \cdot f \in T\}$ and $L'_p = \{g \mid i \cdot g = p\}$ are recognisable languages, effectively computable from $\mathcal{A}$.

(i) If $f \in \mathcal{R}_2(K)$ then $i \cdot f\, f \in T$, and setting $p = i \cdot f$ it follows that $f \in L'_p \cap L_p$. Conversely, if $f \in L'_p \cap L_p$, then $p = i \cdot f$ and $p \cdot f \in T$ and hence $f^2 \in L$. Thus

$$\mathcal{R}_2(K) = \bigcup_{p \in Q} L'_p \cap L_p \quad ,$$

which shows the result.

(ii) For all p, q in Q, write $F_{p,q} = \{f \mid p \cdot f = q\}$, an effectively computable recognisable language. Without loss of generality, other than having to form a finite union at the end of the reasoning, and in order to simplify the notation, we assume that $T = \{t\}$ is a singleton set. Then $L'_p = F_{i,p}$ and $L_p = F_{p,t}$.

If $f \in \mathcal{R}_k(K)$, then $i \cdot f^k = t$, and there exists a $(k+1)$-tuple with states p_0, p_1, ..., p_k, with $p_0 = i$ and $p_k = t$, such that $p_{j-1} \cdot f = p_j$ for all $j \leqslant k$ and hence

$$f \in F_{p_0,p_1} \cap F_{p_1,p_2} \cap \cdots \cap F_{p_{k-1},p_k} = \bigcap_{j\in[k]} F_{p_{j-1},p_j} \quad .$$

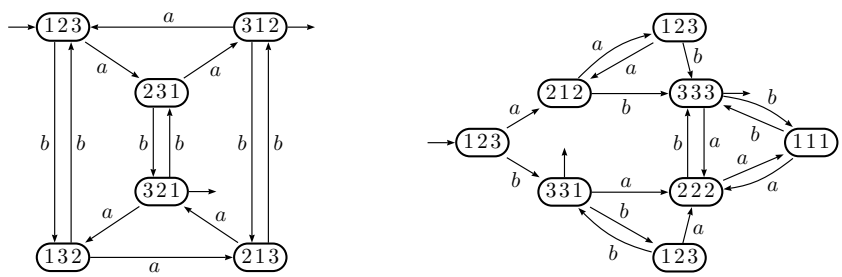
Conversely, for all $(k+1)$-tuples of states $p_0 = i$, p_1, ..., $p_k = t$, every word f in $\bigcap_{j\in[k]} F_{p_{j-1},p_j}$ is such that $f^k \in L$. Hence

$$\mathcal{R}_k(K) = \bigcup_{\substack{(p_0,p_1,\ldots,p_k)\in Q^{k+1}\\ p_0=i,\, p_k=t}} \left[\bigcap_{j\in[k]} F_{p_{j-1},p_j}\right]$$

which proves the result.[83]

(b) *Left regular automaton.*

(i) The automaton $\mathcal{S}$ shown in Figure SE.8.1 is not left regular since we have $1 \cdot a = 1 \cdot ba = 2$, hence $1 \cdot aa = 3 \neq 1 \cdot aba = 1$. The automaton $\mathcal{D}$ is not left regular since $1 \cdot a = 1 \cdot ba = 2$, hence $1 \cdot aa = 1 \neq 1 \cdot aba = 2$. The automata $\mathcal{S}_\mathcal{G}$ and $\mathcal{D}_\mathcal{G}$ are shown in Figure SE.8.2.



(a) The automaton $\mathcal{S}_\mathcal{G}$ (b) The automaton $\mathcal{D}_\mathcal{G}$

Figure SE.8.2: Regularisations of the automata in Figure SE.8.1

(ii) The automaton $\mathcal{A}_\mathcal{G}$ has a unique initial state ε; furthermore, since δ is a *map* from $Q{\times}A$ to Q, Equation (SE.8.2) indeed makes η a map from $\mathfrak{T}(Q) \times A$ to $\mathfrak{T}(Q)$: $\mathcal{A}_\mathcal{G}$ is complete deterministic.

We show, by induction on the length of f, that for all r in $\mathfrak{T}(Q)$, all f in A^* and all p in Q, we have

$$[(r,f)\eta]_p = (r_p, f)\delta \quad . \tag{SE.8.5}$$

Equation (SE.8.5) is true for $f = 1_{A^*}$ since, by definition,

$$\forall p \in Q \quad (r_p, 1_{A^*})\delta = r_p \quad \text{or} \quad (r, 1_{A^*})\eta = r \quad .$$

[83]This proof method anticipates the material of Section II.4 (*cf.* Exam. II.4.4, p. 292 and Exer. II.4.19, p. 293).

We then have, for all r in $\mathfrak{T}(Q)$, all f in A^* and all a in A

$$\begin{aligned}
& (r, f\,a)\eta = ((r,f)\eta, a)\eta && \text{by definition of } (r, f\,a)\eta \text{ and hence} \\
\forall p \in Q \qquad & [(r, f\,a)\eta]_p = [((r,f)\eta, a)\eta]_p && \\
& = ([\eta(r,f)]_p, a)\,\delta && \text{by (SE.8.2)} \\
& = ((r_p, f)\delta, a)\,\delta && \text{by induction hypothesis} \\
& = (r_p, f\,a)\delta && \text{by definition of } (r_p, f\,a)\delta \ .
\end{aligned}$$

In particular, we deduce from (SE.8.5) that

$$\begin{aligned}
L(\mathcal{A}_{\mathcal{G}}) &= \{f \in A^* \mid (\varepsilon, f)\eta \in U\} = \{f \in A^* \mid [(\varepsilon, f)\eta]_i \in T\} && \text{by (SE.8.3)} \\
&= \{f \in A^* \mid (i, f)\delta \in T\} = L(\mathcal{A}) && \text{by (SE.8.5):}
\end{aligned}$$

that is, $\mathcal{A}_{\mathcal{G}}$ is equivalent to $\mathcal{A}$. From the definition of η we obtain, in the standard way

$$\forall r \in \mathfrak{T}(Q)\,,\ \forall f, u \in A^* \qquad (r, u\,f)\eta = ((r,u)\eta, f)\eta \ .$$

If f and g are such that $(\varepsilon, f)\eta = (\varepsilon, g)\eta$, then by (SE.8.5) and since $\varepsilon_p = p$ for all p in Q:

$$\forall p \in Q \qquad (p, f)\delta = (p, g)\delta \ , \qquad \text{and hence} \qquad \forall r \in \mathfrak{T}(Q) \qquad (r, f)\eta = (r, g)\eta \ ,$$

from which we obtain

$$\forall u \in A^* \qquad (\varepsilon, u\,f)\eta = ((\varepsilon, u)\eta, f)\eta = ((\varepsilon, u)\eta, g)\eta = (\varepsilon, u\,g)\eta \ ;$$

that is, $\mathcal{A}_{\mathcal{G}}$ is left regular.

(c) *Second method for the computation of the roots of a language.*

(i) If $g \in \mathcal{R}_2(L(\mathcal{A}))$, then $g^2 \in L(\mathcal{A})$ and $i \cdot g \in V_2$ by the definition of V_2, hence $\mathcal{R}_2(L(\mathcal{A})) \subseteq L(\mathcal{A}_{\mathcal{R}2})$.

Conversely, if $g \in L(\mathcal{A}_{\mathcal{R}2})$, there exists f such that $i \cdot f\,f \in T$ and $i \cdot f = i \cdot g$. Since $\mathcal{A}$ is left regular, this last equality gives $i \cdot f\,f = i \cdot f\,g$ whence we obtain

$$i \cdot f\,f = i \cdot f\,g = (i \cdot f) \cdot g = (i \cdot g) \cdot g = i \cdot g\,g \in T \ ; \qquad \text{(SE.8.6)}$$

that is, that $g \in \mathcal{R}_2(L(\mathcal{A}))$, and hence $L(\mathcal{A}_{\mathcal{R}2}) \subseteq \mathcal{R}_2(L(\mathcal{A}))$.

(ii) We have just shown, with (SE.8.6), that if a word g sends i to a state of V_2, its square sends i to a state of T. To compute V_2 it is therefore enough to consider words of length less than $\|Q\|$, and even just a set of words that allows the graph of $\mathcal{A}$ to be 'covered'.

(iii) We have $1\,2\,3 \cdot a\,a = 3\,1\,2 \in T$ by which $1\,2\,3 \cdot a = 2\,3\,1 \in V_2$.

We verify further that the squares of *all* the other words f of length less than 3 (which is enough to reach all the states of $\mathcal{S}_{\mathcal{G}}$) send 1 2 3 to itself, except $(a\,a)^4$ which sends 1 2 3 to 2 3 1. Hence

$$V_2 = \{2\,3\,1\} \ .$$

In $\mathcal{S}$, we have $1 \cdot a^2 = 3 \in T$, hence $1 \cdot a = 2 \in V_2$. But $1 \cdot b\,a = 2$ and $1 \cdot b\,a\,b\,a = 1 \notin T$. The inclusion $L(\mathcal{S}_{\mathcal{R}2}) \subseteq \mathcal{R}_2(L(\mathcal{S}))$ does not hold (which shows the necessity of considering left regular automata).

(iv) With the above notation, and for any $k \geqslant 3$ set

$$V_k = \{p \in Q \mid \exists f \in A^* \qquad i \cdot f = p \quad \text{and} \quad i \cdot f^k \in T\} \ ,$$

and $\mathcal{A}_{\mathcal{R}k} = \langle\, Q, A, \delta, i, V_k \,\rangle$. The inclusion $\mathcal{R}_k(L) \subseteq L(\mathcal{A}_{\mathcal{R}k})$ is obvious as in (c.i).

Conversely, let g be in $L(\mathcal{A}_{\mathcal{R}k})$; that is, $i \cdot g = p$ and there exists f such that $i \cdot f = p$ and $i \cdot f^k \in T$. Since $\mathcal{A}$ is left regular, we have already seen, by (SE.8.6), that $i \cdot f = i \cdot g$ implies

that $i \cdot f^2 = i \cdot g^2$. By induction on n, we show that $i \cdot f = i \cdot g$ implies that $i \cdot f^n = i \cdot g^n$ for all n; in fact, we have

$$i \cdot f\, f^n = i \cdot f\, g^n = (i \cdot f) \cdot g^n = (i \cdot g) \cdot g^n = i \cdot g\, g^n \quad .$$

Hence in particular $i \cdot g^k \in T$, or $g \in \mathcal{R}_k(L)$ and $L(\mathcal{A}_{\mathcal{R}k}) \subseteq \mathcal{R}_k(L)$. As above, $\mathcal{R}_k(L) = L(\mathcal{A}_{\mathcal{R}k})$ is recognised by a finite automaton and V_k is computable by considering the words of length less than $\|Q\|$, which allows us to cover the graph of $\mathcal{A}$.[84]

Remark SE.8.1 The definition of 'left regularity' of an automaton $\mathcal{A}$ (equation SE.8.1) involves the initial state i of $\mathcal{A}$ and not just any state. From this point of view, the example $\mathcal{S}_{\mathcal{G}}$ is misleading because each state is a *permutation* of the states of $\mathcal{S}$; the example $\mathcal{D}_{\mathcal{G}}$, which contains $\mathcal{D}$ as a sub-automaton, shows clearly that the implication

$$\forall p \in Q\,,\ \forall f, g \in A^* \qquad p \cdot f = p \cdot g \quad \Longrightarrow \quad [\,\forall u \in A^* \quad p \cdot u\, f = p \cdot u\, g\,]$$

does not hold in general in the regularisation of a complete deterministic automaton. □

8.2 ***Iterating factors of a recognisable language.*** *We have to show:*

Proposition 8.1 *The set of iterating factors of a recognisable language is recognisable.*

We write $\mathcal{I}(K)$ for the set of iterating factors of a language K. We take $\mathcal{A} = \langle\, Q, A, E, I, T\,\rangle$ to be a finite automaton and $K = L(\mathcal{A})$.

(a) *For each state p in Q, let $\mathcal{A}_p = \langle\, Q, A, \delta, \{p\}, \{p\}\,\rangle$ be the automaton obtained by replacing in $\mathcal{A}$ both the set I of initial states and the set T of terminal states by p.*

Show that, for every useful state p, $L(\mathcal{A}_p)$ is included in $\mathcal{I}(L(\mathcal{A}))$.

(b) *Show that, for every iterating factor v of $L(\mathcal{A})$, there exists an integer k and a state p in Q such that v^k belongs to $L(\mathcal{A}_p)$.*

(c) *Show that $\mathcal{I}(K)$ is recognised by a finite automaton effectively computable from A.*

Solution: (a) If p is a useful state of $\mathcal{A}$, there exist u and w such that $i \xrightarrow{u} p \xrightarrow{w} t$ is a successful computation in $\mathcal{A}$. For all v in $L(\mathcal{A}_p)$, and every integer n,

$$i \xrightarrow{u} p \xrightarrow{v} p \xrightarrow{w} t \qquad \text{and} \qquad i \xrightarrow{u} \overbrace{p \xrightarrow{v} p \xrightarrow{v} p \cdots p \xrightarrow{v} p}^{n \text{ times}} \xrightarrow{w} t$$

are successful computations in $\mathcal{A}$: v is an iterating factor of $L(\mathcal{A})$.

(b) Set $N = \|Q\|$. If v is an iterating factor of $L(\mathcal{A})$, there exist u and w in A^* such that $f = u\, v^N\, w$ is in $L(\mathcal{A})$. We use the Block Star Lemma – the blocks are the v's. A computation which accepts f can be written

cf. Lem. 1.14, p. 71

$$i \xrightarrow{u} q_0 \xrightarrow{v} q_1 \xrightarrow{v} q_2 \cdots q_{N-1} \xrightarrow{v} q_N \xrightarrow{w} t \quad .$$

The q_i cannot all be distinct and $q_j = q_l = p$ with $j < l$. Hence $v^k \in L(\mathcal{A}_p)$ with $k = l - j$.

(c) Let $\mathcal{A} = \langle\, Q, A, \delta, i, T\,\rangle$ be a complete deterministic automaton which recognises K. From the two previous questions it follows that

$$\left[\bigcup_{p \in Q} L(\mathcal{A}_p)\right] \subseteq \mathcal{I}(K) \subseteq \bigcup_{p \in Q} \left(\bigcup_{1 \leqslant k \leqslant N} \mathcal{R}_k(L(\mathcal{A}_p))\right) \quad .$$

[84] This proof method anticipates the definition of the transition monoid of the automaton $\mathcal{A}$ ($\mathcal{A}_{\mathcal{G}}$ is the Cayley graph of this monoid) and the fact that a language recognised by $\mathcal{A}$ is 'recognised' by its transition monoid (*cf.* Section II.2.2, p. 238 and Prop. II.2.15, p. 252).

More precisely, if $\mathcal{I}_p = \{v \mid \exists u, w \in A^* \quad i \cdot u = p \quad \text{and} \quad u\,v^*\,w \subseteq K\}$, we have $\mathcal{I}(K) = \bigcup_{p \in Q} \mathcal{I}_p$ and, for each useful state p of $\mathcal{A}$,

$$L(\mathcal{A}_p) \subseteq \mathcal{I}_p \subseteq \bigcup_{1 \leqslant k \leqslant N} \mathcal{R}_k(L(\mathcal{A}_p)) \ .$$

This last inclusion *cannot be reversed* but does inspire a line of reasoning based on the construction described in the previous exercise.

Let $\mathcal{A}_{\mathcal{G}} = \langle\, \mathfrak{T}(Q), A, \eta, \varepsilon, U \,\rangle$ be the left regular automaton constructed from $\mathcal{A}$ as in 8.1 (b) above. Setting $U_p = \{r \in \mathfrak{T}(Q) \mid r_p = p\}$ and $\mathcal{A}_{\mathcal{G},p} = \langle\, \mathfrak{T}(Q), A, \eta, \varepsilon, U_p \,\rangle$, we obtain $L(\mathcal{A}_p) = L(\mathcal{A}_{\mathcal{G},p})$.

For each p in Q and each k in $[N]$, let

$$\mathcal{J}_{p,k} = \{v \mid \exists u, w \in A^* \quad i \cdot u = p\,, \ \ p \cdot v^k = p \quad \text{and} \quad u\,v^*\,w \subseteq K\} = \mathcal{I}_p \cap \mathcal{R}_k(L(\mathcal{A}_p)) \ ,$$

and let $W_{p,k}$ be the subset of $\mathfrak{T}(Q)$ defined, in imitation of the V_k above, by

$$W_{p,k} = \{r \in \mathfrak{T}(Q) \mid \exists v \in \mathcal{J}_{p,k} \quad (\varepsilon, v)\eta = r\} \ .$$

Set $\mathcal{A}_{\mathcal{G},p,k} = \langle\, \mathfrak{T}(Q), A, \eta, \varepsilon, W_{p,k} \,\rangle$; we will show that $L(\mathcal{A}_{\mathcal{G},p,k}) = \mathcal{J}_{p,k}$. By definition of $W_{p,k}$, we have $\mathcal{J}_{p,k} \subseteq L(\mathcal{A}_{\mathcal{G},p,k})$.

Conversely, let z be in $L(\mathcal{A}_{\mathcal{G},p,k})$; there exist therefore u, v and w in A^* such that:

$$i \cdot u = p \ , \qquad p \cdot v^k = p \ , \qquad u\,v^*\,w \subseteq K \qquad \text{and} \qquad (\varepsilon, v)\eta = (\varepsilon, z)\eta \ .$$

Since $\mathcal{A}_{\mathcal{G},p,k}$ is left regular, we have

$$(\varepsilon, v^k)\eta = (\varepsilon, z^k)\eta \qquad \text{and} \qquad \forall j \in \mathbb{N} \qquad (\varepsilon, u\,v^j)\eta = (\varepsilon, u\,z^j)\eta \ ,$$

and hence $p \cdot z^k = p \cdot v^k$ and $i \cdot u\,z^j = i \cdot u\,v^j$, from which $u\,z^j\,w \in T$ and $z \in \mathcal{J}_{p,k}$.

A word v is in $\mathcal{J}_{p,k}$ if both $p \cdot v^k = p$ and, writing $q_j = p \cdot v^j$ for all j in $[k]$, there exists a word w such that $q_j \cdot w \in T$; that is, if $[(\varepsilon, w)\eta]_{q_j} \in T$. The two conditions are tested by considering that the set of words which covers the graph of $\mathcal{A}_{\mathcal{G}}$ and $W_{p,k}$ is effectively computable.[85]

8.3 ***Slices of rationals.*** *For each word g in A^*, the* first half *of g, written $\mathsf{FH}(g)$, is the prefix f of g whose length is half that of g (which assumes that g is of even length). By additivity, the first half of a language L, written $\mathsf{FH}(L)$, is the set of first halves of words of L. In the same way, we define the first and second* thirds *of L: $\mathsf{FT}(L)$ and $\mathsf{ST}(L)$, which are written*

$$\mathsf{FT}(L) = \{f \in A^* \mid \exists h \in A^* \qquad |h| = 2\,|f| \quad f\,h \in L\} \ ,$$
$$\mathsf{ST}(L) = \{g \in A^* \mid \exists f, h \in A^* \qquad |f| = |g| = |h| \quad f\,g\,h \in L\} \ .$$

(a) *Show that if L is recognisable, $\mathsf{FH}(L)$ is too, effectively.*

(b) *Do the same for $\mathsf{FT}(L)$.* (c) *Do the same for $\mathsf{ST}(L)$.*

Solution: Let $\mathcal{A} = \langle\, Q, A, \delta, i, T \,\rangle$ be a complete deterministic automaton which recognises L.

(a) We consider the automaton[86] $\mathcal{A}_{\mathcal{B}} = \langle\, \mathfrak{P}(Q), A, \beta, \{T\}, U \,\rangle$ where, for all p in Q and all P included in Q,

$$(p, a)\beta = \{q \in Q \mid \exists x \in A \quad (q, x)\delta = p\,\} \quad \text{and} \quad (P, a)\beta = \bigcup_{p \in P} (p, a)\beta \ ,$$

[85] As with the previous one, this proof anticipates the use of the transition monoid of the automaton $\mathcal{A}$ (*cf.* Section II.2.5, p. 251 and Exer. II.2.26, p. 254). Also like the previous one, it has benefited from careful reading by my students Laure Danthony and Laurent Fousse. The example of the automaton $\mathcal{D}$ which leads to Remark SE.8.1 is due to them.

[86] $\mathcal{B}$ stands for *backward*.

and where the final state is unspecified. We then consider the product of $\mathcal{A}$ with $\mathcal{A}_{\mathcal{B}}$:

$$\mathcal{A}_{\mathcal{H}} = \langle\, Q \times \mathfrak{P}(Q), A, \delta \times \beta, (i,T), V \,\rangle \qquad \text{where, obviously,}$$
$$\big((p,P),a\big)\delta\times\beta = \big((p,a)\delta, (P,a)\beta\big) \quad \text{and} \quad V = \{(p,P) \in Q \times \mathfrak{P}(Q) \mid p \in P\}$$

We verify easily that $\mathsf{FH}(L) = L(\mathcal{A}_{\mathcal{H}})$.

(b) We replace the transition function β above by β':

$$(p,a)\beta' = \{q \in Q \mid \exists x, y \in A \qquad (q, x\,y)\delta = p\}$$

in $\mathcal{A}_{\mathcal{H}}$: $\mathcal{A}_{\mathcal{H}'} = \langle\, Q \times \mathfrak{P}(Q), A, \delta \times \beta', (i,T), V \,\rangle$ and $\mathsf{FT}(L) = L(\mathcal{A}_{\mathcal{H}'})$.

(c) The construction is more complex, and needs the regularisation of $\mathcal{A}$, defined above in Exercise 8.1 (b): $\mathcal{A}_{\mathcal{G}} = \langle\, \mathfrak{T}(Q), A, \eta, \varepsilon, U \,\rangle$. We also define the automaton[87] $\mathcal{A}_{\mathcal{F}} = \langle\, \mathfrak{P}(Q), A, \gamma, \{i\}, U \,\rangle$ where, for all p in Q and all P included in Q,

$$(p,a)\gamma = \{q \in Q \mid \exists x \in A \qquad (p,x)\delta = q\} \quad \text{and} \quad (P,a)\gamma = \bigcup_{p \in P} (p,a)\gamma \ .$$

We then make the product $\mathcal{A}_{\mathcal{F}} \times \mathcal{A}_{\mathcal{G}} \times \mathcal{A}_{\mathcal{B}}$ to give the automaton $\mathcal{A}_{\mathcal{T}}$:

$$\mathcal{A}_{\mathcal{T}} = \langle\, \mathfrak{P}(Q) \times \mathfrak{T}(Q) \times \mathfrak{P}(Q), A, \gamma \times \eta \times \beta, (\{i\}, \varepsilon, T), W \,\rangle \ ,$$

where obviously $\big((P,r,R),a\big)\gamma\times\eta\times\beta = \big((P,a)\gamma, (r,a)\eta, (R,a)\beta\big)$ and where

$$W = \{(P,r,R) \in \mathfrak{P}(Q) \times \mathfrak{T}(Q) \times \mathfrak{P}(Q) \mid \exists p \in P \quad r_p \in R\} \ .$$

We verify easily that $\mathsf{ST}(L) = L(\mathcal{A}_{\mathcal{T}})$.[88]

8.4 **Thin languages.** *Prove:*[89]

Proposition 8.2 *A language is thin if and only if it is recognisable and of bounded growth.*

(a) *Show that if a language L is thin, it is of bounded growth.*

(b) *In the following, L is a language recognised by an automaton $\mathcal{A}$ which we assume to be deterministic and trim.*

Let $\mathcal{C}$ be a strongly connected component of $\mathcal{A}$. Show that if $\mathcal{C}$ is not reduced to a single loop, L is not of bounded growth.

(c) *Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two strongly connected components of $\mathcal{A}$. Show that if there exists a path in $\mathcal{A}$ that leads from a state of $\mathcal{C}_1$ to a state of $\mathcal{C}_2$, then L is not of bounded growth.*

(d) *Deduce that if L is a language of bounded growth, it is thin.*

Solution: (a) Every thin language L can be written in the form

$$L = \bigcup_{i \in I} u_i\, v_i^*\, w_i \ ,$$

[87] $\mathcal{F}$ stands for *forward*.

[88] Seiferas and McNaughton [229] report that the stability of recognisability by 'cutting in two' [$L \in \operatorname{Rec} A^* \Rightarrow \mathsf{FH}(L) \in \operatorname{Rec} A^*$] was conjectured by Ginsburg in 1962. The results above, which generalise cutting in two to cutting in any proportion, and also the method of proof, are due to Stearns and Hartmanis [241]. A more algebraic interpretation is given in [187].

Seiferas and McNaughton generalise the problem further and characterise the relations $\alpha\colon \mathbb{N} \to \mathbb{N}$ such that the operation $\mathsf{F}_\alpha(L) = \{f \mid \exists g \quad |g| \in (|f|)\alpha \text{ and } f\,g \in L\}$ preserves recognisability: F_α *preserves recognisability if and only if* α^{-1} *preserves recognisability,* this last property making sense since $\mathbb{N}$ can be seen as a free monoid with a generator.

[89] For the definitions, see the statement of the exercice on p. 177.

with I finite. For every n, every ray $u_i\, v_i^*\, w_i$ contains at most one word of length n. Hence $l(n) \leqslant \|I\|$.

(b) If $\mathcal{C}$ is not reduced to a single loop, there exists in $\mathcal{C}$ a state p and two distinct transitions $p \xrightarrow{a} q$ and $p \xrightarrow{b} r$. As $\mathcal{A}$ is deterministic, $a \neq b$. As $\mathcal{C}$ is strongly connected, there exists a computation from q to p, with label t_1, and a computation from r to p, with label t_2. Set $y_1 = a\, t_1$ and $y_2 = b\, t_2$, then $v_1 = y_1^{k_1}$ and $v_2 = y_2^{k_2}$, where k_1 and k_2 are chosen such that $|v_1| = |v_2| = \mathrm{ppcm}(|y_1|, |y_2|) = m$. All the words of $\{v_1, v_2\}^n$ are *distinct* and of the same length $n\, m$. There are 2^n.

As $\mathcal{A}$ is trim, there exist u and w in A^*, and t in T, such that $i \xrightarrow{u} p$ and $p \xrightarrow{w} t$ are computations in $\mathcal{A}$. The set $u\,\{v_1, v_2\}^n\, w$ is included in L, and formed from 2^n distinct words which all have the same length: L is not of bounded growth.

(c) Let p_1 be in $\mathcal{C}_1$, p_2 be in $\mathcal{C}_2$, and x be in A^* such that $p_1 \xrightarrow{x} p_2$ is a computation of $\mathcal{A}$. Since $\mathcal{C}_1$ and $\mathcal{C}_2$ are distinct, $x = a\, x'$ is not the empty word, and we can choose p_1 and x such that the state q given by $p_1 \xrightarrow{a} q$ does not belong to $\mathcal{C}_1$.

Let y_1 and y_2 be such that $p_1 \xrightarrow{y_1} p_1$ and $p_2 \xrightarrow{y_2} p_2$ respectively, and let $v_1 = y_1^{k_1}$ and $v_2 = y_2^{k_2}$, where k_1 and k_2 are chosen such that $|v_1| = |v_2| = \mathrm{ppcm}(|y_1|, |y_2|)$. As $\mathcal{A}$ is deterministic, the first letter of y_1, and hence of v_1, is different from the first letter a of x. Thus, for each n, the words $(v_1)^k\, x\, (v_2)^{n-k}$, for $0 \leqslant k \leqslant n$, are all distinct and have the same length.

As $\mathcal{A}$ is trim, there exist u and w in A^*, and t in T, such that $i \xrightarrow{u} p_1$ and $p_2 \xrightarrow{w} t$ are computations in $\mathcal{A}$. The set $\{u\,(v_1)^k\, x\,(v_2)^{n-k}\, w \mid 0 \leqslant k \leqslant n\}$ is included in L, and formed from $n+1$ distinct words which all have the same length: L is not of bounded growth.

(d) According to the two previous questions, a successful computation c of $\mathcal{A}$ has the form

$$i \xrightarrow{u} p \xrightarrow{x} q \xrightarrow{w'} t \quad ,$$

where $t \in T$, and $i \xrightarrow{u} p$ and $q \xrightarrow{w'} t$ contain no internal transition to a strongly connected component, and where $p \xrightarrow{x} q$ is entirely contained in the loop and hence of the form $p \xrightarrow{v^k} p \xrightarrow{x'} q$. The label of c belongs to the ray $u\, v^*\, x'\, w'$ which itself is included in $L(\mathcal{A})$.

The possible choices for p, hence for u, and for q and t, hence for v', x' and w' are finite: $L(\mathcal{A})$ is a finite union of rays.

8.5 ***Number of words of a given length in a recognisable language.***[90]

(a) Examples.

(i) *Characterise the language recognised by the automaton $\mathcal{B}$ of Figure SE.8.3 (a). Deduce the number of words of length n recognised by $\mathcal{B}$, for every integer n.*

(ii) *Characterise the language F recognised by the automaton $\mathcal{F}$ of Figure SE.8.3 (b). For every integer n, we write $f(n)$ for the number of words of length n recognised by $\mathcal{F}$. Express $f(n+2)$ as a function of $f(n+1)$ and $f(n)$. Compute $f(8)$. Express $f(n)$ as a function of n.*

(b) Generalisation. *Let $\mathcal{A} = \langle Q, A, E, i, T \rangle$ be a deterministic automaton and $L = L(\mathcal{A})$ the language that it recognises. For all $p \in Q$, write*

$$L_p = \{f \in A^* \mid \exists t \in T \quad p \xrightarrow{f} t\} \quad .$$

[90] The contents of this exercise and its solution are taken from one of the seminal articles of automata theory and languages by Chomsky and Miller [60]. The function l is there called the *structure function* of the language L.

Write $l(n)$ for the number of words in L of length n and $l_p(n)$ for the number of words in L_p of length n, for all $n \in \mathbb{N}$. [In particular, $L = L_i$ and $l(n) = l_i(n)$.]

(i) *Show that, for all positive n and for all p in Q, $l_p(n)$ is a linear combination with non-negative integer coefficients of the $l_q(n-1)$, for q in Q.*

(ii) *Deduce that the function $l(n)$ satisfies a linear recurrence relation of the form*

$$\forall n \in \mathbb{N} \qquad l(n+k) = z_1\, l(n+k-1) + z_2\, l(n+k-2) + \cdots + z_k\, l(n) \ ,$$

where k is the number of states of $\mathcal{A}$ and the z_j are integers.

(iii) *Explain how to compute the initial conditions: that is, $l(0)$, $l(1)$, ..., $l(k-1)$.*

(c) Application. *Apply the method of the previous question to the computation of the number of words of length n recognised by the automaton $\mathcal{B}$ of question (a).*

(a) automaton $\mathcal{B}$ (b) automaton $\mathcal{F}$

Figure SE.8.3: Two automata

Solution:

(a) (i) A word in A^* is recognised by $\mathcal{B}$ if and only if it contains a b. The *only* word in A^n which is *not* recognised by $\mathcal{B}$ is a^n. The number of words of length n recognised by $\mathcal{B}$ is therefore $2^n - 1$.

(ii) A word u in A^* belongs to F if and only if every b in u is followed by (at least) one a. Let $u \in F$, $|u| = n+2$. Then either

$$u = a\,v \qquad \text{and then} \qquad v \in F, \quad |v| = n+1 \ ,$$

or

$$u = b\,w \qquad \text{and then} \qquad w = a\,w', \quad w' \in F, \quad |w'| = n$$

and hence

$$f(n+2) = f(n+1) + f(n) \ . \tag{SE.8.7}$$

We have $f(0) = 1$ since $1_{A^*} \in F$ and $f(1) = 1$ since $a \in F$, hence

n	0	1	2	3	4	5	6	7	8
$f(n)$	1	1	2	3	5	8	13	21	34

[$f(n)$ is the nth Fibonacci number].

Set $\alpha = \frac{1+\sqrt{5}}{2}$ and $\beta = \frac{1-\sqrt{5}}{2}$, the roots of $X^2 - X - 1$, the characteristic polynomial of the recurrence relation (SE.8.7). Standard theory ensures that there exist x and y, real numbers, such that for all $n \in \mathbb{N}$,

$$f(n) = x\,\alpha^n + y\,\beta^n \ .$$

From

$$f(0) = 1 = x + y \qquad \text{and} \qquad f(1) = 1 = x\,\frac{1+\sqrt{5}}{2} + y\,\frac{1-\sqrt{5}}{2} \ ,$$

we obtain:

$$x = \frac{1}{2}\left(1 + \frac{1}{\sqrt{5}}\right) \qquad \text{and} \qquad y = \frac{1}{2}\left(1 - \frac{1}{\sqrt{5}}\right) \ ,$$

whence

$$f(n) = \frac{1}{2}\left[\frac{1+\sqrt{5}}{\sqrt{5}}\,\alpha^n - \frac{1-\sqrt{5}}{\sqrt{5}}\,\beta^n\right] = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^{n+1} - \left(\frac{1-\sqrt{5}}{2}\right)^{n+1}\right] \ .$$

(b) (i) Let $p \in Q$ and $u = a\,v$ be a word in L_p (where $a \in A$). There exists $q \in Q$ such that $p \xrightarrow{a} q$ is in E and v is a word in L_q. Thus,

$$L_p = \bigcup_{\substack{a \in A,\, q \in Q \\ p \xrightarrow{a} q \in E}} a\,L_q \ . \tag{SE.8.8}$$

The word u and state p being given, the decomposition $u = a\,v$ for $a \in A$ is unique, and the state q, such that $p \xrightarrow{a} q \in E$, is unique *since $\mathcal{A}$ is deterministic*, and if u is of length $n+1$, v is of length n. The equation (SE.8.8) can thus be rewritten

$$l_p(n+1) = \sum_{q \in Q} m_{p,q}\, l_q(n) \ , \tag{SE.8.9}$$

where the $m_{p,q}$ are the integers defined by

$$m_{p,q} = \|\{a \in A \mid p \xrightarrow{a} q \in E\}\| \ .$$

Write $\vec{l}(n)$ for the vector of dimension Q whose pth entry is $l_p(n)$ and M for the square matrix of dimension $Q \times Q$ whose entry at p, q is exactly the integer $m_{p,q}$. Every word in L_p is found exactly once in the right-hand side of (SE.8.9) which can thus be rewritten

$$\forall n \in \mathbb{N} \qquad \vec{l}(n+1) = M\,\vec{l}(n) \ . \tag{SE.8.10}$$

(ii) Let $P(\lambda) = \det(M - \lambda\,I) = (-1)^k \left[\lambda^k - z_1\,\lambda^{k-1} - \cdots - z_{k-1}\,\lambda - z_k\right]$ be the characteristic polynomial of the matrix M, where the z_i are integers. The matrix M satisfies the equation $P(M) = 0$ by the Cayley–Hamilton Theorem,[91] so

$$M^k - z_1\,M^{k-1} - \cdots - z_{k-1}\,M - z_k\,I = 0 \ .$$

If we apply this operator, identically equal to zero, to the vector $\vec{l}(n)$, it follows, since $M^j\,\vec{l}(n) = \vec{l}(n+j)$, that

$$\vec{l}(n+k) - z_1\,\vec{l}(n+k-1) - \cdots - z_{k-1}\,\vec{l}(n+1) - z_k\,\vec{l}(n) = 0 \ .$$

Taking just the ith entry ($l_i(n) = l(n)$), and taking the terms other than $l(n+k)$ to the other side, we obtain exactly the recurrence relation we want.

(iii) We have $l_p(0) = 1$ if $p \in T$ and $l_p(0) = 0$ otherwise. We then compute the $\vec{l}(j)$, for $1 \leqslant j \leqslant k-1$, by $\vec{l}(j) = M^j\,\vec{l}(0)$, from which we obtain the $l(j) = l_i(j)$.

(c) In the case of the automaton $\mathcal{B}$, the system (SE.8.9) takes the form

$$\forall n \in \mathbb{N} \qquad \begin{aligned} l_p(n+1) &= l_p(n) + l_q(n) \\ l_q(n+1) &= 2\,l_q(n) \end{aligned}$$

It corresponds to the matrix $M = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix}$ whose characteristic equation is $\lambda^2 - 3\,\lambda + 2 = 0$ and we verify that we indeed have the recurrence

$$\forall n \in \mathbb{N} \qquad 2^{n+2} - 1 = 3\left(2^{n+1} - 1\right) - 2\,(2^n - 1) \ .$$

[91] Another classic of linear algebra.

NOTES & REFERENCES

The majority of the contents of this chapter is standard material and is covered in many books on automata theory, of course, but also in texts on compilation, algorithms and in more general computer science works. Here, then, is a partial, arbitrary and chronological list, limited to books specifically devoted to automata, augmented by the relevant chapters of two recent handbooks.

M. Gross and A. Lentin, *Introduction aux grammaires formelles*. Gauthier-Villars, 1967. [109]

J. E. Hopcroft and J. D. Ullman, *Formal Languages and their Relation to Automata*. Addison-Wesley, 1969. [121]

S. Eilenberg, *Automata, Languages and Machines*, volume A. Academic Press, 1974. [80]

M. Harrison, *Introduction to Formal Language Theory*. Addison-Wesley, 1974. [112]

J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, 1979. [122]

D. Perrin, Finite automata. In *Handbook of Theoretical Computer Science*, Vol. B. Elsevier, 1990. [183]

J. M. Howie, *Automata and Languages*. Oxford, 1991. [123]

D. Kozen, *Automata and Computability*. Springer, 1997. [138]

S. Yu. Regular languages. In *Handbook of Formal Languages*, vol. 1. Elsevier, 1997. [251]

J. E. Hopcroft, R. Motwani and J. D. Ullman, *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley, 2000. (Third Ed. 2007). [120]

1. What is an 'automaton'?

To put forward the graph point of view is to follow the work of Eilenberg, whose book is the only one in the list above to proceed thus. That said, everything in this section is standard material. It is worth mentioning that Proposition 1.6 appeared in [100], the principle of star lemmas in the seminal paper [15], and the characterisation of ambiguous automata (Lemma 1.16) in [29, Prop. IV.1.6].

2. Rational languages

Theorem 2.1, the basis of the theory of finite automata, is due to S. Kleene [133], hence its name. The BMC algorithm for the calculation of a rational expression of the language recognised by a finite automaton was published in [43]; it was called the 'state elimination method' and presented in its present more accessible form in [250, 251]. The MNY algorithm is the method given in all other books on the subject (except for [66]); the original reference is [169].

3. The functional point of view

This is the most widely adopted point of view. The construction of the minimal automaton by successive refinement of partitions is due to Moore and leads naturally to an $O(n^2)$ algorithm. A more involved algorithm due to J. Hopcroft [119] performs the same construction with a complexity of $O(n \log n)$. This algorithm, whose proof was simplified by Gries [108] is also given in [4] and, in a highly detailed manner, in [23].

Proposition 3.13 and the minimisation method for automata which is derived from it are due to J. Brzozowski [41].

Theorem 3.3 and its proof are due to Ehrenfeucht, Parikh and Rozenberg [79]. A demonstration of Ramsey's Theorem (Theorem 3.4, originally proved in [191]) can be found in many treatises on combinatorics and graph theory, for example [76].

4. Rational expressions

Our presentation of standard identities follows the one that Krob [139] gave of Conway's system [66]. Theorem 4.2 can be found in [139] and [66], with a less elementary proof.

The problem of the axiomatisation of rational expressions has been studied since the beginnings of automata theory. Redko [194] gave a proof that a complete system for rational expressions over a single-letter alphabet is necessarily infinite. Contrary to what one might rashly conclude, one cannot conclude that the same holds for any alphabet, and Conway showed that a complete system of axioms must contain an infinity of identities for an alphabet containing at least two letters [66, Th. 13.9], one of the main results of his monograph. Conway's method is to associate a rational identity with each finite monoid, in particular with each finite group, and to show that the identities corresponding to *simple groups* are independent from one another. Krob showed that, as conjectured by Conway, this set of identities is complete [139].

By allowing as well as substitution a rule of inference which rehashes Arden's Lemma (and which is therefore valid for equivalent expressions), Salomaa constructed a *finite axiomatisation* of rational expressions in [216].

The comparison between the results of the BMC and MNY algorithms seems to be original. The notion of derivative of an expression and the definition of the derivative automaton via Theorem 4.3 are due to J. Brzozowski [42].

5. From expressions to automata

Constructing an automaton from a rational expression and finding a factor described by an expression in a text are two facets of a problem which has spawned much work and an abundant literature. This section is merely an invitation to this rich domain. Compared to more generalist books on algorithms that avoid some of these questions, the chapter by A. Aho in the *Handbook of Theoretical Computer Science* [3] and two recent works [68, 175] deal with them comprehensively, and give the original references.

The automaton that we call the 'standard automaton of an expression' is usually attributed to Glushkov [104] and called 'Glushkov's automaton' or the 'position automaton'. In [169], McNaughton and Yamada compute its determinisation directly; Berry and Sethi describe it with the aid of expression derivation [26, 30]; the effective computation of this automaton is the subject of many works ([38] *etc.*). The mode of construction given here, which will be continued in Chapter III with that of '$\mathbb{K}$-standard' automata is adapted from [49]. *cf. Def. 2.8, p. 414*

The concept of 'derived terms' is a remarkable idea due to V. Antimirov [8]; Antimirov calls them 'partial derivatives' of E with the explanation that they are 'parts of derivatives' of E. This seemed to me an unhappy nomenclature, since the derivative expressions of E are already 'partial' inasmuch as they are the result of a derivation with respect to *one* of the letters, and because 'partial derivatives' further overloads an established mathematical term.

6. Star height

The definition of loop complexity and Theorem 6.1 were adapted from Eggan's seminal article [77] where the notion of star height appeared for the first time. The proof given for Theorem 6.1 is that of [155]. Example 6.3 is taken from [168] (*cf.* also Section II.8).

Theorem 6.2 and its proof are due to Dejean and Schützenberger, and taken pretty much *verbatim* from the original paper [73]. The same proof is also reproduced in [217]. In Sec-

cf. Cor. II.8.1, p. 342

tion II.8 we will give McNaughton's more conceptual proof for the same result [168].

The proof that star height is computable in general was made by K. Hashiguchi [117], by a formidable extension of the methods which had already enabled him to show that it is decidable whether a rational language is of height 1 or strictly greater than 1 [115, 116].[92]

7. A field of automata

In this section we did not seek to give the history of the theory of automata, and hence to cover the first models of finite automata, in particular the neural networks of McCulloch and Pitts [165], which were reconsidered and analysed in Kleene's seminal article [133]. On this subject one can consult Dominique Perrin's excellent synthesis [184].

The model of automata which we called 'Rabin–Scott' was defined by these authors in [190]. Theorem 7.1 is due to Shepherdson [233]; the proof we give follows [219].

[92] *Note added in translation:* We already quoted D. Kirsten's work [132] which gives a more accessible proof for this result (p. 157).

Chapter II

The power of algebra

Automata over an arbitrary monoid

We look afresh at the notion of an automaton, a labelled graph, supposing that the labels are no longer just the letters of an alphabet, generator of a free monoid, but the elements of an arbitrary monoid.

We aim to show to what extent the results and properties that we proved in the previous chapter rely on the 'freedom' of the label monoid, and how much holds in general; that is, how much is due to the notion of an automaton over a monoid. Thus the distinction between a *rational set* and a *recognisable set* will appear, a distinction without a difference in the case of free monoids.

However, as in the previous chapter, automata are still used to define *subsets*: an element m of a monoid M is accepted, or not, by an automaton $\mathcal{A}$; that is, belongs or does not belong to the subset of M defined by $\mathcal{A}$. In other words, the underlying algebraic structure is $\mathfrak{P}(M)$, the semiring of subsets of M, in contrast to the next chapter where we will not only ask whether an element is accepted or not by an automaton but will *count* the number of ways in which the element can be accepted.

The rational and recognisable subsets of a monoid are defined and studied in Sections 1 and 2 respectively. In the third section, we put in place the notion of a *morphism* of automata which will enable us to treat not only the subsets of a monoid but also automata over a monoid with all the power and clarity of algebraic methods. We describe in the fourth section a remarkable construction, due to J. H. Conway, that associates with each recognisable subset of a monoid an automaton which is in some sense *universally minimal*. The fifth section presents another algebraic structure: the *well quasi-order* (a generalisation of *partial well-order*), widely used in all aspects of theoretical computer science and which we will use on several occasions, in this chapter and those that follow. It will allow us to prove a beautiful result of Ehrenfeucht, Haussler and Rozenberg on a family of rewriting systems.

The bases are set. They allow us to use the notions of rational and recognisable subsets and their properties in the rest of the work, in a variety of situations and contexts which speaks to their fundamental character. The objective is twofold. The first part is to give us powerful tools to prove the properties of rational languages – which are also recognisable. Sections 2 and 8 give several illustrations. On this subject, I must express some regrets, and give my apologies to the reader. As I said in the preface, I have not dealt at all with that part of the theory that deals with the characterisation and *classification* of rational languages by the properties of their

syntactic monoids, which is indeed a domain in which the power of algebra shines. This field, called 'variety theory', is already covered in several works; I have given some references in the bibliographic notes.

The second part of the objective is to be able to study rational subsets in multiplicative structures other than the free monoid, and thereby to extend the power of finite automata. Along these lines, the supplementary sections discuss rational subsets in the free group, which among other things gives an introduction to the mathematical study of pushdown automata, and rational subsets in free commutative monoids, which appear in many contexts, in particular in connection with verification and so-called 'timed automata'. The study of rational subsets in free products, which extends that of rational subsets in the free group, is postponed for reasons of technique – proof technique – to the next chapter.

Contents of the chapter

1. Automata and rational sets
2. Actions and recognisable sets
3. Morphisms and coverings
4. The universal automaton
5. The importance of being well ordered

Deeper & broader

6. Rationals in the free group
7. Rationals in commutative monoids
8. Star height of group languages

Solutions to the exercises

Notes & references

1 Automata and rational sets

The aim of this section is to generalise the notion of an automaton over an alphabet A to that of an automaton over a monoid M, and, as in the previous chapter, to characterise the subsets defined by these automata. Throughout this chapter, M and N denote monoids, *a priori* arbitrary, and $\alpha\colon M \to N$ a morphism.

1.1 Automata over a monoid

Bis repetita placent.

Cicero

The definition of automata (finite or not) *over an arbitrary monoid* is copied from that of automata *over an alphabet*. An automaton $\mathcal{A}$ over M is a graph labelled with the elements of M, which has two sets of distinguished states,

$$\mathcal{A} = \langle\, Q, M, E, I, T \,\rangle \ ,$$

where, as we expect, Q is the set of states, M the *label monoid* (or *input monoid*), I and T are the sets of initial and final states, and E is the set of *transitions*, or labelled arcs. Each transition is a triple (*source*, *label*, *destination*); that is, E is canonically equipped with three maps: ι, ε and τ:

$$\iota\colon E \longrightarrow Q \ , \qquad \tau\colon E \longrightarrow Q \ , \qquad \text{and} \qquad \varepsilon\colon E \longrightarrow M \ .$$

The state $e\iota$ is the source of the transition e, $e\tau$ is its destination and $e\varepsilon$ its label. *A priori*, two distinct transitions may have the same source, same destination and same label, but, bending the language and to simplify the notation,[1] we will usually identify a transition e with the triple (p, m, q), which we will also write $p \xrightarrow{m} q$ (or $p \xrightarrow[\mathcal{A}]{m} q$ if necessary), where $p = e\iota$, $q = e\tau$ and $m = e\varepsilon$, and hence E with a subset of $Q \times M \times Q$.

A *computation* is a path in the graph $\mathcal{A}$; it is *successful* if its source is in I and its destination is in T. The label, written $|c|$, of a computation c in $\mathcal{A}$,

$$c \ := \ p_0 \xrightarrow{m_1} p_1 \xrightarrow{m_2} p_2 \cdots p_{n-1} \xrightarrow{m_n} p_n \ ,$$

is the product of the successive transition labels of c: $|c| = m_1 m_2 \cdots m_n$ and we also write $c \ := \ p_0 \xrightarrow[\mathcal{A}]{m_1 m_2 \cdots m_n} p_n$. An element of M is *accepted* by $\mathcal{A}$ – we avoid, for reasons that will become obvious, saying 'recognised' by $\mathcal{A}$ – if it is the *label of a successful computation*.

Definition 1.1

(i) The *behaviour* of an automaton $\mathcal{A}$, written $|\mathcal{A}|$, is the set of elements of M accepted by $\mathcal{A}$:

$$|\mathcal{A}| = \{m \in M \mid \exists p \in I \,,\ \exists q \in T \quad p \xrightarrow[\mathcal{A}]{m} q\} \ .$$

(ii) Two automata over M are *equivalent* if they have the same behaviour: that is, if they accept the same subset of M. □

[1] Nonetheless we will write nothing that contradicts the general definition.

All is thus consistent with the previous chapter: an automaton *over the alphabet* A is simply an automaton *over the monoid* A^* in our new terminology. There is a difference all the same: the set of labels of an automaton over M is potentially infinite, which leads us to modify the definition of finite automata.

cf. Def. I.1.2, p. 53

Definition 1.2 An automaton $\mathcal{A}$ over M, $\mathcal{A} = \langle Q, M, E, I, T \rangle$, is *finite* if E, the set of transitions, is finite. □

Since automata over M are labelled graphs, all the definitions and properties that deal with the 'graph side' of automata carry over from the case of automata over A which we covered in the previous chapter: the definitions of an accessible, co-accessible and useful state, of a trim automaton, and even, with some precautions, the iteration lemmas, extend directly. *Every automaton is equivalent to a trim automaton.* If E is finite, the set of useful states of $\mathcal{A}$ is finite and we can always assume in this case that Q is finite. Finally, the automata over a monoid M have as before a natural graph representation.

cf. Sec. I.1.3, p. 66

Example 1.1 Figure 1.1 shows an automaton over $\{a,b\}^* \times \{c\}^*$. The reader can verify without difficulty that it accepts the set $V_1 \cup W_1$ where V_1 and W_1 are defined by $V_1 = \{(a^m b^n, c^m) \mid n, m \in \mathbb{N}\}$ and $W_1 = \{(a^m b^n, c^n) \mid n, m \in \mathbb{N}\}$. □

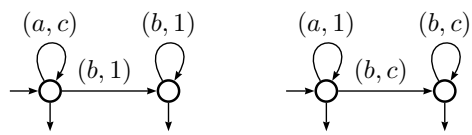


Figure 1.1: An automaton $\mathcal{A}_s$ over $\{a,b\}^* \times \{c\}^*$

To characterise the behaviour of finite automata over an arbitrary monoid we must first define the rational subsets of such a monoid.

Exercises

- 1.1 Let $\mathbb{N}^2 = \{a\}^* \times \{b\}^*$. Show that the set $(a,b)^* = \{(a^n, b^n) \mid n \in \mathbb{N}\}$ is accepted by a finite automaton over $\mathbb{N}^2$.

- 1.2 Demonstrate with an example that the construction of the cartesian product used in Chapter I to accept the intersection of behaviours of two automata does not work for automata over arbitrary (that is, non-free) monoids. Observe that the same remark holds for automata with spontaneous transitions.

cf. Prop. I.1.4, p. 61

1.3 State the iteration lemmas that hold for the subset accepted by a finite automaton over M.

1.2 Rational sets

Here again the definitions given in the case of free monoids extend easily to the case of arbitrary monoids. We will do this adopting an algebraic point of view.

1.2.1 The semiring $\mathfrak{P}(M)$

The *product* of two subsets P and Q of M is defined by

$$P\,Q = \{p\,q \mid p \in P, q \in Q\} \ .$$

The set $\mathfrak{P}(M)$ equipped with union and product operations is a semiring. It is a *complete* semiring: the union operation is defined for any set of operands, possibly infinite; that is, an expression of the form

cf. Exer. 0.5.2, p. 29.c

$$\bigcup_{i\in I} P_i$$

is always well defined, and the laws of associativity and distributivity extend to this case:

cf. Equat. (I.2.4), p. 85

$$\bigcup_{j\in J}\left(\bigcup_{i\in I_j} P_i\right) = \bigcup_{i\in\{\bigcup_{j\in J} I_j\}} P_i \qquad \text{and} \qquad \left(\bigcup_{i\in I} P_i\right)\cdot\left(\bigcup_{j\in J} R_j\right) = \bigcup_{(i,j)\in I\times J} (P_i R_j) \ .$$

The set $\{\emptyset, \{1_M\}\}$ is a sub-semiring of $\mathfrak{P}(M)$, isomorphic to $\mathbb{B}$, with which it will from now on often be identified.

Like every map, a morphism $\alpha\colon M \to N$ *extends by additivity* to a map, still written α, from $\mathfrak{P}(M)$ to $\mathfrak{P}(N)$. It is then immediate to verify:

Lemma 1.1 *If $\alpha\colon M \to N$ is a morphism on monoids, then*

$$\alpha\colon \mathfrak{P}(M) \to \mathfrak{P}(N)$$

is a morphism on complete semirings. ■

We next define the *star* operation on $\mathfrak{P}(M)$:

$$P^* = \bigcup_{n\in\mathbb{N}} P^n = \{p_1 p_2 \cdots p_n \mid n \geq 0, p_i \in P\} \ .$$

By definition, $P^0 = 1_M$ and, according to the definition, P^* is closed under product: P^* is a submonoid; conversely, each P^n, and hence P^*, is contained in every submonoid which contains P. If $\langle P \rangle$ denotes the submonoid generated by P, we have just shown that:

cf. Sec. 0.2

Property 1.1 $P^* = \langle P \rangle$. ■

This property gives a first algebraic interpretation of the star operation: *it is the transformation into the generated submonoid.* Hence we easily obtain:

Lemma 1.2 *Let $\alpha\colon M \to N$ be a morphism on monoids and P be contained in M. Then $(P^*)\alpha = (P\alpha)^*$.* ■

1.2.2 Rational operations and subsets

The three operations *union*, *product* and *star* are called *rational operations* on M (in fact, on $\mathfrak{P}(M)$). As in the previous chapter, we are naturally led to the following series of definitions:

cf. Def. I.2.2, p. 86

Definitions 1.3
(a) A family of subsets of M is *rationally closed* if it is closed under the rational operations. That is, the family $\mathcal{F}$ is rationally closed if and only if

$$\forall X, Y \subseteq M \qquad X, Y \in \mathcal{F} \quad \Longrightarrow \quad X + Y\,,\ X\,Y\,,\ X^* \in \mathcal{F}\ .$$

(b) The *rational closure* of a family $\mathcal{F}$ of subsets of M is the *smallest* family of subsets of M that contains $\mathcal{F}$ and which is rationally closed. It is written $\operatorname{Rat}\mathcal{F}$.
(c) A subset of M is called *rational* if it belongs to *the rational closure of finite subsets* of M. The family of rational subsets of M is written $\operatorname{Rat} M$. □

Example 1.2 A rational subset in the free commutative monoid.
Let $\mathbb{N}^2 = \{a\}^* \times \{b\}^*$. Then $(a,b)^* = \{(a^n, b^n) \mid n \in \mathbb{N}\} \in \operatorname{Rat}\mathbb{N}^2$. □

Example 1.1 (continued) Rational subsets in a direct product of free monoids. The two subsets V_1 and W_1 defined in the previous section are rational subsets of $\{a,b\}^* \times \{c\}^*$:

$$V_1 = \{(a^m b^n, c^m) \mid n, m \in \mathbb{N}\} = (a,c)^*(b,1)^* \quad \text{and}$$
$$W_1 = \{(a^m b^n, c^n) \mid n, m \in \mathbb{N}\} = (a,1)^*(b,c)^*\ .$$ □

1.2.3 Rational expressions

Like those of automaton or rational subset, the notions of *rational expression over a monoid*, of the subset denoted by an expression, or of the interpretation of an expression in a monoid, replicate those defined on free monoids. We write them out for completeness, but with no further comment.

cf. Sec. I.4, p. 123

cf. Def. I.4.1, p. 124

Definition 1.4 A *rational expression over* M is a *formula* obtained recursively from the elements of M and the functions $\{\mathsf{0}, \mathsf{1}, +, \cdot, *\}$ in the following manner:
(i) $\mathsf{0}$, $\mathsf{1}$, and m, for all m in M, are (atomic) rational expressions;
(ii) If E and F are two rational expressions, then $(\mathsf{E} + \mathsf{F})$, $(\mathsf{E} \cdot \mathsf{F})$, and (E^*) are rational expressions.
We write $\mathsf{Rat}\mathsf{E}\,M$ for the set of rational expressions over M. □

We will use the usual conventions to simplify the notation of expressions: associativity of sum and product, and operator precedence.

Definition 1.5 To each rational expression E in $\mathsf{Rat}\mathsf{E}\,M$, a corresponding subset of M, written $|\mathsf{E}|$, is defined recursively by: *cf. Def. I.4.2, p. 125*

(i) for atomic expressions:

$$|0| = \emptyset \ , \quad |1| = 1_M \ , \quad \text{and} \quad |m| = m \quad \text{for all } m \text{ in } M \ ; \tag{1.1}$$

(ii) for composite expressions:

$$|(\mathsf{E}+\mathsf{F})| = |\mathsf{E}| \cup |\mathsf{F}| \ , \quad |(\mathsf{E}\cdot\mathsf{F})| = |\mathsf{E}|\cdot|\mathsf{F}| \quad \text{and} \quad |(\mathsf{E}^*)| = \{|\mathsf{E}|\}^* \ . \tag{1.2}$$

We will say that E *denotes* the subset $|\mathsf{E}|$. □

The analogue of Proposition I.4.1 gives: *cf. p. 126*

Proposition 1.3 *A subset of M is rational if and only if it is denoted by a rational expression over M.* ■

From this we conclude, since there is only a finite number of atoms in an expression:

Corollary 1.4 *Every rational subset of M is contained in a finitely generated submonoid of M.* ■

In particular, we have:

Corollary 1.5 *A monoid is a rational subset of itself if and only if it is finitely generated.* ■

Let E be a rational expression over a set $\mathbf{X}$ of variables. An *interpretation* of E in M is a map τ from $\mathrm{Var}\,\mathsf{E}$ to $\mathfrak{P}(M)$, and hence a morphism from $(\mathrm{Var}\,\mathsf{E})^*$ to $\mathfrak{P}(M)$. The *result* of the interpretation τ of E, written $[\mathsf{E}]\tau$, is defined as for the result of an interpretation in A^*, by induction on the depth of E. And we have:

Proposition 1.6 *Let $\mathcal{F}$ be a family of subsets of M. A subset R of M belongs to the rational closure of $\mathcal{F}$,* $\mathrm{Rat}\,\mathcal{F}$*, if and only if there exists a rational expression* E *over a set of variables, and an interpretation of* E *which takes values in $\mathcal{F}$, whose result is R.* ■

1.2.4 Image under morphism

An elementary but fundamental property is that the family of rational subsets is *closed under morphism*. The following proposition puts this more precisely.

Proposition 1.7 *Let $\alpha\colon M \to N$ be morphism on monoids.*

(i) *The image under α of a rational over M is a rational over N.*

(ii) If α is surjective, *every rational over N is the image of a rational over M.*

In other words,

(i) $R \in \mathrm{Rat}\,M \quad \Longrightarrow \quad R\alpha \in \mathrm{Rat}\,N$.

(ii) α *surjective* $\quad \Longrightarrow \quad \forall U \in \mathrm{Rat}\,N\,,\ \exists R \in \mathrm{Rat}\,M \quad U = R\alpha$.

Proof. (i) Let $\mathcal{D}$ be the family of subsets of M whose image under α is a rational over N (that is, $P \in \mathcal{D} \Rightarrow P\alpha \in \operatorname{Rat} N$). By Lemma 1.1, $\mathcal{D}$ is closed under union and product; by Lemma 1.2, $\mathcal{D}$ is closed under star: $\mathcal{D}$ is rationally closed. As $\mathcal{D}$ contains all finite subsets, it contains $\operatorname{Rat} M$.

(ii) Let $\mathcal{E}$ be the family of subsets of N such that there exists a rational subset in M whose image under α they are (that is, $V \in \mathcal{E} \Rightarrow \exists R \in \operatorname{Rat} M \ R\alpha = V$). As above, $\mathcal{E}$ is rationally closed. Since α is surjective, $\mathcal{E}$ contains all the singletons and hence $\operatorname{Rat} N$. ■

This proposition will be used many times. By considering special cases, most often where M is a free monoid A^*, we can 'lift' the study of properties of rational subsets in arbitrary monoids into the corresponding ones of the same subsets in a free monoid.

1.2.5 Intersection and inverse morphism

We began by observing the similarities between rational subsets of an arbitrary monoid and those of a free monoid; without further ado we must now consider their distinguishing features.

The family of rational subsets of a monoid M is not in general closed *either under inverse morphism, or under intersection*, contrary to the family of rational subsets of a free monoid. If M is not finitely generated, $\operatorname{Rat} M$ is not closed under complement. If M is finitely generated, and if $\operatorname{Rat} M$ is not closed under intersection, it is not closed under complement either (because it is closed under union and contains M). But $\operatorname{Rat} M$ can be closed under intersection without being closed under complement.

Example 1.2 (continued) Let $\alpha\colon \{a,b\}^* \to \{a\}^* \times \{b\}^*$ be the canonical morphism (that is, $a\alpha = (a,1)$ and $b\alpha = (1,b)$). The inverse image of $(a,b)^*$,

$$((a,b)^*)\alpha^{-1} = \{f \in \{a,b\}^* \mid |f|_a = |f|_b\} = Z_1 \ ,$$

cf. Exer. I.1.22, p. 73 is not rational in $\{a,b\}^*$. □

Example 1.1 (continued) In $\{a,b\}^* \times \{c\}^*$, the intersection of

$$V_1 = (a,c)^*(b,1)^* \qquad \text{and} \qquad W_1 = (a,1)^*(b,c)^*$$

is equal to

$$V_1 \cap W_1 = \{(a^n b^n, c^n) \mid n \in \mathbb{N}\}$$

which is not rational since its projection over $\{a,b\}^*$,

$$(V_1 \cap W_1)\pi_1 = \{a^n b^n \mid n \in \mathbb{N}\} = P_1 \ ,$$

cf. Rem. I.1.10, p. 72 is not (otherwise there would be a contradiction with Proposition 1.7). □

However, certain (non-free) monoids M have the property that, like $\operatorname{Rat} A^*$, $\operatorname{Rat} M$ is an effective Boolean algebra. Several of these clearly interesting cases are studied in the supplementary sections of this chapter (and of the next).

cf. p. 301, p. 323, and p. 476

Exercises

1.4 Let $\alpha\colon M \to N$ be a morphism of monoids.

(a) Verify that $\alpha\colon \mathfrak{P}(M) \to \mathfrak{P}(N)$ is a morphism of semirings (Lemma 1.1).

(b) Verify that for all P included in M, $(P^*)\alpha = (P\alpha)^*$ (Lemma 1.2).

• 1.5 Show, using a simple example, that P can be a rational subset M contained in a (rational) submonoid N of M without necessarily being a rational subset of N.

• 1.6 Let M be a monoid without divisors of the identity (that is, $pq = 1_M \Leftrightarrow p = q = 1_M$). Show that if R is a rational set of M, then so is $R \setminus 1_M$.

• 1.7 Give an example of a (finitely generated) monoid whose rational subsets are closed under intersection but not under complement.

1.3 Behaviour of finite automata

We are now in a position to characterise the subsets accepted by a finite automaton over a monoid M.

Theorem 1.1 *A subset of a monoid M is rational if and only if it is accepted by a finite automaton over M.*[2]

We could give a direct proof of this statement by reworking the previous chapter's proof of Kleene's Theorem or that of the Fundamental Theorem of Automata from the next chapter, of which it is merely a special case. To avoid tiresome restatement, we shall instead *deduce* it from Theorem I.2.1 by a sort of lifting. Let us first define an elementary construction on automata, which is interesting in itself.

cf. Th. I.2.1, p. 87

cf. Th. III.2.1, p. 409

Let $\alpha\colon M \to N$ be a morphism and $\mathcal{A} = \langle Q, M, E, I, T \rangle$ an automaton over M. We will say that the automaton $\mathcal{B} = \langle Q, N, F, I, T \rangle$ over N is the *relabelling*[3] of $\mathcal{A}$ under α if $\mathcal{B}$ is *a copy* of $\mathcal{A}$, except that the label of each transition in $\mathcal{B}$ is the image under α of the label of the corresponding transition in $\mathcal{A}$:

$$F = \{(p, m\alpha, q) \mid (p, m, q) \in E\} \ .$$

Since $\mathcal{B}$ is a copy of $\mathcal{A}$, there is a bijection between the transitions,[4] between the computations, and between the successful computations of $\mathcal{B}$ and of $\mathcal{A}$, and we therefore have:

[2] N.B. This proposition is not standard. Its disturbing similarity with certain statements of Kleene's Theorem, which holds only for (finitely generated) free monoids, can make it confusing. We will explain in greater detail in Subsection 2.2, p. 238.

[3] We will not say that $\mathcal{B}$ is the image of $\mathcal{A}$ under the *morphism* α because we reserve the term *automaton morphism* for a different and much richer notion to be studied in Section 3.

[4] Here is an uncontrived example where an automaton can have two distinct transitions with the same source, destination and label; *cf.* the comment at the beginning of Sec. 1.1.

Lemma 1.8 *Let $\alpha\colon M \to N$ be a morphism, $\mathcal{A}$ an automaton over M, and $\mathcal{B}$ the relabelling of $\mathcal{A}$ under α. Then the behaviour of $\mathcal{B}$ is the image under α of the behaviour of $\mathcal{A}$; that is,*

$$|\mathcal{B}| = |\mathcal{A}|\alpha \ . \qquad \blacksquare$$

Note that if α is surjective and $\mathcal{A} = \langle Q, N, E, I, T \rangle$ is an automaton over N, we can go the other way: the automaton $\mathcal{C} = \langle Q, M, G, I, T \rangle$ is a copy of $\mathcal{A}$ except that each of its transition labels is an element of M in the inverse image under α of the corresponding transition label in $\mathcal{A}$ (and this is possible since α is surjective). It is $\mathcal{A}$ that is a relabelling of $\mathcal{C}$ under α and we have

$$|\mathcal{A}| = |\mathcal{C}|\alpha \ .$$

Proof of Theorem 1.1. Let $\mathcal{A} = \langle Q, M, E, I, T \rangle$ be an automaton over M. For clarity, we rename the set E to Z: that is, Z is an alphabet (finite, since $\mathcal{A}$ is finite) whose letters are the elements of E. Let $\eta\colon Z \to M$ be the map that projects the label of a transition:

$$(p, m, q)\eta = m \ .$$

This map extends to a morphism $\eta\colon Z^* \to M$ which we can assume is surjective (that is, we can take $M = Z^*\eta$). From $\mathcal{A}$, we construct the automaton $\mathcal{C}$ over alphabet Z:

$$\mathcal{C} = \langle Q, Z, D, I, T \rangle$$

such that $\mathcal{A}$ is its relabelling under η; the set of its transitions is given by

$$D = \{\big(p, (p, m, q), q\big) \mid (p, m, q) \in E = Z\} \qquad \text{and} \qquad |\mathcal{A}| = |\mathcal{C}|\eta \ .$$

The behaviour of $\mathcal{C}$, a finite automaton over Z, is a rational subset of Z^* (Theorem I.2.1) whose image under a morphism is a rational subset of M (Proposition 1.7). Thus, the condition is sufficient.

Conversely, let U be a rational subset of M. Without loss of generality, we can assume that M is finitely generated and hence that there exists a finite alphabet A and a surjective morphism $\alpha\colon A^* \to M$. By Proposition 1.7, there exists a rational subset R of A^* such that $R\alpha = U$. By Kleene's Theorem, there exists an automaton $\mathcal{A}$ over A whose behaviour is equal to R. Let $\mathcal{B}$ be the relabelling of $\mathcal{A}$ by α. It is an automaton over M whose behaviour is equal to $|\mathcal{A}|\alpha = U$. $\blacksquare$

Example 1.1 (continued) Let us return to the automaton $\mathcal{A}_s$ of Figure 1.1 and apply to it the construction above. We obtain (Figure 1.2) an automaton $\mathcal{C}_s$ over the alphabet $C = \{x, y, z, t, u, v\}$ whose behaviour is easily computed:

$$|\mathcal{C}_s| = x^* \cup x^* y\, z^* \cup t^* \cup t^* u\, v^* \ .$$

The morphism η_1 from C^* to M is defined by

$$x\eta_1 = (a, c) \ , \quad y\eta_1 = z\eta_1 = (b, 1) \ , \quad t\eta_1 = (a, 1) \quad \text{and} \quad u\eta_1 = v\eta_1 = (b, c) \ ,$$

and gives

$$\begin{aligned} |\mathcal{A}_s| = |\mathcal{C}_s|\eta_1 &= (a, c)^* \cup (a, c)^*(b, 1)\,(b, 1)^* \cup (a, 1)^* \cup (a, 1)^*(b, c)\,(b, c)^* \\ &= (a, c)^*(b, 1)^* \cup (a, 1)^*(b, c)^* \ . \end{aligned} \qquad \square$$

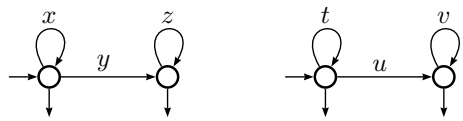


Figure 1.2: The automaton $\mathcal{C}_s$, lifting of $\mathcal{A}_s$

Theorem 1.1 and its proof open the way to a natural extension (which we already hinted at in Section I.2.4 and which anticipates what we shall see in Chapter III in a more general setting). Let $\mathcal{A} = \langle Q, M, E, I, T \rangle$ be an automaton over M whose labels are no longer elements of M but *subsets* of M. The label of a computation c of $\mathcal{A}$, *cf. p. 97, p. 402*

$$c := p_0 \xrightarrow{U_1} p_1 \xrightarrow{U_2} p_2 \cdots \xrightarrow{U_n} p_n \ ,$$

is the subset of M, still written $|c|$, that is the product of the successive transition labels of c:

$$|c| = U_1 U_2 \cdots U_n \ .$$

The behaviour of $\mathcal{A}$ is, as before, the *union* of the labels of successful computations and such an $\mathcal{A}$ is called a *generalised finite automaton over* M if the set E of transitions is finite.

Theorem 1.2 *A subset of a monoid M belongs to the rational closure of a family of subsets $\mathcal{F}$ if and only if it is the behaviour of a generalised finite automaton whose labels are elements of $\mathcal{F}$.*

Proof. Let $\mathcal{A} = \langle Q, M, E, I, T \rangle$ be a generalised finite automaton over M. Let $\mathbf{X}$ be a set of variables in bijection with E; for each $\mathbf{x}$ in $\mathbf{X}$, write $\mathbf{x}\tau$ for the transition label of $\mathcal{A}$ that corresponds to $\mathbf{x}$ in this bijection: we have thus defined an interpretation in M of every rational expression over $\mathbf{X}$.

Let $\mathcal{C} = \langle Q, \mathbf{X}, D, I, T \rangle$ be obtained, as in the proof of Theorem 1.1, from $\mathcal{A}$ by replacing each transition by a transition labelled by the corresponding variable. We have

$$|\mathcal{A}| = [|\mathcal{C}|]\tau \ .$$

Conversely, every element R of $\operatorname{Rat}\mathcal{F}$ is the result $[\mathsf{F}]\tau$ of a rational expression F under an interpretation τ whose value is in $\mathcal{F}$ (Proposition 1.6).

We have $[\mathsf{F}]\tau = (L[\mathsf{F}])\,\tau$ (Proposition I.4.4) and $L[\mathsf{F}]$ is the behaviour of a finite automaton $\mathcal{C}$ over $\mathbf{X} = \operatorname{Var}\mathsf{F}$. By replacing each label $\mathbf{x}$ by $[\mathbf{x}]\tau$ we transform $\mathcal{C}$ into a generalised finite automaton $\mathcal{A}$ whose behaviour is R. ■

Remark 1.1 We could have stated this result in Chapter I (for $M = A^*$) as a corollary of Proposition I.4.4 with the same proof. I did not do so in order not to overuse generalised automata in that chapter. □ *cf. p. 128*

Contrary to what we saw with automata over a free monoid, the notions of deterministic automaton and of determinisation *make no sense* for automata over an arbitrary monoid M. We can see here the reason that the family $\operatorname{Rat} M$ is not closed under intersection. These notions are replaced by that of the action of M on a set, which we will study in the next section.

Exercise

• 1.8 **Local languages.** Let us first make a definition:

Definition 1.6 A *local language* (over an alphabet) is a set of words whose first letter, last letter and all the factors of length 2 (that is, pairs of letters in the words) belong respectively to given subsets; that is,

$$L \subseteq A^* \text{ is local} \iff \exists D, F \subseteq A\,,\ \exists V \subseteq A^2$$
$$L = (DA^* \cap A^*F) \setminus (A^*VA^*) \quad . \qquad \square$$

(Note that in fact we mention in the formula the set V of factors of length 2 which *do not appear* in the words of L.)

(a) $A = \{a, b\}$. Show that the following languages are local:

$$a^+ \ , \quad \{a\,b\}^+ \ , \quad a^+ \cup b^+ \ , \quad \{a\} \qquad \text{and} \qquad \{a\,b\} \ .$$

Show that the following languages are not local:

$$a^+ \cup \{a\,b\}^+ \ , \quad a^+\{a\,b\}^+ \qquad \text{and} \qquad \{a\,a\} \ .$$

(b) Show that the family of local languages is closed under intersection and star. Show that it is not closed under union or product.

(c) Show that the set of factors of a local language is a local language.

(d) Show that a language L over A^* is local if and only if for all f, g, h and k in A^* and all a in A, we have that if $f\,a\,g$ and $h\,a\,k$ are in L then $f\,a\,k$ is in L.

(e) Prove:

Proposition 1.9 *A language is rational if and only if it is the image of a local language under a (strictly alphabetic) morphism.*

1.4 Unambiguous rational sets

In this subsection, we will go beyond the set point of view which prevails in this chapter and consider the subsets of a monoid which can be constructed with rational operations in such a way that each of their elements is obtained only *once*.

In a sense this preoccupation is premature, and would be better in the next chapter, '*Weighted* automata', where we will have the concepts we need to deal with this class of problems and place their result in perspective. But we can also tackle the notion of ambiguity, or non-ambiguity, in a more naive manner and without needing a general definition of formal power series.

1.4.1 Definitions

On M – in fact on $\mathfrak{P}(M)$ – we define the *unambiguous rational operations*, a specialisation of rational operations for which we shall use the same symbols:

(i) *disjoint union*: $X \cup Y$ is 'disjoint' if $X \cap Y = \emptyset$;

(ii) *unambiguous product*: $X \cdot Y$ is 'unambiguous' if:

$$\forall x, x' \in X\,,\ \forall y, y' \in Y \qquad x\,y = x'\,y' \implies x = x' \quad \text{and} \quad y = y' \ ;$$

(iii) *unambiguous star*: $X^* = \bigcup_{n\in\mathbb{N}} X^n$ is unambiguous if each of the products X^n is unambiguous and the union is unambiguous: that is, if the X^n are pairwise disjoint.

In other words, X^* is an unambiguous star if X generates a free submonoid from M with base X:

$$x_1x_2\cdots x_n = x'_1x'_2\cdots x'_m \quad\Longrightarrow\quad \left\{\begin{array}{ll} & n = m \\ \text{and} & x_i = x'_i \quad \text{for all } i,\ 1 \le i \le n. \end{array}\right.$$

A rational operation which is not unambiguous is called *ambiguous*.[5] As in Section 1.2, the following definitions are natural.

Definitions 1.7

(i) A family $\mathcal{F}$ of subsets of M is *closed* under the unambiguous rational operations if the result of every unambiguous rational operation performed on elements of $\mathcal{F}$ is also in $\mathcal{F}$:

$$\forall X, Y \subseteq M\,,$$
$$X, Y \in \mathcal{F} \quad\Longrightarrow\quad \left\{\begin{array}{lll} X \cap Y = \emptyset & \Rightarrow & X \cup Y \in \mathcal{F}\ , \\ X \cdot Y \quad \text{unambiguous} & \Rightarrow & X \cdot Y \in \mathcal{F}\ , \\ X^* \quad \text{unambiguous star} & \Rightarrow & X^* \in \mathcal{F}\ . \end{array}\right.$$

(ii) The *unambiguous rational closure* of a family $\mathcal{F}$ of subsets of M is the *smallest* family of subsets of M that contains $\mathcal{F}$ and which is closed under the unambiguous rational operations. It is written $\mathrm{URat}\,\mathcal{F}$.

(iii) A subset of M is *unambiguously rational* if it belongs to the *unambiguous rational closure of finite subsets* of M. The family of unambiguous rational subsets of M is written $\mathrm{URat}\,M$. □

Definition 1.8 A rational expression E over M is called *unambiguous* if every rational operation that is used in the computation of the subset denoted by E is unambiguous.

We write $\mathsf{URatE}\,M$ for the set of unambiguous rational expressions over M. □

The analogue of Propositions I.4.1 and 1.3 is:

Proposition 1.10 *A subset of M is unambiguously rational if and only if it is denoted by an unambiguous rational expression over M.* ■

1.4.2 The family $\mathrm{URat}\,M$

Every unambiguous rational subset is rational but in general the converse is not true: that is, the inclusion

$$\mathrm{URat}\,M \subseteq \mathrm{Rat}\,M$$

[5]The terminology is slightly dubious. It is not the union, product or star operations which are ambiguous or unambiguous, but their instances: $X \cup Y$ is ambiguous or unambiguous, *etc.*

is *strict* in general. A rational subset which is not unambiguous will be called *inherently ambiguous* or more simply *ambiguous*. In certain cases, the most important being that of free monoids, every rational subset is unambiguous:

Theorem 1.3 *For every alphabet A,* $\mathrm{URat}\, A^* = \mathrm{Rat}\, A^*$.

In fact, this result is a direct consequence of the theorem, a generalisation of Theorem 1.1, that says that the behaviour of an automaton is described by a rational expression not only in terms of the elements accepted but also for the weight with which each element is accepted (which we shall show in the following chapter). We give here an elementary proof of Theorem 1.3.

cf. Th. III.2.1, p. 409

Proposition 1.11 *Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an unambiguous automaton over A. The rational expression given by the McNaughton–Yamada algorithm applied to $\mathcal{A}$ is unambiguous.*

cf. p. 94

Proof. Let us reconsider the proof of Proposition I.2.7. We set

$$R_{p,q} = \{f \in A^* \mid p \xrightarrow[\mathcal{A}]{f} q\} \ , \quad E_{p,q} = \{a \in A \mid p \xrightarrow[\mathcal{A}]{a} q\} \ ,$$

and we have $L(\mathcal{A}) = \bigcup_{i \in I, t \in T} R_{i,t}$. This union is unambiguous since, as $\mathcal{A}$ is unambiguous, for every word f in $L(\mathcal{A})$, there exists a unique i in I and a unique t in T such that $i \xrightarrow{f} t$ (Definition I.1.11 (ii)).

We order Q, from 1 to $n = \|Q\|$, and define the $R^{(k)}_{p,q}$, $0 \leqslant k \leqslant n$, by induction:

$$R^{(0)}_{p,q} = E_{p,q}$$
$$R^{(k+1)}_{p,q} = R^{(k)}_{p,q} + R^{(k)}_{p,k+1}(R^{(k)}_{k+1,k+1})^* R^{(k)}_{k+1,q}$$

and we have:

$$R_{p,q} = \begin{cases} R^{(n)}_{p,q} & \text{if} \quad p \neq q \\ R^{(n)}_{p,q} + 1_{A^*} & \text{if} \quad p = q \end{cases}$$

The $R^{(k)}_{p,q}$ describe the labels of computations that go from p to q without passing through intermediate states greater than k. If any of the rational operations used in the equations above is ambiguous, then we have, for some pair of states (p, q) and some f in A^*, two distinct computations in $\mathcal{A}$ with source p, destination q and label f: this contradicts the hypothesis that $\mathcal{A}$ is unambiguous. ∎

Proof of Theorem 1.3. Every rational in A^* can be accepted by a deterministic, hence unambiguous, automaton over A^*, and the proposition above enables us to deduce the conclusion. ∎

We will see later in the chapter other monoids which still obey the equality $\mathrm{URat}\, M = \mathrm{Rat}\, M$. The following statement, whose proof is left as an exercise, shows that this equality is not always true.

Fact 1.1 *If* $M = \{a,b\}^* \times \{c\}^*$ *, then* $\mathrm{URat}\, M \subsetneqq \mathrm{Rat}\, M$ *. More precisely, the subset* $V_1 \cup W_1 = (a,c)^*(b,1)^* \cup (a,1)^*(b,c)^*$ *is inherently ambiguous.* ■

Proposition 1.7 also has its analogue for unambiguous rational subsets. More precisely, we have the following proposition whose proof we leave as an exercise.

Proposition 1.12 *Let* $\alpha\colon M \to N$ *be a morphism on monoids.*

(i) *The image under* α *of an unambiguous rational of* M *over which* α *is injective is an unambiguous rational of* N*.*

(ii) If α is surjective*, every unambiguous rational of* N *is the injective image of an unambiguous rational of* M*.*

In other words:

(i) $R \in \mathrm{URat}\, M$ *and* α *injective over* $R \quad \Longrightarrow \quad R\alpha \in \mathrm{URat}\, N$*.*

(ii) α *surjective* $\Longrightarrow \forall U \in \mathrm{URat}\, N,\ \exists R \in \mathrm{URat}\, M \quad U = R\alpha$ *and* α *injective over* U*.* ■

Exercises

1.9 Let $X \in \mathrm{URat}\, M$ and F be a finite subset of M. Show that $X \setminus F \in \mathrm{URat}\, M$.

1.10 Let $\alpha\colon M \to N$ be a morphism of monoids and X and Y two non-empty subsets of M. Show that the following two conditions are equivalent:

(i) the product XY is unambiguous and α is injective over XY;

(ii) the product $(X\alpha)(Y\alpha)$ is unambiguous (in N) and α is injective over X and Y.

1.11 Prove Proposition 1.12.

• 1.12 Prove Fact 1.1.

• 1.13 **Rationally enumerable monoids.** Let M be a monoid. Let $\alpha\colon A^* \to M$ be a surjective morphism; we will say that a set R of A^* is a *cross-section* (of A^*) for α if α is injective over R and $R\alpha = M$: that is, if R is a set of representatives for the equivalence $\alpha\alpha^{-1}$.

(a) Show that if there exists a rational cross-section for α, then there exists a rational cross-section for every surjective morphism $\beta\colon B^* \to M$. We will say in this case that M is *rationally enumerable.*

(b) Show that M is rationally enumerable if and only if $M \in \mathrm{URat}\, M$.

2 Actions and recognisable sets

We start by defining the notion of *action*, which extends to arbitrary monoids that of a deterministic automaton over a free monoid and which leads to that of a recognisable subset. We then assure ourselves that these definitions are consistent with those made in the previous chapter for the free monoid and prove in all generality the closure properties which we showed for recognisable languages. We then present the process of action minimisation in the appropriate algebraic framework and finish with some applications of this point of view.

2.1 Actions on a set

2.1.1 Definition

A *(right) action*[6] of a monoid M on a set S is a map

$$\delta\colon S\times M \longrightarrow S \ ,$$

which is consistent with multiplication in M:

$$\begin{aligned} &\forall s\in S && (s,1_M)\delta = s \ , \\ &\forall s\in S\,,\ \forall m,m'\in M && ((s,m)\delta,m')\delta = (s,m\,m')\delta \ . \end{aligned} \tag{2.1}$$

If, in the pairs (s,m) which are arguments of δ, we fix m, we obtain a map from S to S. Thus, every action δ from M to S defines a map

$$\check{\delta}\colon M \longrightarrow \mathfrak{T}_S \ ,$$

from M to the monoid $\mathfrak{T}_S$ of maps from S to itself, by

$$\forall s\in S\,,\ \forall m\in M \qquad (s)[m\check{\delta}] = (s,m)\delta \ . \tag{2.2}$$

Equation (2.1) implies that $\check{\delta}$ is a *morphism* (of monoids), and that M is *represented by maps from S to itself*. The *transition monoid* of an action δ of M on S is the image $M\check{\delta}$ of M in $\mathfrak{T}_S$. Conversely, every morphism from M to $\mathfrak{T}_S$ defines, by (2.2), an action of M on S.

cf. Exer. 2.1

Notation To avoid an excess of parentheses and symbols, we will identify $m\check{\delta}$ and m and will write $s\cdot m$ for $s[m\check{\delta}]$, or for $(s,m)\delta$; Equation (2.1) then becomes

$$\forall s\in S\,,\ \forall m,m'\in M \qquad (s\cdot m)\cdot m' = s\cdot m\,m' \ . \tag{2.3}$$

We will write $s\underset{S}{\cdot} m$ if there is a risk of ambiguity as to the set on which M is acting. We will stick to $(s,m)\delta$ if the ambiguity relates to the action itself.

We define in a dual manner a *left action* of M on S: a map from $M\times S$ to S, written $m\cdot s$, which satisfies the relation

$$\forall s\in S\,,\ \forall m,m'\in M \qquad m'\cdot(m\cdot s) = m'm\cdot s \ . \tag{2.4}$$

Remark 2.1 Every action δ of M on S naturally induces an action of every submonoid M' of M on S, the 'trace' on M' of δ, which is also written δ. □

Examples 2.1 (i) Every monoid M defines an *action on itself* by multiplication – or *translation* – on the right:

$$\forall p,m\in M \qquad p\cdot m = p\,m \ ;$$

this is called *(right) regular representation* of M on itself.

[6] This is the generalisation to monoids of the more classical notion of the action of a group on a set. We shall return to this point in Subsection 6.1, p. 301.

This representation is *faithful*: that is, the corresponding morphism of M in $\mathfrak{T}_M$ is *injective*, since M is a monoid[7] (and hence $1_M \cdot m \neq 1_M \cdot m'$ if $m \neq m'$). Thus M is *isomorphic to the transition monoid* of this action.

(ii) Every morphism $\alpha\colon M \to N$ defines likewise an action of M on N:

$$\forall n \in N\,,\ \forall m \in M \qquad n \cdot m = n(m\alpha)\ . \tag{2.5}$$

This map satisfies (2.3) because α is a morphism and multiplication in N is associative. (The regular representation above corresponds to the identity morphism ι from M to itself.) The transition monoid is equal to $M\alpha$, if we use the above isomorphism to identify N and its right regular representation. □

Example 2.2 We showed in Chapter I that every automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ defines an action of A^* on $\mathfrak{P}(Q)$, and that every deterministic automaton $\mathcal{A} = \langle Q, A, E, I, T \rangle$ defines an action of A^* on Q. We shall return to this point. □ *cf. Sec. I.3, p. 101* *cf. Sec. 2.2, p. 238*

2.1.2 Matrix representation of actions

If S is *finite*, the monoid $\mathfrak{T}_S$ of maps from S to itself identifies naturally with the monoid of *row-monomial* boolean matrices of dimension S. Furthermore, the transition monoid of an action of M on S hence identifies with a monoid of *row-monomial* boolean matrices (square matrices, of dimension S), a submonoid of $\mathfrak{T}_S$. Dually, and still assuming that S is finite, the transition monoid of a *left action* of M on S identifies with a monoid of *column-monomial* boolean matrices (of dimension S).

Example 2.3 (Continuation of Example I.1.1) The deterministic automaton $\mathcal{D}_1$ that recognises $L_1 = A^* a\, b\, A^*$ defines an action δ_1 of A^* on $S_1 = \{p, q, r\}$. The corresponding morphism $\check{\delta}_1$ is defined by the two matrices $a\check{\delta}_1$ and $b\check{\delta}_1$ of $\mathbb{B}^{S_1 \times S_1}$:

$$a\check{\delta}_1 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad b\check{\delta}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

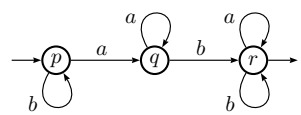


Fig. 2.1: The automaton $\mathcal{D}_1$

which allows the effective computation of the transition monoid M_1. To this end, set $x = a\check{\delta}_1$ and $y = b\check{\delta}_1$; we obtain

$$x\,x = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = x \qquad \text{and} \qquad y\,y = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = y\ .$$

We set:

$$y\,x = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = t \qquad \text{and} \qquad x\,y = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = z\ ,$$

[7] This would not necessarily be the case if M were only a semigroup.

and we also verify that this last element is a zero:

$$x\,z = z\,x = y\,z = z\,y = z$$

We thus obtain the multiplication table of M_1 shown opposite. □

1	x	y	t	z
x	x	z	z	z
y	t	y	t	z
t	t	z	z	z
z	z	z	z	z

Let δ be an action of M on S; if G is a generating set of M, it is enough to know the value of δ on $S\times G$ to be able to deduce, by (2.1) (or (2.3)), the complete definition of δ on $S\times M$. On the other hand, and if M is not a free monoid, *it is not true* that an arbitrary map from $S\times G$ to S allows us to define an action of M on S.

cf. Exer. 2.3

Example 2.4 Action of a non-free monoid. Let A and B be two alphabets, $M = A^*\times B^*$, and $S_2 = \{p,q,r,s\}$. The set $C = (A\times\{1_{B^*}\})\cup(\{1_{A^*}\}\times B)$ is a generating set of M. We *verify* that the map $\delta_2\colon S_2\times C\to S_2$ defined by

cf. Exer. 2.2

$$p\cdot(a,1) = q\cdot(a,1) = q \qquad r\cdot(a,1) = q\cdot(1,b) = s$$
$$p\cdot(1,b) = r\cdot(1,b) = r \qquad s\cdot(a,1) = s\cdot(1,b) = s$$

for all a in A and all b in B indeed corresponds to an action of M on S_2. The matrix representation of this action is given by

$$\forall a\in A \qquad (a,1)\check{\delta}_2 = \begin{pmatrix} 0&1&0&0\\ 0&1&0&0\\ 0&0&0&1\\ 0&0&0&1 \end{pmatrix}, \quad \forall b\in B \qquad (1,b)\check{\delta}_2 = \begin{pmatrix} 0&0&1&0\\ 0&0&0&1\\ 0&0&1&0\\ 0&0&0&1 \end{pmatrix}.$$ □

2.1.3 Subsets recognised by an action

In fact, what we call an *action of M on S* usually consists not only of a map $\delta\colon S\times M\to S$ but also of *a distinguished element* s_0 of S; that is, a nullary operation (from S to itself). This choice is in general sufficiently canonical to justify its not being more explicit.

In this way, the regular representation of M on itself is accompanied by the choice of the unit element 1_M, like the nullary operation of the monoid; the action of M on a monoid N defined by a morphism is accompanied by the choice of 1_N.

Definition 2.1 A subset P of M is *recognised* by an action of M on S if there exists a subset R of S such that P is the set of elements of M that send the distinguished element s_0 to R:

$$P = \{m\in M \mid s_0\cdot m\in R\}\ .$$

A subset P of M is *recognisable* if it is recognised by an action of M on a *finite* set.

The family of recognisable subsets of M is written $\operatorname{Rec} M$. □

Example 2.4 (continued) In $S_2 = \{p,q,r,s\}$, we take p as the element distinguished by the action δ_2. The subset $A^*\times\{1_{B^*}\}$ is recognised by $\{p,q\}$ (understanding that we mean the action δ_2); the set $\{1_{A^*}\}\times B^*$ is recognised by $\{p,r\}$. They are both recognisable in $A^*\times B^*$. □

Remark 2.2 Let P be a subset recognised by an action δ of M on S. Following Remark 2.1, $P \cap M'$ is recognised by the action δ of M' on S, for all submonoids M' of M. □

We have thus in particular:

Property 2.1 *Let P be recognisable in M. For all submonoids M' of M, $P \cap M'$ is recognisable in M'.* ■

Note that in this statement neither M' nor $P \cap M'$ is necessarily recognisable in M.

Definition 2.1 is slightly out of line with usual definitions.[8] I have stated it this way the better to illustrate a symmetry between actions and automata, and between recognisable subsets and rational subsets, a symmetry which results, in the case of free monoids, in an identity, by its very definition, or alternatively by the effect of determinisation on automata (*cf.* below, Section 2.2).

But we must also recognise (!) that Definition 2.1 has a hidden drawback: the notion of action is directed: right action, left action; as the same should apply to the notion of recognisability, should we not speak of 'left recognisable' and 'right recognisable' sets? Proposition 2.1 below shows that in fact the definition gives a perfectly symmetric characterisation of recognisable sets. *cf. also Exer. 2.6*

Proposition 2.1 *A subset P of M is* recognisable *– that is, it is recognised by an action of M on a finite set S – if and only if there exists a morphism α on M in a finite monoid N such that*

$$P = P\alpha\alpha^{-1} \quad .$$

Proof. Property 2.2 shows that the condition is necessary in the case where S, and hence $\mathfrak{T}_S$, is finite.

Property 2.2 *Let $\check{\delta}\colon M \to \mathcal{T}$ be the morphism defined by (or which defines) an action δ of M on S. If P is a subset of M recognised by δ, then $P = P\check{\delta}\check{\delta}^{-1}$.*

Proof. Let s_0 be the distinguished element of S (by δ) and R the subset that recognises P:

$$P = \{m \mid s_0 \cdot m \in R\} \quad .$$

Let Q in $\mathfrak{T}_S$ be defined by

$$Q = \{\xi \in \mathfrak{T}_S \mid (s_0)\xi \in R\} \quad .$$

We then have $P = \{m \in M \mid m\check{\delta} \in Q\} = Q\check{\delta}^{-1}$, hence the conclusion. ■

[8] If the definition of a recognisable subset in a monoid may be said to be 'usual'. In Berstel, *Transductions of Context-Free Languages* [27], and Pin, *Variétés de langages formels* [185], for example, it is the characterisation of Proposition 2.1 that is taken as the definition; in Eilenberg, *op. cit.*, it is that of the upcoming Theorem 2.3.

Property 2.3 shows that the condition is sufficient in the case where the monoid N is finite.

Property 2.3 *Let $\alpha\colon M \to N$ be a morphism. A subset P of M is recognised by the action defined by α if and only if $P = P\alpha\alpha^{-1}$.*

Proof. If P is recognised by the action α, there exists, by definition, a subset Q of N such that

$$P = \{m \mid 1_N \cdot m\alpha = m\alpha \in Q\} = Q\alpha^{-1}$$

so that $P\alpha = Q\alpha^{-1}\alpha = Q$ and $P\alpha\alpha^{-1} = P$.

Conversely, if $P = P\alpha\alpha^{-1}$, then $P = \{m \mid 1_N \cdot m\alpha \in P\alpha\}$ and P is recognised by the action of M on N defined by α. ■

Hence Proposition 2.1 is proved. ■

This characterisation of recognisable subsets often has the advantage of being easier to apply, by being better suited to certain proofs. It will therefore be useful to have the following terminology at our command, which is entirely justified by Properties 2.2 and 2.3.

cf. Sec. 2.5

A subset P of M is *recognised by a morphism* α of M in N if

$$P = P\alpha\alpha^{-1} \quad .$$

In other words, P is *recognised by* α if P is a saturated subset (that is, P is a union of classes) for the map equivalence of α, or if there exists a subset Q of N such that $P = Q\alpha^{-1}$ (we also say, in the case where α is implied by the context, that Q recognises P). More briefly, P is *recognised by* N if there exists a morphism α of M in N which recognises P. If P is saturated by a congruence ρ, P is recognised by M/ρ.

Example 2.3 (continued) Let $\check{\delta}_1\colon A^* \to M_1$ be the morphism defined by the action δ_1; we have

$$L(\mathcal{A}_1) = A^*a\,b\,A^* = z\check{\delta}^{-1} \quad .$$

More broadly, the map equivalence of $\check{\delta}_1$ determines a partition of A^* shown in the adjacent figure, the inverse images of elements of M_1. □

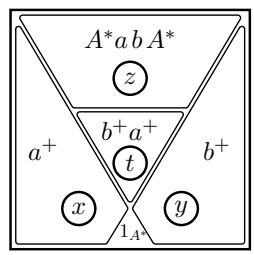


Fig. 2.2: Classes mod $\check{\delta}_1$ in A^*

Example 2.4 (continued) Let $U_1 = \{1, 0\}$ be *the* 2-element monoid with a zero, A and B two alphabets and $\alpha\colon A^* \times B^* \to U_1 \times U_1$ the morphism defined by

$$(a, 1_{B^*})\alpha = (0, 1) \quad \text{and} \quad (1_{A^*}, b)\alpha = (1, 0)$$

for all a in A and b in B. The subsets $A^* \times 1_{B^*} = (U_1 \times \{1\})\alpha^{-1}$ and $1_{A^*} \times B^* = (\{1\} \times U_1)\alpha^{-1}$ are recognisable in $A^* \times B^*$. □

We conclude directly from Proposition 2.1 the following:

Corollary 2.2 *The family of recognisable subsets is closed under inverse morphism (that is, if* $\alpha\colon M \to N$ *and* $Q \in \operatorname{Rec} N$ *, then* $Q\alpha^{-1} \in \operatorname{Rec} M$ *).*

Proof. Let $\beta\colon N \to H$ be a morphism in a finite monoid H which recognises Q : $Q = Q\beta\beta^{-1}$. We have, setting $P = Q\alpha^{-1}$ (which means that $Q = P\alpha$),

$$P = Q\alpha^{-1} = Q\beta\beta^{-1}\alpha^{-1} = P\alpha\beta\beta^{-1}\alpha^{-1} = P\alpha\beta(\alpha\beta)^{-1}$$

and P is recognised by the morphism $\alpha\beta$ (in the finite monoid H). ■

This last statement admits (when α is surjective) a converse which we will prove in Subsection 2.4 and use frequently thereafter. *cf. Cor. 2.12, p. 250*

Exercises

- 2.1 Let $\delta\colon S\times M \to S$ and $\tau\colon M \to \mathfrak{T}_S$ be two maps such that

$$\forall s \in S\,,\ \forall m \in M \qquad (s)[m\tau] = (s,m)\delta \ .$$

Verify that δ is an action if and only if τ is a morphism. (In other words, verify the assertions that follow Equation (2.2).)

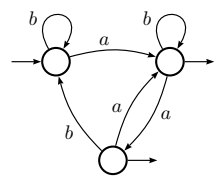
- 2.2 Verify that the map δ_2 given in Example 2.4 really does correspond to an action.

- 2.3 More generally, let A be an alphabet (not necessarily finite), R a set (not necessarily finite) of pairs (u,v) of words from A^*, and M the quotient of A^* by the congruence generated by R. Let S be an arbitrary set and δ a map from $S\times A$ to S. Show that δ defines an action of M on S if, for all s in S and every pair (u,v) in R, we have $(s,u)\delta = (s,v)\delta$.

2.4 Prove:

Proposition 2.3 *A subset of* $\mathbb{N}$ *is recognisable if and only if it is a finite union of arithmetic progressions.*

- 2.5 Compute the transition monoid defined by the deterministic automaton in the adjacent figure.



- 2.6 Let P be a subset recognised by a (right) action of M on S. Describe a *left* action of M on a set T which also recognises P and such that T is finite when S is.

- 2.7 **'Lead or Delay' action**

Let B be an alphabet and H_B the set defined by $H_B = (B^*\times 1_{B^*}) \cup (1_{B^*}\times B^*) \cup \{\mathbf{0}\}$. We define a map ψ from $B^*\times B^*$ to H_B by

$$\forall f,g \in B^* \qquad (f,g)\psi = \begin{cases} (g^{-1}f, 1_{B^*}) & \text{if} \quad g \leqslant f \\ (1_{B^*}, f^{-1}g) & \text{if} \quad f \leqslant g \\ \mathbf{0} & \text{otherwise} \ , \end{cases}$$

then a map ω_B from $H_B\times(B^*\times B^*)$ to H_B by

$$\forall f,g,u,v \in B^* \qquad \begin{cases} ((f,1_{B^*}),(u,v))\,\omega_B & = & (f\,u,v)\psi \\ ((1_{B^*},g),(u,v))\,\omega_B & = & (u,g\,v)\psi \\ (\mathbf{0},(u,v))\,\omega_B & = & \mathbf{0} \ . \end{cases}$$

(a) Verify that ω_B is an action of $B^* \times B^*$ on H_B. Interpret – that is, justify – the name 'Lead or Delay'.

(b) What is the transition monoid of this action if $\|B\| \geqslant 2$? if $\|B\| = 1$?

2.2 Recognisable here, recognisable there

Def. 2.1 The definition of recognisable subsets is not just 'slightly out of line', it is problematic since, as stated for an arbitrary monoid M, it does not coincide with that given for a free monoid A^*. Before going any further, this point must be cleared up.

2.2.1 Consistency

We defined the recognisable subsets of a monoid by actions, and the recognisable languages, or recognisable subsets of a free monoid, by automata, which we shall see must not be confused with actions. Are we consistent? Yes, in the sense that we have given the *same name* to the *same objects*.

A *deterministic* automaton over an alphabet A: $\mathcal{A} = \langle Q, A, E, q_0, T \rangle$ could be written in the form

$$\mathcal{A} = \langle Q, A, \delta, q_0, T \rangle \ ,$$

where δ is an action of A^* on Q (Property I.3.3). The automaton in addition specifies first, the distinguished element q_0 naturally associated with the action δ, and second the subset T that will recognise the subset $L(\mathcal{A})$. Thus, via the process of determinisation (Proposition I.3.2), every recognisable subset of A^* is recognised by a deterministic finite automaton over A, and hence by an action of A^* on a finite set.

Conversely, every action δ of A^* on a set S is completely defined by a map from $S \times A$ to S and the choice of a distinguished element s_0 of S. Every subset P of A^* recognised by the action δ corresponds to a subset R of S and coincides with the subset recognised (or accepted) by the automaton

$$\mathcal{A} = \langle S, A, \delta, s_0, R \rangle \ .$$

The knot is tied and there really is an exact correspondence between the subsets of a free monoid A^* recognised by a finite automaton over A and the subsets recognised by an action of A^* on a finite set, and hence the subsets recognised by a morphism from A^* to a finite monoid. This special case of Proposition 2.1 deserves to be stated explicitly.

Proposition 2.4 *A language L over A^* is recognised by a finite automaton if and only if there exists a morphism α from A^* to a* finite *monoid such that*

$$L = L\alpha\alpha^{-1} \ . \qquad \blacksquare$$

2.2.2 Kleene's Theorem

In Chapter I, we defined the finite automata over A under the assumption that the alphabet A is finite. This assumption is no longer made *a priori* for the *definition* of

a finite automaton over A^* but it is still necessary for Kleene's Theorem, which we can restate: *cf. Th. I.2.1*

Theorem 2.1 *If A is a finite alphabet,* $\operatorname{Rat} A^* = \operatorname{Rec} A^*$.

In § 2.2.4 below, we will explain the role played by the finiteness of the alphabet in this result. Beforehand, by going back and forth with morphisms between arbitrary monoids and free monoids, we shall obtain some important results about recognisable sets.

Proposition 2.5 [**McKnight**] *Every recognisable subset of a finitely generated monoid is rational; that is, if M is finitely generated, then* $\operatorname{Rec} M \subseteq \operatorname{Rat} M$.

Proof. Let $\alpha\colon A^* \to M$ be a surjective morphism; A is (or can be chosen to be) finite if M is finitely generated. Let $P \in \operatorname{Rec} M$; $P\alpha^{-1}$ is recognisable in A^* (Corollary 2.2) hence rational, and $(P\alpha^{-1})\alpha$ is rational in M by Proposition 1.7 (i). ■ *cf. p. 223*

Proposition 2.6 *The intersection of a rational subset and a recognisable subset is rational.*

Proof. Let $P \in \operatorname{Rec} M$ and $R \in \operatorname{Rat} M$; R is included in a finitely generated submonoid M' of M (Corollary 1.5), $P \cap M'$ is recognisable in M' (Property 2.1) and $P \cap R = (P \cap M') \cap R$: we can assume without loss of generality that M is finitely generated. *cf. also Prop. 3.10, p. 268*

If $\alpha\colon A^* \to M$ is a surjective morphism, there exists L, a rational over A^*, such that $L\alpha = R$ (Proposition 1.7 (ii)). We have *cf. Exer. 0.1.3*

$$P \cap R = (P\alpha^{-1} \cap L)\alpha \ ,$$

which completes the proof since $P\alpha^{-1}$ is recognisable in A^* hence also $P\alpha^{-1} \cap L$, which is also rational and we apply Proposition 1.7 (i) once more. ■

Remark 2.3 Proposition 2.5 can also be seen as a direct corollary of Proposition 2.6: a finitely generated monoid M is a rational subset of itself and an arbitrary subset of M is equal to its intersection with M. □

2.2.3 Matrix representation of automata over A^*

The black square that ends Proposition 2.4 tells the reader that the proof of this statement, and hence the *construction*, from an automaton $\mathcal{A}$, of a morphism α that recognises $L(\mathcal{A})$, is the natural and immediate consequence of the reasoning that precedes it. This construction is actually quite a performance: starting from $\mathcal{A}$ we go to an equivalent deterministic automaton, then compute the transition monoid of this automaton seen as an action. It is therefore interesting to note that we can obtain the result directly, and accordingly even more so that we now introduce a notion that will be used constantly: the *representation* of a (finite) automaton over a free monoid *by matrices*.

Definition 2.2 Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be a finite automaton over A; the *(boolean) matrix representation* of $\mathcal{A}$ is the triple (λ, μ, ν) where

$$\mu \colon A^* \to \mathbb{B}^{Q \times Q}$$

is the morphism defined by

$$\forall a \in A \qquad a\mu_{p,q} = \begin{cases} 1 & \text{if} \quad (p, a, q) \in E \\ 0 & \text{otherwise} \end{cases} , \tag{2.6}$$

and where $\lambda \in \mathbb{B}^{Q \times 1}$ and $\nu \in \mathbb{B}^{1 \times Q}$ are the two (boolean) vectors, respectively row and column, of dimension Q, defined by

$$\lambda_q = \begin{cases} 1 & \text{if} \quad q \in I \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \nu_p = \begin{cases} 1 & \text{if} \quad p \in T \\ 0 & \text{otherwise} \end{cases} . \qquad \square$$

The morphism μ warrants being called a *representation of the automaton* $\mathcal{A}$ since, for any word f, the entry (p, q) of a matrix $f\mu$ is 1 if and only if there exists in $\mathcal{A}$ a computation from p to q labelled f; that is:

Property 2.4

$$\forall f \in A^* \qquad f\mu_{p,q} = \begin{cases} 1 & \textit{if} \quad p \xrightarrow[\mathcal{A}]{f} q \\ 0 & \textit{otherwise} \end{cases} . \tag{2.7}$$

Proof. By induction on the length of words f. By definition (that is, by (2.6)), (2.7) is true for $|f| = 1$ (and also $|f| = 0$). As

$$(f\,a)\mu_{p,q} = \sum_{r \in Q} (f\mu_{p,r})\,(a\mu_{r,q}) \ ,$$

we indeed have $(f\,a)\mu_{p,q} = 1$ if and only if there exists r in Q such that

$$p \xrightarrow[\mathcal{A}]{f} r \qquad \text{and} \qquad r \xrightarrow[\mathcal{A}]{a} q \quad . \qquad \blacksquare$$

From this we deduce:

Property 2.5 $\qquad L(\mathcal{A}) = \{ f \in A^* \mid \lambda \cdot f\mu \cdot \nu = 1 \}$. $\blacksquare$

Example 2.3 (continued) The representation $(\lambda_1, \mu_1, \nu_1)$ of the automaton $\mathcal{A}_1$ (Figure 2.3) which recognises $L_1 = A^* a\,b\,A^*$ is

$$\lambda_1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \ , \quad a\mu_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \ , \quad b\mu_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \ , \quad \nu_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \ .$$

We compute for example:

$$(a\,b\,a\,b)\mu_1 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

from which $\lambda_1 \cdot (a\,b\,a\,b)\mu_1 \cdot \nu_1 = 1$. $\square$

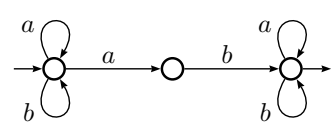


Fig. 2.3: The automaton $\mathcal{A}_1$

It is clear that if $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is *deterministic* the matrix representation which we have just defined coincides with the morphism of A^* in the row-monomial boolean matrices of dimension Q that correspond to the action of A^* on Q defined by $\mathcal{A}$. The following definition is thus natural, and consistent with that of the transition monoid of an action.

Definition 2.3 Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton and (λ, μ, ν) its matrix representation. The *transition monoid* of $\mathcal{A}$ is the image of A^* under μ. □

The definition of the matrix representation of an automaton and the properties that we have just proved enable us to give a self-contained proof of Proposition 2.4.

Proof of Proposition 2.4. [Alternative] Using the above notation, set

$$R = \{r \in \mathbb{B}^{Q \times Q} \mid \lambda \cdot r \cdot \nu = 1\} \ .$$

It follows that

$$f \in L(\mathcal{A}) \quad \Longleftrightarrow \quad \lambda \cdot f\mu \cdot \nu = 1 \quad \Longleftrightarrow \quad \exists r \in R \quad f\mu = r \ ,$$

or $L(\mathcal{A}) = R\mu^{-1}$.

Conversely, let L be a subset of A^* recognised by a morphism $\alpha \colon A^* \to N$ (where N is a finite monoid) and let S the subset of N such that $S = L\alpha$. We then construct the automaton $\mathcal{N}$ over A:

$$\mathcal{N} = \langle N, A, E, \{1_N\}, S \rangle \ ,$$

where

$$E = \{(n, a, n') \mid a \in A \ n, n' \in N \quad n' = n(a\alpha)\} \ .$$

By induction on the length of words we have, for all f in A^*, $1_N \xrightarrow[\mathcal{N}]{f} f\alpha$, and hence $L(\mathcal{N}) = \{f \in A^* \mid f\alpha \in S\} = S\alpha^{-1} = L$. ■

Remark 2.4 The automaton $\mathcal{N}$ above is *deterministic*: for each n in N, and each a in A, there exists a unique $n' = n\,(a\alpha)$ such that (n, a, n') is in E. This demonstration is thus also a(nother) proof that every automaton over A is equivalent to a deterministic automaton: let $\mathcal{A}$ be an arbitrary automaton over A; its behaviour $|\mathcal{A}|$ is recognised by its transition monoid N from which we construct the deterministic automaton $\mathcal{N}$ that accepts $|\mathcal{A}|$. □

2.2.4 Automaton of an action

We return to the problem of Proposition 2.5 and show that every *action* of M on a set can be viewed as an *automaton* on M.

Property 2.6 *Let δ be an action of M on S, with s_0 the distinguished element of S and R contained in S. The automaton $\mathcal{D}_\delta = \langle S, M, F, s_0, R \rangle$ defined by the transition set*

$$F = \{(s, m, s \cdot m) \mid s \in S \,,\ m \in M\}$$

accepts the subset P recognised by R (and by δ).

Proof. By construction,

$$s_0 \xrightarrow{m} s = s_0 \cdot m \quad , \tag{$*$}$$

is a transition of $\mathcal{D}_\delta$ for all m in M. If s is in R, the computation $(*)$ is successful and its label m is in $|\mathcal{D}_\delta|$. Conversely, let

$$s_0 \xrightarrow{m_1} s_1 \xrightarrow{m_2} s_2 \cdots s_{k-1} \xrightarrow{m_k} s_k \quad , \tag{$**$}$$

be a computation in $\mathcal{D}_\delta$. Its label is $m = m_1 m_2 \cdots m_k$ and we verify, by induction on k, that, since δ is an action,

$$s_k = s_0 \cdot m = s_0 \cdot (m_1 m_2 \cdots m_k) = (\cdots((s_0 \cdot m_1) \cdot m_2) \cdots) \cdot m_k \quad .$$

If m is in $|\mathcal{D}_\delta|$ – that is, if the computation $(**)$ is successful – then $s_k = s_0 \cdot m$ is in R and hence m is in P. ■

The automaton that we have just constructed is completely 'tied up' in the sense that we never have to follow a computation (of length greater than 1) to read a label: there is already a transition that bears that label. A moment's thought shows that we can define, still based on the action δ, an automaton that is more like our idea of what an automaton should be. This is the generalisation of the construction made from the action defined by a morphism of a free monoid in a finite monoid, for the proof of Proposition 2.4.

Property 2.7 *Let G be a generating set of M and δ an action of M on S as above. The automaton $\mathcal{D}_\delta^{(G)}$ obtained from $\mathcal{D}_\delta$ by retaining only the transitions labelled by elements of G:*

$$\mathcal{D}_\delta^{(G)} = \langle\, S, M, F', s_0, R \,\rangle \quad \textit{with} \quad F' = \{(s, g, s \cdot g) \mid s \in S\,,\ g \in G\} \quad ,$$

is equivalent to $\mathcal{D}_\delta$.

Proof. If G is a generating set, every m in M can be written $m = g_1 g_2 \cdots g_n$ (possibly not uniquely) and as δ is an action we have

$$s_0 \cdot m = (\cdots((s_0 \cdot g_1) \cdot g_2) \cdots) \cdot g_n = s$$

and the *transition* $s_0 \xrightarrow[\mathcal{D}_\delta]{m} s$ is replaced by the *computation* $s_0 \xrightarrow[\mathcal{D}_\delta^{(G)}]{g_1 g_2 \cdots g_n} s$. ■

We will say that $\mathcal{D}_\delta$, or $\mathcal{D}_\delta^{(G)}$ as the case may be, is the *automaton of the action* δ. And we note that Property 2.7 is *itself* a proof of Proposition 2.5: if M is finitely generated, we can choose G to be a finite set, *the automaton $\mathcal{D}_\delta^{(G)}$ is finite*, and the subset that it recognises is rational.

Example 2.4 (continued) If we apply the above constructions to the action δ_2, we obtain the two automata of Figure 2.4. □

The definition of the automaton of an action and its properties naturally suggest a statement which contains Kleene's Theorem.

Proposition 2.7 *Let M be a monoid.*

(i) *If M is finitely generated, then* $\operatorname{Rec} M \subseteq \operatorname{Rat} M$.

(ii) *If M is free, then* $\operatorname{Rat} M \subseteq \operatorname{Rec} M$.

Proof. Assertion (i) is a restatement of Proposition 2.5 and a paraphrase of Properties 2.6 and 2.7 (whose proof is independent of Kleene's Theorem). The proof of (ii) begins with the application of Theorem 1.1 that tells that a rational set of M is the behaviour of a finite automaton over M (which is thus understood to be the core of Kleene's Theorem). The sequel, and as M is a free monoid A^*, consists of the determinisation of an automaton over A^*, once the transition labels have been turned into elements of A. ■

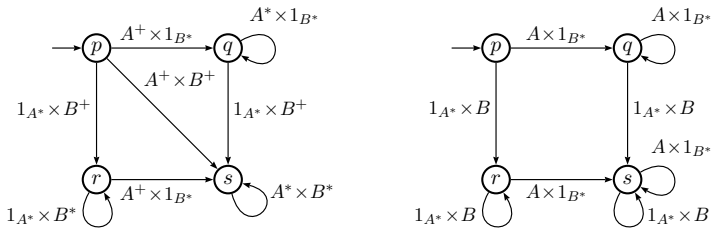


Figure 2.4: The two automata $\mathcal{D}_{\delta_2}$ and $\mathcal{D}_{\delta_2}^{(C)}$

Exercises

• 2.8 Compute the transition monoid of the automaton $\mathcal{A}_1$.

• 2.9 (a) Let $\mathcal{A}$ be a deterministic automaton with n states. What is the maximum size of its transition monoid? Can this bound be reached?

(b) Do the same for an automaton $\mathcal{A}$ that is not necessarily deterministic.

2.10 In Remark 2.4, we noted that Proposition 2.4 essentially contains the proof that every (finite) automaton over A is equivalent to a (finite) deterministic automaton over A. Compare the complexity (in terms of the number of states of the deterministic automaton constructed) of this method with that of the subset construction.

• 2.11 Show that every rational language is recognised by a *left regular deterministic* automaton: that is, by an automaton $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ such that *cf. Exer. I.8.1, p. 175.*

$$\forall f, g \in A^* \qquad i \cdot f = i \cdot g \quad \Longrightarrow \quad [\, \forall u \in A^* \quad i \cdot u\, f = i \cdot u\, g \,] \ .$$

2.12 Show that if A is an infinite alphabet, then $\operatorname{Rec} A^* = \operatorname{Rat} \mathfrak{P}(A)$.

2.3 Elementary operations on recognisable subsets

We will show here not only several closure properties of recognisable sets but also those operations on actions to which these properties correspond. We can do the same *cf. Exer. 2.13* starting from the definition of subsets recognised by morphisms in finite monoids.

2.3.1 Boolean operations

Property 2.8 *The complement of a subset P of M recognised by an action of M on S is recognised by the same action.*

Proof. If $P = \{m \in M \mid s_0 \cdot m \in R\}$, then the complement of P in M is $Q = \{m \mid s_0 \cdot m \notin R\} = \{m \mid s_0 \cdot m \in (S \setminus R)\}$. ■

Property 2.9 *Let P and P' be two subsets of M recognised by two actions δ of M on S and δ' of M on S' respectively. Then $P \cap P'$, and $P \cup P'$, are recognised by the action $\delta \times \delta'$ of M on $S \times S'$ defined by*

$$\forall (s, s') \in S \times S', \ \forall m \in M \qquad (s, s') \underset{S\times S'}{\cdot} m = (s \underset{S}{\cdot} m, s' \underset{S'}{\cdot} m) \tag{2.8}$$

(and with (s_0, s'_0) chosen as the distinguished element).

Proof. We first verify that (2.8) indeed makes $\delta \times \delta'$ an action:

$$\begin{aligned}\Big((s, s') \underset{S\times S'}{\cdot} m\Big) \underset{S\times S'}{\cdot} m' &= \Big((s \underset{S}{\cdot} m, s' \underset{S'}{\cdot} m)\Big) \underset{S\times S'}{\cdot} m' = \Big((s \underset{S}{\cdot} m) \underset{S}{\cdot} m', (s' \underset{S'}{\cdot} m) \underset{S'}{\cdot} m'\Big) \\ &= (s \underset{S}{\cdot} m\, m', s' \underset{S'}{\cdot} m\, m') = (s, s') \underset{S\times S'}{\cdot} m\, m' \ .\end{aligned}$$

We then have:

$$\begin{aligned}P \cap P' &= \{m \in M \mid s_0 \underset{S}{\cdot} m \in R \quad \text{and} \quad s'_0 \underset{S'}{\cdot} m \in R'\} \\ &= \{m \in M \mid (s_0, s'_0) \underset{S\times S'}{\cdot} m \in R \times R'\} \ .\end{aligned}$$

cf. Exer. I.1.33, p. 82 We conclude for $P \cup P'$ either directly as above, or by invoking Property 2.8 and using de Morgan's Law: $P \cup P' = \complement\left[\complement P \cap \complement P'\right]$. ■

Remark 2.5 The trivial action of a monoid M on a set with one element ($s_0 \cdot m = s_0$ for all m) implies that, for all M, M itself, and the empty set, are two recognisable subsets. □

From these two properties and the above remark we deduce:

Corollary 2.8 $\operatorname{Rec} M$ *is a Boolean algebra, effective if M is finitely generated.*[9] ■

2.3.2 Inverse morphism

cf. p. 237 We have already seen (Corollary 2.2) that the family $\operatorname{Rec} M$ is closed under inverse morphism. We can give this property a more precise statement.

Property 2.10 *Let $\alpha \colon M \to N$ be a morphism and ε an action of N on S. Let Q be a subset of N recognised by ε. Then $Q\alpha^{-1}$ is recognised by the action δ of M on S defined*[10] *by*

$$\forall s \in S, \ \forall m \in M \qquad (s, m)\delta = (s, m\alpha)\varepsilon \ . \tag{2.9}$$

[9] More precisely, if multiplication in M is effective.

[10] Here is an example where writing an action as a dot would be ambiguous.

Proof. We first verify that (2.9) indeed makes δ an action:

$$\big((s,m)\delta, m'\big)\,\delta = \big((s,m\alpha)\varepsilon, m'\alpha\big)\,\varepsilon = (s, m\alpha\, m'\alpha)\varepsilon = (s, m\,m'\alpha)\varepsilon = (s, m\,m')\delta \ .$$

We then have $Q\alpha^{-1} = \{m \in M \mid (s_0, m\alpha)\varepsilon \in R\} = \{m \in M \mid (s_0, m)\delta \in R\}$. ■

2.3.3 Quotient

We define, in a arbitrary monoid M as in the free monoid, the left (or right) quotient of a subset by an element:

$$\forall m \in M\,,\ \forall P \subset M \qquad \begin{aligned} m^{-1}P &= \{n \in M \mid m\,n \in P\} \quad \text{and} \\ P\,m^{-1} &= \{n \in M \mid n\,m \in P\} \ . \end{aligned}$$

The difference from the case of the free monoid is that we can no longer speak of 'prefix' (or 'suffix') or say that the quotient of p by m is obtained by 'deleting' in p the prefix equal to m. The two operations extend additively:

$$\forall K, P \subset M \qquad \begin{aligned} K^{-1}P &= \bigcup_{m\in K} m^{-1}P = \{n \in M \mid \exists m \in K \quad m\,n \in P\} \ , \\ P\,K^{-1} &= \bigcup_{m\in K} P\,m^{-1} = \{n \in M \mid \exists m \in K \quad n\,m \in P\} \ . \end{aligned}$$

When we come to consider the quotients of recognisable subsets, it is easier to use the definition of recognition by morphism. We then have: *cf. also Exer. 2.14*

Property 2.11 *Let P in M be recognised by $\alpha\colon M \to N$. For all subsets K of M, $K^{-1}P$ (resp. $P\,K^{-1}$) is recognised by α.*

Proof. For all m in M and all Q included in N, we have

$$\big[(m\alpha)^{-1}Q\big]\,\alpha^{-1} = \{n \in M \mid m\alpha\, n\alpha \in Q\} = \{n \in M \mid m\,n \in Q\alpha^{-1}\} = m^{-1}(Q\alpha^{-1}) \ .$$

Thus, setting $Q = P\alpha$, we have, for all K in M,

$$\begin{aligned} [(K\alpha)^{-1}Q]\alpha^{-1} &= [\bigcup_{m\in K}((m\alpha)^{-1}Q)]\alpha^{-1} = \bigcup_{m\in K}([(m\alpha)^{-1}Q)]\alpha^{-1}) \\ &= \bigcup_{m\in K}(m^{-1}(Q\alpha^{-1})) = \quad K^{-1}P \ . \end{aligned}$$

An identical series of equalities holds for $P\,K^{-1}$. ■

Corollary 2.9 *The family* $\mathrm{Rec}\,M$ *is closed under left (and right) quotient by an arbitrary subset of M.* ■

Recall Remark I.1.6: if P is recognisable, we know that $K^{-1}P$ and $P\,K^{-1}$ are recognisable, but if K is not recursive we do not know how to give an effective expression for it. *cf. p. 64*

2.3.4 Morphism and product

The family of recognisable subsets *is not closed* under direct morphism.

Example 2.5 Let α be the canonical morphism from $A^* = \{a,b\}^*$ to $\mathbb{Z}$: $a\alpha = +1$ and $b\alpha = -1$. The singleton $\{1_{A^*}\}$ is recognisable in A^* but $\{1_{A^*}\}\alpha = \{0\}$ is not recognisable in $\mathbb{Z}$, since, for example, $\{0\}\alpha^{-1} = Z_1 = \{f \in A^* \mid |f|_a = |f|_b\}$ is not recognisable in A^* (and we apply Corollary 2.2). □

cf. also Prop. 6.1, p. 302

The family of recognisable subsets is not, in general, closed under product or star. It is an interesting problem to determine whether these closure properties do actually hold for certain classes of monoids.

cf. Exer. 2.29

Example 2.6 Let U be the monoid $\{u,x,z\}$ where u is the neutral element, and z the zero, with x such that $xx = z$ and $M = \{(u,0),(x,0)\} \cup z \times \mathbb{Z}$, a submonoid of $U \times \mathbb{Z}$. The singleton $(x,0)$ is recognisable in M but none of $(x,0)^2$, $(x,0)^*$ or $(x,0)^+$ is. □

See solution of Exer. 2.16 for source.

Exercises

- 2.13 Show directly the equivalents of Properties 2.8, 2.9 and 2.10 for subsets recognised by morphisms. That is:

  (a) The complement of a subset recognised by $\alpha\colon M \to N$ is also recognised by α.

  (b) Let P and P' be two subsets of M, recognised respectively by $\alpha\colon M \to N$ and $\alpha'\colon M \to N'$. Then the morphism $\alpha{\times}\alpha'\colon M \to N{\times}N'$ defined by $(m)\alpha{\times}\alpha' = (m\alpha, m\alpha')$ recognises $P \cap P'$ (and $P \cup P'$).

  (c) Let Q be a subset N recognised by $\beta\colon N \to S$; then $P = Q\alpha^{-1}$ is recognised by $\alpha\beta$.

- 2.14 Let P be a subset of M recognised by an action δ of M on S. Show that for every subset K of M, PK^{-1} is recognised by δ.

- 2.15 Show that the definition of the left quotient (resp. right quotient) in a monoid coincides in a group with multiplication on the left (resp. on the right) by the 'standard' inverse.

- 2.16 Verify the assertions of Example 2.6.

- 2.17 Let M be an equidivisible monoid, m be in M and P and Q be two (arbitrary) subsets of M. Prove: $m^{-1}(PQ) = (m^{-1}P)Q \cup (P^{-1}m)^{-1}Q$ and $m^{-1}P^+ = \left[(P^{*-1}m)^{-1}P\right] P^*$.

2.4 Minimisation

We have seen that we can associate a *minimal automaton* with each deterministic automaton. It is natural that it should be the same for actions. The presentation of the minimisation process for actions makes clear its *algebraic nature*[11] and allows us to clarify the respective position of a certain number of notions and results (quotients, minimal actions, syntactic monoids, *etc.*).

[11] This process is an instance of the 'First Theorem of Isomorphism', *cf.* N. Bourbaki, *Algèbre* [37, Chap. I].

This subsection is awash with universal algebra and the proofs, which are mechanical, and for the most part left as exercises. However, its result is a double *characterisation* of recognisable sets which will be frequently used thereafter and which we can state in anticipation.[12]

Theorem 2.2 *Let A be a finite alphabet and α a surjective morphism from A^* to a monoid N. A subset Q of N is recognisable if and only if $Q\alpha^{-1}$ is a rational subset of A^*.*

Theorem 2.3 [**Myhill–Nerode**] *Let P be a subset of a monoid M. The following conditions are equivalent:*[13]

(i) *the set of left quotients of P is finite;*

(ii) *the set of right quotients of P is finite;*

(iii) *the syntactic monoid of P is finite;*

(iv) *the subset P is recognisable.*

What follows is not the most rapid proof of these statements since we choose to put in place some other equally useful results.

2.4.1 Action morphisms

Let δ and η be two actions of M on S and T respectively. An *(action) morphism* of δ on η is a map φ from S to T which commutes with the effect of the action and respects the choice of the distinguished element:

$$\forall s \in S\,,\ \forall m \in M \qquad (s \underset{S}{\cdot} m)\varphi = s\varphi \underset{T}{\cdot} m \quad \text{and} \quad s_0\varphi = t_0\ .$$

$$\begin{tikzcd} S \arrow[r, "m\check{\delta}"] \arrow[d, "\varphi"'] & S \arrow[d, "\varphi"] \\ T \arrow[r, "m\check{\eta}"] & T \end{tikzcd}$$

That is, the adjacent diagram commutes for all m in M. The actions of M and their morphisms form a *category*.[14] We verify in particular:

Properties 2.12

(i) *For every action δ of M on S, there exists a (unique) map σ from M to S which is an action morphism from the regular representation of M on itself to δ (and which we call the* canonical morphism *of δ).*

(ii) *Let δ and η be two actions from M to S and T and let σ and τ be their two canonical morphisms. If φ is an action morphism from δ to η, then $\tau = \sigma\varphi$.* ■

This property is illustrated by the commutative diagram below – where the multiplication (on the right) by m is written $m\check{\iota}$ for consistency (*cf.* Example 2.1 (b)).

[12] Even if the syntactic monoid of a subset has not yet been defined.

[13] This is the *definition* of recognisable subsets in Eilenberg, *op. cit.* (*cf.* Note 8, p. 235).

[14] I do not think that category theory helps much, at least not with our problems. I have therefore not given the relevant definitions. For those who know them already, I have also given the statements in categorical form. For an introduction to category theory, see [171, 167].

It amounts to saying[15] that *the regular representation of M on itself is the initial object of the category of all the actions of M on sets.*

In fact, as we said for morphisms and congruences (of monoids), the actions and right regular equivalences are one and the same thing considered from two different points of view:

$$\begin{tikzcd}
S \arrow[rr, "\check{m\delta}"] \arrow[dd, "\varphi"'] & & S \arrow[dd, "\varphi"] \\
& M \arrow[ul, "\sigma"'] \arrow[dl, "\tau"] \arrow[r, "\check{m\iota}"] & M \arrow[ur, "\sigma"'] \arrow[dr, "\tau"] & \\
T \arrow[rr, "\check{m\eta}"'] & & T
\end{tikzcd}$$

Property 2.13 *The map equivalence of the canonical morphism of M's regular representation on an arbitrary action of M is a right regular equivalence of M. Conversely, every right regular equivalence of M canonically induces an action of M on the quotient set of classes.*

Proof. Suppose, with the notation of Properties 2.12, that p and q, elements of M, are such that $p\sigma = q\sigma$, and let m be an arbitrary element of M. From $(p\sigma) \underset{S}{\cdot} m = (q\sigma) \underset{S}{\cdot} m$ we deduce $(p \underset{M}{\cdot} m)\sigma = (q \underset{M}{\cdot} m)\sigma$; that is, $(p\,m)\sigma = (q\,m)\sigma$; in other words, σ is right regular.

Now let α be a right regular equivalence relation on M. Write M/α for the quotient set, and $[p]_\alpha$ for the class of p modulo α. The action of M on M/α is defined by

$$[p]_\alpha \cdot m = [p\,m]_\alpha \ .$$

This equation indeed defines a map on M/α because α is right regular. This map is an action by associativity in M. ■

2.4.2 Minimal action

Let P be a fixed set in M. Recall that we write $m^{-1}P$ for the *(left) quotient* of P by m:

$$m^{-1}P = \{p \mid mp \in P\} \ .$$

The monoid M acts naturally (on the right) on the set Q_P of (left) quotients of P:

$$Q_P = \{m^{-1}P \mid m \in M\} \ .$$

For all m and n in M, we define $(m^{-1}P) \cdot n$ by

$$(m^{-1}P) \cdot n = n^{-1}(m^{-1}P) = (m\,n)^{-1}P \ .$$

It is indeed an action, which we will write ρ_P since, for all m, n and n' in M

$$(m^{-1}P \cdot n) \cdot n' = ((m\,n)^{-1}P) \cdot n' = (m\,n\,n')^{-1}P = (m^{-1}P) \cdot n\,n' \ .$$

The distinguished element of ρ_P is $q_0 = (1_M)^{-1}P = P$ and ρ_P recognises P: setting $R = \{m^{-1}P \mid 1_M \in m^{-1}P\}$, we obtain

$$m \in P \quad \Longleftrightarrow \quad 1_M \in m^{-1}P \quad \Longleftrightarrow \quad q_0 \cdot m \in R \ .$$

[15] A pedantic but concise statement.

We also write ρ_P for the relation on M defined by

$$m \, \rho_P \, m' \quad \Longleftrightarrow \quad m^{-1}P = m'^{-1}P \ ,$$

and we easily prove:

Property 2.14 *The relation ρ_P is the coarsest right regular equivalence on M that saturates P.* ■

There follows:

Proposition 2.10

(i) *For all actions δ of M on S which recognise P, there exists a unique action morphism of δ on ρ_P.*

(ii) *Let δ and η be two actions of M on S and T respectively which recognise P, and let σ and τ be the two morphisms of δ and η on ρ_P defined above. If φ is an action morphism of δ on η, then $\sigma = \tau\varphi$.* ■

In other words,[16] *the action ρ_P is the terminal object of the category of actions of M which recognise P*, which is shown by the adjacent commutative diagram.

$$\begin{tikzcd}
S \arrow[rr, "m\check{\delta}"] \arrow[dr, "\sigma"] \arrow[dd, "\varphi"'] & & S \arrow[dl, "\sigma"'] \arrow[dd, "\varphi"] \\
& M/\rho_P \arrow[r, "m\check{\iota}"] & M/\rho_P \\
T \arrow[ur, "\tau"] \arrow[rr, "m\check{\eta}"] & & T \arrow[ul, "\tau"']
\end{tikzcd}$$

2.4.3 Syntactic congruence and monoid

We can give a 'recognition by morphisms' version of what we have just proved for actions that recognise a given subset P. Therefore let P be once again a fixed subset of M. A pair (u, v) in $M \times M$ is a *context* of m (in M) for P if $u\,m\,v$ is in P; we write $\mathsf{C}_P(m)$ for the set of contexts of m for P:

$$\mathsf{C}_P(m) = \{(u, v) \mid u\,m\,v \in P\} \ ,$$

and σ_P for the relation on M defined by

$$m \, \sigma_P \, m' \quad \Longleftrightarrow \quad \mathsf{C}_P(m) = \mathsf{C}_P(m') \ .$$

Property 2.15

The relation σ_P is the coarsest congruence on M that saturates P. ■

We call σ_P the *syntactic congruence* of P, and the monoid M/σ_P, the quotient of M by σ_P, also written M_P, is called the *syntactic monoid* of P.

Proposition 2.11 *Let P be a subset of M and σ the canonical morphism from M to M/σ_P. If $\alpha\colon M \to N$ is a surjective morphism which recognises P, there exists a unique morphism $\theta\colon N \to M/\sigma_P$ such that $\sigma = \alpha\theta$ and that θ recognises $P\alpha$.*[17]

[16] A concise but pedantic statement.

[17] That is, the morphism $\sigma\colon M \to M/\sigma_P$ is the *terminal object* of the category of morphisms $\alpha\colon M \to N$ that recognise P.

Proof. The congruence $\alpha\alpha^{-1}$ saturates P; it is therefore coarser than σ_P. For each element n of N we find a corresponding class of σ_P that contains $n\alpha^{-1}$. This map from N to M/σ_P, which we write θ, is such that $\sigma = \alpha\theta$. It follows that θ is a morphism:

$$\begin{tikzcd} M \arrow[r, "\alpha"] \arrow[d, "\sigma_P"'] & N \arrow[dl, dashed, "\theta"] \\ M/\sigma_P & \end{tikzcd}$$

$$(n\,n')\theta = (m\,m')\alpha\theta = (m\,m')\sigma = m\sigma\,m'\sigma = m\alpha\theta\,m'\alpha\theta = n\theta\,n'\theta \ ,$$

since α is surjective, and that θ recognises $P\alpha$, since

$$P\alpha = P\sigma\sigma^{-1}\alpha = P\alpha\theta\theta^{-1}\alpha^{-1}\alpha = (P\alpha\theta)\theta^{-1} \ . \qquad \blacksquare$$

cf. p. 237

Hence we immediately obtain the promised converse of Corollary 2.2:

Corollary 2.12 *Let $\alpha\colon M \to N$ be a* surjective *morphism and Q a subset of N. If $Q\alpha^{-1}$ is a recognisable subset of M, then Q is a recognisable subset of N.* ■

Proof of Theorem 2.2. Theorem 2.2 is the instance of this corollary when $M = A^*$, via Kleene's Theorem, and characterises the recognisable subsets of finitely generated monoids. ■

Remark 2.6 Armed with Proposition 2.11 and following reasonably wide usage, we will say that a property or construction on a subset P of a monoid *is syntactic* if, when P is recognised by a surjective morphism α, the same property or construction holds for $P\alpha$.

For example, taking the quotient (left or right) by an element or an arbitrary subset is a syntactic operation (Property 2.11). □

Remark 2.7 Imitating the definition of $\mathsf{C}_P(m)$, we write that the left quotient $m^{-1}P$ is also the set of *right contexts* of m for P, which we can write $\mathsf{RC}_P(m)$. □

The link between the minimal action and the syntactic monoid is proven by the following proposition:

Proposition 2.13 *Let P be a subset of M. The syntactic monoid of P, M/σ_P, is the transition monoid of the action ρ_P (that is, the transition monoid of the canonical action of M on M/ρ_P).* ■

Proof of Theorem 2.3. We must also note that we can define a *left* action of M on the set of *right* quotients of P, that this action is the minimal left action that recognises P, and that the syntactic monoid M/σ_P is also the transition monoid of this action (since its definition has left–right symmetry). We have thus proved Theorem 2.3. ■

cf. Sec. 3.2, p. 261 and Exer. 3.9, p. 264

Anticipating the next section, we can also state the following result, which gives an effective procedure for computing the minimal action of a recognisable subset.

Proposition 2.14 *Let G be a finite generating set of M. Let δ be an action of M on a finite set S which recognises P, effectively described by the automaton $\mathcal{D}_\delta^{(G)}$. Then the minimal action ρ_P is obtained by applying Moore's algorithm to $\mathcal{D}_\delta^{(G)}$.* *cf. Propty 2.7, p. 242*

Exercises

2.18 Let δ be an action of M on S and let the associated morphism (of monoids) be $\tau\colon M \to \mathfrak{T}_S$. Show that there exists an action morphism of the action defined by τ on δ.

• 2.19 Verify Property 2.12.

• 2.20 Verify Property 2.14.

2.21 Prove Proposition 2.10.

• 2.22 Verify Property 2.15.

• 2.23 Verify Proposition 2.13.

• 2.24 **Disjunctive subset.** We say that a subset P of a monoid M is *disjunctive* if the syntactic congruence P is the identity.

(a) Verify that, for all u and v in M and all subsets P of M,

$$(u^{-1}P)\,v^{-1} = u^{-1}(P\,v^{-1})$$

which we will therefore write $u^{-1}P\,v^{-1}$ and call the *bi-quotient*.

(b) Show that P is disjunctive if and only if the Boolean algebra generated by the bi-quotients of P is the whole of $\mathfrak{P}(M)$.

(c) Give an example of a monoid which contains no disjunctive subset.

2.25 Show that Pal is a disjunctive subset of A^*. What more can you say?

2.5 Algebra at work

The characterisation of recognisable subsets, and hence of rational languages, as the inverse images of morphisms in finite monoids gives a powerful method of investigation and proof. It opens up the theory of rational languages, and hence of automata, to the results and methods of the theory of finite monoids.

We thus discover an entire area of automata theory that can be developed in two directions: first by the development of techniques for establishing the rationality (indeed, the recognisability) of certain languages, and second by finding correspondences between combinatorial operations on languages and their effect on the algebraic structure of the corresponding syntactic monoids. The latter theme has been explored thoroughly in other works and will not be tackled here. *cf. bibliographic notes*

2.5.1 Two examples

The properties that we mentioned at the end of the last chapter come to a great extent from the techniques derived from recognition by a finite monoid. For example, *cf. Sec. I.8, p. 175*

consider this: for every integer k and every language L over A^*, write $\mathcal{R}_k(L)$ for the set of words whose kth power is in L; that is, $\mathcal{R}_k(L)$ is a sort of 'kth root of L', and we have:

Proposition 2.15 *If L is rational, $\mathcal{R}_k(L)$ is rational.*

Compare this with the solutions of Exer. I.8.1, p. 205.

Proof. Let $\alpha\colon A^* \to M$ be a morphism that recognises L (with M finite); that is, there exists a subset P of M such that $L = P\alpha^{-1}$, and $\mathcal{R}_k(P) = \{m \in M \mid m^k \in P\}$. We verify trivially that $\mathcal{R}_k(L) = (\mathcal{R}_k(P))\,\alpha^{-1}$. ■

We also note that this proof yields an effective computation of $\mathcal{R}_k(L)$. We now give another example.

Definition 2.4 A *substitution* is a relation σ from A^* to M such that the map

$$\sigma\colon A^* \to \mathfrak{P}(M) \ ,$$

is a (monoid) morphism; that is, σ is completely defined by the $a\sigma$, the subsets of M for all a in A, and by the fact that $(f\,g)\sigma = (f\sigma)\,(g\sigma)$. □

Proposition 2.16 *The inverse image of a recognisable subset of M in an arbitrary substitution of a free monoid in M is a rational language.*

Proof. Let $\sigma\colon A^* \to M$ be a substitution and let $\alpha\colon M \to N$ be a morphism in a finite monoid that recognises a subset P of M: $P = Q\alpha^{-1}$. We have

$$\begin{aligned} P\sigma^{-1} &= \{f \in A^* \mid f\sigma \cap P \neq \emptyset\} = \{f \in A^* \mid f\sigma \cap Q\alpha^{-1} \neq \emptyset\} \\ &= \{f \in A^* \mid f\sigma\alpha \cap Q \neq \emptyset\} \ . \end{aligned}$$

Let W be the subset of $\mathfrak{P}(N)$; that is, the family of subsets of N, defined by

$$W = \{V \in \mathfrak{P}(N) \mid V \cap Q \neq \emptyset\} \ .$$

cf. Lem. 1.1, p. 221

The morphism α extends to a morphism $\alpha\colon \mathfrak{P}(M) \to \mathfrak{P}(N)$ (morphism of multiplicative monoids) and it follows that

$$L = P\sigma^{-1} = W(\sigma\alpha)^{-1} \ ,$$

where $\sigma\alpha\colon A^* \to \mathfrak{P}(N)$ is a morphism of monoids. Thus L is recognised by $\sigma\alpha$ and is recognisable, hence rational, since $\mathfrak{P}(N)$ is finite when N is. ■

The specialisation to free monoids of Proposition 2.16 gives:

Corollary 2.17 *The inverse image of a rational language in an arbitrary substitution on a free monoid in another set is a rational language.* ■

What adds to the interest of the corollary (or of the proposition) is that the substitution σ is arbitrary; each $a\sigma$ can easily be non-rational (or non-recognisable), can even be non-recursive, in which case we fall back into the paradox already mentioned: we know that $L\sigma^{-1}$ is rational but we do not necessarily know how to decide whether it is empty.

cf. Rem. I.1.6, p. 64

Proposition 2.16 has another interesting consequence to which we shall return in Section 4.

Corollary 2.18 *Let $\sigma\colon A^* \to M$ be an (arbitrary) substitution and P a recognisable subset of M. Then the set L of words of A^* whose image under σ is* contained *in P:* *cf. Th. 4.5, p. 291*

$$L = \{f \in A^* \mid f\sigma \subseteq P\} \ ,$$

is rational.

Proof. We verify easily that $L = \complement_{A^*}\left[\left(\complement_M P\right)\sigma^{-1}\right]$. ■

2.5.2 Recognisable subsets included in a product

The following property is rather technical. It and its immediate corollaries will be used so often that they may aptly be described as fundamental. It will also appear as the corollary of the results in Section 4, but it is simpler to give a direct presentation of it immediately. *cf. Exer. 4.5*

Proposition 2.19 *Let P be a recognisable subset of M contained in the product XY of two subsets X and Y of M. There exists a finite set I of terms $R_i S_i$ where R_i is included in X and S_i is included in Y such that*

$$P = \bigcup_{i\in I} R_i S_i \ .$$

If X and Y are recognisable, then so are the R_i and the S_i.

Proof. Let $\alpha\colon M \to H$ be a morphism which recognises P with H *finite*.[18] Write

$$V = X\alpha \ , \quad W = Y\alpha \qquad \text{and} \qquad Q = P\alpha \ .$$

Since Q is finite and $P = \bigcup_{q\in Q} q\alpha^{-1}$, we can assume without restriction that Q is reduced to a singleton: $Q = \{q\}$. Let I be the set, which is finite since H is, defined by

$$I = \{(v,w) \mid v \in V \ , \quad w \in W \quad vw = q\} \ .$$

For each (v,w) in I, let

$$R_{(v,w)} = X \cap v\alpha^{-1} \qquad \text{and} \qquad S_{(v,w)} = Y \cap w\alpha^{-1} \ . \tag{2.10}$$

It remains to show that

$$P = q\alpha^{-1} = \bigcup_{(v,w)\in I} R_{(v,w)}\, S_{(v,w)} \ .$$

The inclusion from right to left is immediate: $(R_{(v,w)})\alpha = v$, $(S_{(v,w)})\alpha = w$ and $vw = q$. In the other direction, every element p of P can, by the hypothesis, be written $p = xy$ with x in X and y in Y; hence p belongs to $R_{(x\alpha,y\alpha)}\, S_{(x\alpha,y\alpha)}$.

If X and Y are recognisable, then the R_i and S_i, defined by (2.10), are recognisable by virtue of Corollaries 2.8 and 2.2. ■

Remark 2.8 If $\operatorname{Rec} M$ is closed under product, the (necessary) condition given in Proposition 2.19 for P to be recognisable in M is obviously sufficient. □

[18] We do not use N here because of Corollary 2.20, and there are not many letters left.

cf. Exam. 2.4, p. 234

Let $M = A^* \times B^*$; since $A^* \times \{1_{B^*}\}$ and $\{1_{A^*}\} \times B^*$ are recognisable and $M = (A^* \times \{1_{B^*}\})(\{1_{A^*}\} \times B^*)$, we conclude from Proposition 2.19 that *a recognisable subset of* $A^* \times B^*$ *is a finite union of direct products* $(R \times S)$ *where* R *is a recognisable subset of* A^* *and* S *is a recognisable subset of* B^*. In fact, this statement holds for all direct products $M \times N$ and we obtain the following even stronger property.

Corollary 2.20 *A subset* P *of a direct product* $M \times N$ *of two monoids is recognisable if and only if there exists a finite family* $\{T_i\}_{i \in I}$ *of recognisable subsets of* M *and a finite family* $\{U_i\}_{i \in I}$ *of recognisable subsets of* N *such that*

$$P = \bigcup_{i \in I} T_i \times U_i \quad .$$

Proof. The condition is necessary. We take $X = M \times 1_N$ and $Y = 1_M \times N$ and apply the previous proposition:

$$P = \bigcup_{i \in I} R_i \, S_i = \bigcup_{i \in I} [T_i \times 1_N] \; [1_M \times U_i] = \bigcup_{i \in I} T_i \times U_i \quad .$$

The subsets X and Y are not necessarily recognisable in $M \times N$ and we cannot use the last assertion of Proposition 2.19.

Let $\mu \colon M \to M \times N$ be the morphism defined by $m\mu = (m, 1_N)$. With the above notation, we have

$$R_{(v,w)} = M \times 1_N \cap v\alpha^{-1} \qquad \text{from which} \qquad T_{(v,w)} = (v)\,[\mu\alpha]^{-1}$$

and $T_{(v,w)}$, and similarly $U_{(v,w)}$, are recognisable.

The condition is sufficient. Let $\pi_M \colon M \times N \to M$ and $\pi_N \colon M \times N \to N$ be the two canonical projections. For all T included in M and all U included in N we have

$$T \times U = (T \times N) \cap (M \times U) = T\pi_M^{-1} \cap U\pi_N^{-1} \quad .$$

Thus, $T \times U$ is recognisable in $M \times N$ when T and U are recognisable in M and N respectively, and the conclusion follows. ■

Exercises

• 2.26 Write $\mathcal{I}(L)$ for the set of iterating factors of L. Show that $\mathcal{I}(L)$ is rational when L is (Proposition I.8.1).

cf. p. 177

2.27 Show that $\mathbb{Z} \times 1_{\mathbb{Z}}$ is not a recognisable subset of $\mathbb{Z} \times \mathbb{Z}$.

2.28 Show that $\operatorname{Rec} M \times N$ is closed under product if both $\operatorname{Rec} M$ and $\operatorname{Rec} N$ are.

cf. Sec. 0.4 and 0.6

• 2.29 Let M be an *equidivisible* monoid.

(a) Show that $\operatorname{Rec} M$ is closed under product.

(b) Show that $\operatorname{Rec} M$ is closed under $+$; that is, $P \in \operatorname{Rec} M \;\Rightarrow\; P^+ \in \operatorname{Rec} M$.

• 2.30 Let \$ be a symbol which, by convention, does not belong to any alphabet.[19] For each alphabet A, write $A_\$$ for the set $A \cup \{\$\}$. Let δ_3 be the minimal action of $A_\$^* \times B_\$^*$ which recognises the subset $A^*\$ \times B^*\$$. Compute $\mathcal{D}_{\delta_3}^{(C)}$, the automaton of δ_3. *cf.* § *2.2.4, p. 241*

3 Morphisms and coverings

We shall now define the *automata morphism*, an idea that is already implicit in the definition of a (deterministic) minimal automaton and of minimal action by the expedient of the action morphism, which works for automata over an arbitrary monoid.[20] A certain class of so-called *Out-surjective* morphisms allows us to generalise the definition of minimal objects to these automata. The notion of a *covering*, even more precise,[21] will show itself to be perfectly suited to several constructions on automata that we shall be led to make in the following chapters. We shall in particular shortly present (Subsection 3.4) the construction of a remarkable covering for automata over a free monoid which we will use several times to deal with weighted automata (Chap. III) and with functions realised by automata (Chap. V).

The choice of a term from topology – covering – is neither a coincidence nor a misappropriation. The idea that underlies this notion is to be able to deduce from the correspondence of the 'local' properties of two automata a correspondence between their 'global' behaviours.

For the rest of this section, $\mathcal{A} = \langle Q, M, E, I, T \rangle$ and $\mathcal{B} = \langle R, M, F, J, U \rangle$ are two automata over (the monoid) M.

3.1 Automata morphisms

We saw at the very start of the chapter that a transition of an automaton is fundamentally a triple (*source*, *label*, *destination*); that is, E is canonically equipped with three maps ι, ε and τ:

$$\iota\colon E \longrightarrow Q \ , \qquad \tau\colon E \longrightarrow Q \ , \qquad \text{and} \qquad \varepsilon\colon E \longrightarrow M \ ,$$

(in other words, the state $e\iota$ is the source of the transition e, $e\tau$ is its destination and $e\varepsilon$ its label), even if for convenience E is usually identified with a subset of $Q \times M \times Q$. We shall use this extremely formal point of view in the definitions that follow.

[19] The symbol that we shall call the *endmarker* in Chapter IV (Section 5.1).

[20] Even if, in this section, the examples and exercises relate solely to automata over free monoids (but not necessarily deterministic automata).

[21] It was originally defined for *graphs* by J. R. Stallings (in *Inventiones Math.*, 1983, [239]) and extends naturally, as we shall see, to automata.

3.1.1 Definitions and examples

A *morphism* φ from $\mathcal{B}$ to $\mathcal{A}$ is in fact a *pair of maps*, both written φ, one between the states and the other between the transitions of two automata:

$$\varphi\colon R \longrightarrow Q \qquad \text{and} \qquad \varphi\colon F \longrightarrow E \ ,$$

which satisfy the three properties

$$\varphi \circ \iota = \iota \circ \varphi \qquad \text{and} \qquad \varphi \circ \tau = \tau \circ \varphi \ , \tag{3.1}$$

$$\varphi \circ \varepsilon = \varepsilon \ , \tag{3.2}$$

$$J\varphi \subseteq I \qquad \text{and} \qquad U\varphi \subseteq T \ . \tag{3.3}$$

That is, the three following diagrams are commutative.[22]

$$\begin{tikzcd} F \arrow[r, "\varphi"] \arrow[d, "\iota"'] & E \arrow[d, "\iota"] \\ R \arrow[r, "\varphi"'] & Q \end{tikzcd} \qquad \begin{tikzcd} F \arrow[r, "\varphi"] \arrow[d, "\tau"'] & E \arrow[d, "\tau"] \\ R \arrow[r, "\varphi"'] & Q \end{tikzcd} \qquad \begin{tikzcd} F \arrow[rr, "\varphi"] \arrow[dr, "\varepsilon"'] & & E \arrow[dl, "\varepsilon"] \\ & M & \end{tikzcd}$$

The conditions (3.1) imply that *the image of a computation in $\mathcal{B}$ is a computation in $\mathcal{A}$*, or put another way, that φ is a *graph morphism*.[23] Condition (3.2) implies that *the label of a computation in $\mathcal{B}$ is equal to the label of the image of this computation in $\mathcal{A}$*, in other words, that φ is a *morphism on labelled graphs*. Finally, conditions (3.3), which make φ an *automata morphism*, imply that *the image of a successful computation in $\mathcal{B}$ is a successful computation in $\mathcal{A}$*. We can therefore in particular state:

Property 3.1 *If $\varphi\colon \mathcal{B} \to \mathcal{A}$ is a morphism, then $|\mathcal{B}| \subseteq |\mathcal{A}|$.* ■

Since the definition of a morphism is symmetric with respect to the source and destination of transitions, we have:

Property 3.2 *If $\varphi\colon \mathcal{B} \to \mathcal{A}$ is a morphism, $\varphi\colon \mathcal{B}^{\mathsf{t}} \to \mathcal{A}^{\mathsf{t}}$ is also a morphism.* ■

Examples 3.1 Figure 3.1 shows three pairs of automata, with the intuitive convention that the vertical direction is a morphism from each upper automaton to the corresponding lower automaton. □

Example 3.2 The standard construction of the *(direct) product* of two automata (over a free monoid A^*) gives a common and very useful example of automata mor-

cf. § I.1.2.2, p. 61

[22] The reader may be surprised that I should define $\varphi\colon \mathcal{B} \to \mathcal{A}$ and not the more natural $\varphi\colon \mathcal{A} \to \mathcal{B}$, and the first few subsections will only increase any discomfort. In most applications (of the notion of morphism) we start from an automaton $\mathcal{A}$ and construct an automaton $\mathcal{B}$ – this is the natural order – and the virtue of the construction will be the existence of a morphism from $\mathcal{B}$ to $\mathcal{A}$, hence our notation. Seek reassurance in Subsections 3.3, 3.4, *etc.*

[23] In the sense of Stallings, *art. cit.*

phisms. The product of $\mathcal{A} = \langle Q, A, E, I, T \rangle$ and $\mathcal{A}' = \langle Q', A, E', I', T' \rangle$ is by definition the automaton $\mathcal{A}\times\mathcal{A}' = \langle Q\times Q', A, G, I\times I', T\times T' \rangle$ where the set of transitions G is defined by

$$G = \{\big((p,r), a, (q,s)\big) \mid (p,a,q) \in E \ , \quad (r,a,s) \in E'\} \ .$$

The projections $\pi_{\mathcal{A}}$ and $\pi_{\mathcal{A}'}$ of the set $Q\times Q'$ on the first and second components respectively, combined with the corresponding maps from G to E and E', are morphisms of $\mathcal{A}\times\mathcal{A}'$ in $\mathcal{A}$ and in $\mathcal{A}'$ respectively. □

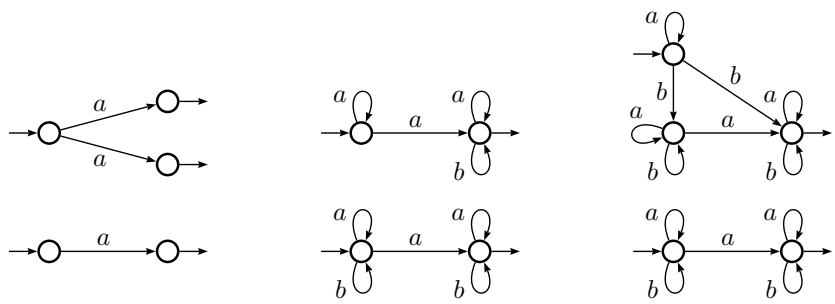


Figure 3.1: Three automata morphisms

Their definition itself shows that automata morphisms can be composed; its wording calls for some comments.

Property 3.3 *If $\varphi\colon \mathcal{B} \to \mathcal{A}$ and $\psi\colon \mathcal{C} \to \mathcal{B}$ are automata morphisms, then $\psi\varphi\colon \mathcal{C} \to \mathcal{A}$ is also an automata morphism.* ■

Remark 3.1 It seems that we should have used a simpler definition and said that a morphism $\varphi\colon \mathcal{B} \to \mathcal{A}$ is a map from R to Q such that if (r,a,s) is in F then $(r\varphi, a, s\varphi)$ is in E. However, φ is *also* and *first* a map from F to E, hence our choice of definition. It is vital to remember this in what follows. □

Remark 3.2 We might have imagined, especially after the proof of Theorem 1.1, that an automata morphism would open the way to a morphism between possibly different input monoids. This is obviously possible, but it seems more useful to have the separate notions of relabelling and of automata morphisms that do not change labels. □ *cf. p. 225*

In what follows we shall generally abbreviate 'automata morphism' to 'morphism'. When the conditions (3.3) are not satisfied, we shall explicitly say that φ is a 'labelled graph morphism', or a 'graph morphism' if condition (3.2) is not satisfied either.

Remark 3.3 Given a map $\varphi\colon \mathcal{B} \to \mathcal{A}$, we can ensure that the conditions (3.3) hold in a roundabout way, and also reduce the notion of automata morphism to that of a labelled graph morphism. To every automaton $\mathcal{A} = \langle Q, M, E, I, T \rangle$, we relate,

by a sort of normalisation, an automaton $\mathcal{A}_{\mathsf{n}}$ to which we have added two new states – $i_{\mathcal{A}}$, an initial state, and $t_{\mathcal{A}}$, a final state – and some transitions, *labelled with* 1_M, which go from $i_{\mathcal{A}}$ to each initial state of $\mathcal{A}$ and from each final state of $\mathcal{A}$ to $t_{\mathcal{A}}$:

$$\mathcal{A}_{\mathsf{n}} = \langle\, Q \cup \{i_{\mathcal{A}}, t_{\mathcal{A}}\}, M, E_{\mathsf{n}}, i_{\mathcal{A}}, t_{\mathcal{A}} \,\rangle \ ,$$
$$E_{\mathsf{n}} = E \cup \{(i_{\mathcal{A}}, 1_M, i) \mid \forall i \in I\} \cup \{(t, 1_M, t_{\mathcal{A}}) \mid \forall t \in T\} \ .$$

We verify easily that $\mathcal{A}_{\mathsf{n}}$ is equivalent to $\mathcal{A}$. If φ is a map from $\mathcal{B}$ to $\mathcal{A}$, we extend it to a map φ_{n} from $\mathcal{B}_{\mathsf{n}}$ to $\mathcal{A}_{\mathsf{n}}$ by taking

$$(i_{\mathcal{B}})\varphi_{\mathsf{n}} = i_{\mathcal{A}} \qquad \text{and} \qquad (t_{\mathcal{B}})\varphi_{\mathsf{n}} = t_{\mathcal{A}} \ .$$

We then verify, just as easily, that $\varphi\colon \mathcal{B} \to \mathcal{A}$ is an automata morphism if and only if $\varphi_{\mathsf{n}}\colon \mathcal{B}_{\mathsf{n}} \to \mathcal{A}_{\mathsf{n}}$ is a labelled graph morphism. □

3.1.2 Conformal morphisms

One of the subjects dealt with in this section is the study of the conditions under which the inclusion in Property 3.1 becomes an equality. The first and most general is almost a tautology; it uses the convention of the previous remark.

Definition 3.1 We will say that a morphism $\varphi\colon \mathcal{B} \to \mathcal{A}$ is *conformal* if every computation of $\mathcal{A}_{\mathsf{n}}$ is the image of at least one computation of $\mathcal{B}_{\mathsf{n}}$. □

As the example in Figure 3.2 shows, a morphism is not necessarily conformal, even if the two maps φ are surjective. The two following properties are immediate.

Property 3.4 *If $\varphi\colon \mathcal{B} \to \mathcal{A}$ is a conformal morphism, the behaviour of $\mathcal{B}$ is equal to that of $\mathcal{A}$: $|\mathcal{B}| = |\mathcal{A}|$.* ■

Property 3.5 *The composition of two conformal morphisms is a conformal morphism.* ■

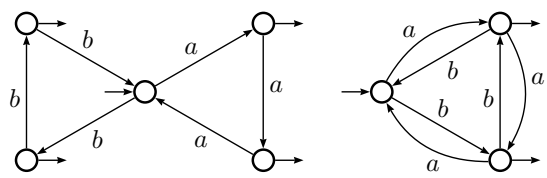


Figure 3.2: A non-conformal morphism (the morphism is the horizontal direction).

cf. Th. 8.2, p. 343 The notion of a conformal morphism will be used, in all its generality, in Section 8. We shall now look at some properties of morphisms that enable us to construct some particular conformal morphisms.

3.1.3 Local properties

To define the 'local' properties of automata mentioned at the start of the section, we must first determine what we understand by that. For each state q of $\mathcal{A}$, we write[24] $\mathsf{Out}_{\mathcal{A}}(q)$ the set of transitions of $\mathcal{A}$ whose source is q – that is, the transitions that 'leave' from q:

$$\mathsf{Out}_{\mathcal{A}}(q) = \{e \in E \mid e\iota = q\} \ ,$$

and we call it the *outgoing bouquet* of q. Dually,[25] we define the *incoming bouquet* of q, $\mathsf{In}_{\mathcal{A}}(q)$, as the set of transitions of $\mathcal{A}$ whose destination is q – that is, of transitions that 'arrive' at q:

$$\mathsf{In}_{\mathcal{A}}(q) = \{e \in E \mid e\tau = q\} \ .$$

The two notions $\mathsf{Out}_{\mathcal{A}}$ and $\mathsf{In}_{\mathcal{A}}$ correspond to the transformation from an automaton to its transpose:

$$\mathsf{In}_{\mathcal{A}}(q) = \mathsf{Out}_{\mathcal{A}^{\mathsf{t}}}(q) \ .$$

Remark 3.4 We could conclude that we can discard the notion of an incoming bouquet and all the definitions and statements that go with it (and which are tedious, with their cumbersome prefixes). We shall persevere nonetheless as it is *necessary* to be able to make statements directly about a given automaton $\mathcal{A}$ rather than some about $\mathcal{A}$ and some about $\mathcal{A}^{\mathsf{t}}$; sometimes we will even talk about both at once. On the other hand, we shall not insist each time on the fact that a dual definition or proposition is equivalent to the direct equivalent on a transposed automaton. □

If φ is a morphism from $\mathcal{B}$ to $\mathcal{A}$ then, for all r in R, φ is a map from $\mathsf{Out}_{\mathcal{B}}(r)$ to $\mathsf{Out}_{\mathcal{A}}(r\varphi)$, and from $\mathsf{In}_{\mathcal{B}}(r)$ to $\mathsf{In}_{\mathcal{A}}(r\varphi)$.

The above considerations apply to morphisms of (labelled) graphs. For automata morphisms, we must also take into account the initial and final states or, along the lines of Remark 3.3, the fact that φ_{n} is a map from $\mathsf{Out}_{\mathcal{B}}(i_{\mathcal{B}})$ to $\mathsf{Out}_{\mathcal{A}}(i_{\mathcal{A}})$ and from $\mathsf{In}_{\mathcal{B}}(t_{\mathcal{B}})$ to $\mathsf{In}_{\mathcal{A}}(t_{\mathcal{A}})$ – which is exactly the content of conditions (3.3).

We will say that φ is *Out-surjective* if for each r in R the restriction of φ to $\mathsf{Out}_{\mathcal{B}}(r)$ is surjective from $\mathsf{Out}_{\mathcal{B}}(r)$ onto $\mathsf{Out}_{\mathcal{A}}(r\varphi)$. In particular, every state u of $\mathcal{B}$ whose image $u\varphi$ is a final state of $\mathcal{A}$ is a final state of $\mathcal{B}$: in other words, $\mathsf{Out}_{\mathcal{A}}(u\varphi)$ contains a transition $(u\varphi, 1_M, t_{\mathcal{A}})$, thus $\mathsf{Out}_{\mathcal{B}}(u)$ contains a transition $(u, 1_M, t_{\mathcal{B}})$. That is, $t\varphi^{-1}$ is contained in U for each t in T; and then, by (3.3), $T\varphi^{-1} = U$. Furthermore, if p is an initial state of $\mathcal{A}$, there is *at least* one initial state of $\mathcal{B}$ in $p\varphi^{-1}$: if $(i_{\mathcal{A}}, 1_M, p)$ is in $\mathsf{Out}_{\mathcal{A}}(i_{\mathcal{A}})$ there must be at least one transition $(i_{\mathcal{B}}, 1_M, r)$ in $\mathsf{Out}_{\mathcal{B}}(i_{\mathcal{B}})$ with r in $p\varphi^{-1}$.

Likewise, we shall say that φ is *Out-injective* if for each r in R the restriction of φ to $\mathsf{Out}_{\mathcal{B}}(r)$ is injective from $\mathsf{Out}_{\mathcal{B}}(r)$ to $\mathsf{Out}_{\mathcal{A}}(r\varphi)$. In particular, if p is an initial

[24] Stallings (*art. cit.*) uses the notation $\mathrm{Star}_{\mathcal{A}}(q)$, which is reminiscent of the set of 'rays' that leave q; however, it is impossible to use it in our context where the star operation is ubiquitous.

[25] The appellation 'dual' is here slightly improper and really applies only if $\mathcal{A}$ is an automaton over A^*.

state of $\mathcal{A}$, there is *at most* one initial state of $\mathcal{B}$ in $p\varphi^{-1}$: if $(i_{\mathcal{A}}, 1_M, p)$ is in $\mathsf{Out}_{\mathcal{A}}(i_{\mathcal{A}})$ there cannot be more than one transition $(i_{\mathcal{B}}, 1_M, r)$, with r in $p\varphi^{-1}$, in $\mathsf{Out}_{\mathcal{B}}(i_{\mathcal{B}})$. Moreover, this condition makes no additional restrictions on the final states.

We say that φ is *Out-bijective* if it is both Out-injective and Out-surjective. We shall return to these morphisms in more detail in Section 3.3.

In the same vein for the dual notions we shall say that φ is *In-surjective* (resp. *In-injective*) if for each r in R the restriction of φ to $\mathsf{In}_{\mathcal{B}}(r)$ is surjective on $\mathsf{In}_{\mathcal{A}}(r\varphi)$ (resp. injective). If φ is In-surjective, then $i\varphi^{-1}$ is contained in J for each i in I (that is, by (3.3) $I\varphi^{-1} = J$) and for each t in T, there exists *at least one* s in U such that $s\varphi = t$. If φ is In-injective, there exists *at most one* s in U such that $s\varphi = t$ for each t in T. We shall say that φ is *In-bijective* if it is both In-injective and In-surjective and we shall return to these morphisms in greater detail in Section 3.3.

In Figure 3.1, the morphism in (a) is Out-surjective and In-surjective; in (b) it is Out-injective and In-injective; in (c) it is Out-bijective.

Property 3.6 *The composition of two Out-surjective morphisms (resp. Out-injective, Out-bijective, In-surjective, In-injective, In-bijective) is Out-surjective (resp. Out-injective, Out-bijective, In-surjective, In-injective, In-bijective).* ■

Remark 3.5 It could be that an automata morphism $\varphi\colon \mathcal{B} \to \mathcal{A}$ induced, for example, a surjection from $\mathsf{Out}_{\mathcal{B}}(r)$ to $\mathsf{Out}_{\mathcal{A}}(r\varphi)$, for all r in R, without the corresponding conditions on the initial and final states being satisfied, or without its having been proved that they are. We shall then say that φ is an Out-surjective *labelled graph morphism*. We shall use the equivalent phrase for the other types of morphism. □

Exercises

• 3.1 Verify (a) Property 3.3,

(b) the assertion of Example 3.2, and (c) those of Remark 3.3.

• 3.2 For each state q of an automaton $\mathcal{A}$ over M, $\mathcal{A} = \langle Q, M, E, I, T \rangle$, write

$$\mathsf{Past}_{\mathcal{A}}(q) = \{m \in M \mid \exists i \in I \quad i \xrightarrow[\mathcal{A}]{m} q\} \quad \text{and} \quad \mathsf{Fut}_{\mathcal{A}}(q) = \{m \in M \mid \exists t \in T \quad q \xrightarrow[\mathcal{A}]{m} t\} \ ,$$

cf. Sec. 4, p. 273 'the past' and 'the future' of q in $\mathcal{A}$. Verify that if $\varphi\colon \mathcal{B} \to \mathcal{A}$ is a morphism, then for every state r of $\mathcal{B}$ we have $\mathsf{Past}_{\mathcal{B}}(r) \subseteq \mathsf{Past}_{\mathcal{A}}(r\varphi)$ and $\mathsf{Fut}_{\mathcal{B}}(r) \subseteq \mathsf{Fut}_{\mathcal{A}}(r\varphi)$.

3.3 Let δ and η be two actions of M on S and T respectively, and φ an action morphism of δ on η. Verify that if δ and η are viewed as automata $\mathcal{D}$ and $\mathcal{H}$ over M, as explained in Example 2.2, then φ is an automata morphism from $\mathcal{D}$ to $\mathcal{H}$.

In the following exercises, we consider automata over a free monoid A^*.

3.4 Show that the canonical map from a deterministic automaton to its minimal automaton is an automata morphism.

• 3.5 Verify that in general there is no automata morphism between a (non-deterministic) automaton and an equivalent deterministic automaton [by considering for example the automaton $\mathcal{A}_1$ from Example I.1.1].

3.2 Quotients of automata

As announced, Out-surjective morphisms allow us to reverse the inclusion of Property 3.1: that is, to relate the behaviour of the image to that of the source.

3.2.1 Out-surjective morphisms

Proposition 3.1 *Let $\varphi\colon \mathcal{B} \to \mathcal{A}$ be an Out-surjective automata morphism. For every computation c in $\mathcal{A}$ whose source p is in the image of φ* and for all s in R such that $s\varphi = p$, *there exists at least one computation d of $\mathcal{B}$ whose source is s satisfying $d\varphi = c$.*

Proof. By induction on the length of c. The property is trivial if c is a computation of length 0. Let

$$c \;:=\; p \xrightarrow[\mathcal{A}]{f} q \xrightarrow[\mathcal{A}]{m} r$$

be a computation (in $\mathcal{A}$), where (q, m, r) is a transition. By the induction hypothesis, for all s such that $s\varphi = p$ there exists at least one computation $e \;:=\; s \xrightarrow[\mathcal{B}]{f} t$ such that $e\varphi \;:=\; p \xrightarrow[\mathcal{A}]{f} q$ (hence $t\varphi = q$). Since φ is Out-surjective, there exists at least one transition $t \xrightarrow[\mathcal{B}]{m} u$ with source t whose image under φ is $q \xrightarrow[\mathcal{A}]{m} r$. The computation $d \;:=\; s \xrightarrow[\mathcal{B}]{f} t \xrightarrow[\mathcal{B}]{m} u$ has the required properties. ■

From this we obtain a series of consequences for automata morphisms.

Corollary 3.2 *If $\varphi\colon \mathcal{B} \to \mathcal{A}$ is an Out-surjective morphism, then $|\mathcal{B}| = |\mathcal{A}|$.*

Proof. If the source of a computation c of $\mathcal{A}$ is an initial state i, there exists at least one j in J such that $j\varphi = i$ and, by Proposition 3.1, at least one computation d of $\mathcal{B}$ with source j such that $d\varphi = c$. Now if c is a successful computation, its destination q is a final state, the destination t of d which belongs to $q\varphi^{-1}$ is also a final state, and d is indeed a successful computation. ■

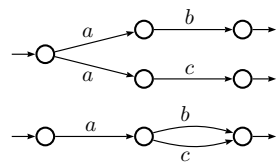
Corollary 3.3 *Let $\varphi\colon \mathcal{B} \to \mathcal{A}$ be an Out-surjective morphism. Let c be a computation of $\mathcal{A}$ and d an arbitrary computation of $\mathcal{B}$ such that $d\varphi = c$. For every computation c' that extends c there exists a computation d' which extends d and for which $d'\varphi = c'$.* ■

Corollary 3.4 *Let $\varphi\colon \mathcal{B} \to \mathcal{A}$ be an Out-surjective morphism. Then:*

(i) *φ is conformal if and only if φ is (globally) surjective – that is, $R\varphi = Q$;*

(ii) *if $\mathcal{A}$ is accessible, φ is conformal.* ■

A conformal morphism is not necessarily Out-surjective as the example in the adjacent figure shows.



Corollary 3.4 says that the really interesting case is that in which the morphism is globally surjective: it is natural enough to consider accessible automata. That is what we shall do now.

3.2.2 Totally surjective morphisms

Definition 3.2 We shall say that an automaton $\mathcal{A}$ is a *quotient* of $\mathcal{B}$ if there exists an automata morphism $\varphi\colon \mathcal{B} \to \mathcal{A}$ that is Out- and globally surjective. To simplify, we shall say that φ is *totally* surjective.[26] □

Figure 3.3 shows an automaton and one of its quotients (p and q are sent to p', r and s to r'). We have, for automata and their quotients, the same type of property as for actions and their morphisms: the existence of a minimum object.

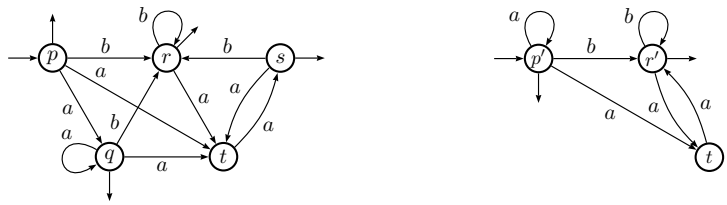


Figure 3.3: An automaton and a quotient

Proposition 3.5 *Every automaton $\mathcal{B}$ has a minimum quotient $\mathcal{C}$, unique up to isomorphism, which is the quotient of all the quotients of $\mathcal{B}$. More precisely, there exists $\psi\colon \mathcal{B} \to \mathcal{C}$, a totally surjective morphism, such that if $\varphi\colon \mathcal{B} \to \mathcal{A}$ is a totally surjective morphism, there exists $\theta\colon \mathcal{A} \to \mathcal{C}$, such that $\psi = \varphi\theta$.*

Proof. Suppose, as a simplification, that the automata under consideration have no distinct transitions which have the same source, label and destination; that is, that the transitions of $\mathcal{A}$ are subsets of the product $Q \times M \times Q$, which implies that the map $\varphi\colon F \to E$ is entirely determined by the map $\varphi\colon R \to Q$ itself defined, up to isomorphism, by the map equivalence $\varphi\varphi^{-1}$.

Let $\varphi\colon \mathcal{B} \to \mathcal{A}$ and $\varphi'\colon \mathcal{B} \to \mathcal{A}'$ be two totally surjective morphisms and let ψ be the equivalence relation on R generated by $\varphi\varphi^{-1}$ and $\varphi'\varphi'^{-1}$. We shall show that ψ determines an *Out*-surjective automata morphism (it is by definition globally surjective).

Write $[r]$ for the class modulo ψ of the state r of $\mathcal{B}$. By definition, a transition $([r], m, [s])$ appears in the quotient of $\mathcal{B}$ by ψ if *there exists* p in $[r]$ and t in $[s]$ such that (p, m, t) is a transition of $\mathcal{B}$. We must show that *for any* q in $[r]$ (that is, any q equivalent to p) *there exists* u in $[s]$ such that (q, m, u) is in F. The state q is in $[r]$ if and only if there exists a series of states $r_0 = p$, $r_1, \ldots, r_n = q$ such that, for all i in $\{0, 1, \ldots, n-1\}$, r_i is equivalent to r_{i+1} modulo $\varphi\varphi^{-1}$ or modulo $\varphi'\varphi'^{-1}$.

Since φ and φ' are Out-surjective, if there exists, for any i in $\{0, 1, \ldots, n-1\}$, a state s_i and a transition (r_i, m, s_i) in F, then there exists an s_{i+1} equivalent to s_i

[26] In the next chapter, we shall study weighted automata in a semiring $\mathbb{K}$ and define $\mathbb{K}$-quotients. The automata of this chapter can be considered as weighted automata in the Boolean semiring $\mathbb{B}$ and in fact the quotients that we have just defined are indeed $\mathbb{B}$-quotients. In the literature on 'transition systems', totally surjective morphisms are called *simulations*; see the bibliographic notes.

modulo $\varphi\varphi^{-1}$ or modulo $\varphi'\varphi'^{-1}$ (like r_i and r_{i+1}), such that (r_{i+1}, m, s_{i+1}) is in F. Setting $s_0 = t$, there therefore exists a state $u = s_n$ such that (q, m, u) is in F and ψ is Out-surjective.

The same reasoning proves that the upper bound of any set, even infinite, of totally surjective morphisms is also a totally surjective morphism; the maximum element of the set of all such morphisms corresponds to the minimum quotient of $\mathcal{B}$. ■

3.2.3 Moore's algorithm

It remains to determine how, at least in the case of finite automata, we can *compute* the minimum quotient of an automaton. We shall in fact generalise Moore's algorithm as described for the minimisation of (complete) deterministic automata. But, contrary to the situation in that case, we cannot define a (Nerode) equivalence on the states from a *global condition*. We shall include a series of local conditions which amounts to the description of the algorithm.

cf. Sec. I.3.3, p. 111

Let $\mathcal{B} = \langle R, M, F, J, U \rangle$ be an automaton over M. We define a series $\{\mathcal{R}_0, \mathcal{R}_1, \ldots\}$ of *nested partitions* of R in the following manner:

- $\mathcal{R}_0 = \{R \setminus U, U\}$;
- for all k, $\mathcal{R}_{k+1}$ is *the* refinement of $\mathcal{R}_k$ such that, for all p and q in R, p and q are joint in $\mathcal{R}_k$ and disjoint in $\mathcal{R}_{k+1}$,
  - (i) if there exists a transition (p, m, s) in F such that:
    - (a) either there is no transition with source q and label m in F;
    - (b) or for every transition (q, m, t) in F, s and t are disjoint in $\mathcal{R}_k$;
  - (ii) or the symmetric conditions obtained by swapping the roles of p and q hold.

The algorithm halts when $\mathcal{R}_k$ cannot be refined by the above criterion. Write l for this last index, if it exists. If R is finite, l exists and is less than $\|R\|$.

Notice that the starting point of the algorithm corresponds to a first step from the trivial partition $\mathcal{R}_{-1} = \{R\}$ taking into account only the transitions $(p, 1_M, t_{\mathcal{B}})$ that exist if and only if p is in U.

Write ψ_k for the canonical (surjective) map from R to the classes of $\mathcal{R}_k$. The refinement condition from $\mathcal{R}_k$ to $\mathcal{R}_{k+1}$ expresses precisely either that ψ_k is not a morphism (case (b)) or that ψ_k is not Out-surjective (case (a)) and, conversely, that ψ_l is an Out-surjective morphism.

Theorem 3.1 *The quotient $\mathcal{C} = \mathcal{B}/\psi_l$ computed by Moore's algorithm is the minimum quotient of $\mathcal{B}$.*

Proof. We have just seen that ψ_l is an Out-surjective morphism; it remains to show that $\mathcal{C}$ has no proper quotient. Suppose on the contrary that there exists an Out-surjective morphism $\alpha\colon \mathcal{C} \to \mathcal{D}$ (that is not globally injective) and write $\varphi = \psi_l \circ \alpha$.

For each pair of states p and q of $\mathcal{B}$ not equivalent modulo ψ_l but equivalent modulo φ, there exists a k such that p and q are joint in $\mathcal{R}_k$ and disjoint in $\mathcal{R}_{k+1}$

(which is why the last remark was important). Let us choose such a pair – one exists since α is not globally injective – which gives a *minimum* k.

For every transition (p, m, s) there exists a transition (q, m, t) since α is an Out-surjective morphism and $s\varphi = t\varphi$. Since p and q are disjoint in $\mathcal{R}_{k+1}$, there exists at least one such pair of transitions such that s and t are disjoint in $\mathcal{R}_k$, which contradicts the hypothesis of minimality in the choice of the pair p, q. ■

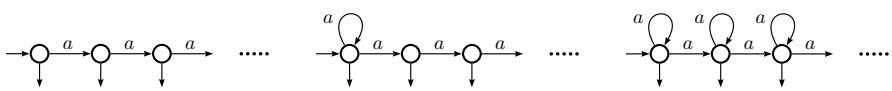
Remark 3.6 Let $\mathcal{B}$ be an automaton over A, complete and deterministic, which defines an action of A^* on the set of its states R. Moore's algorithm as described here is identical to that described in Chapter I and a totally surjective morphism is an action morphism. The notions of minimal automaton, of minimal action and of minimal quotient coincide in this case.

We stress the fact that, computed with the same algorithm, the minimal automaton of a language and the minimum quotient of an automaton (not necessarily deterministic) are fundamentally different: the first is canonically associated with the language, whereas the second is associated with the automaton, not with its behaviour. □

Exercises

3.6 Prove Corollary 3.3.

3.7 Describe the quotients of the following automata:



cf. Sec. I.5, p. 147 and p. 151

3.8 Let E be a rational expression over the alphabet A. Show that $\mathcal{A}_{\mathsf{E}}$, the derived term automaton of E, is a quotient of $\mathcal{S}_{\mathsf{E}}$, the standard automaton of E.

cf. p. 251

3.9 Prove Proposition 2.14.

3.10 **Bisimulation.** Let $\mathcal{B} = \langle R, M, F, J, U \rangle$ and $\mathcal{B}' = \langle R', M, F', J', U' \rangle$ be two automata. A *bisimulation* between $\mathcal{B}$ and $\mathcal{B}'$ is a relation β from R to R' ($\widehat{\beta} \subseteq R \times R'$) such that for all (p, p') in $\widehat{\beta}$ and every transition (p, m, q) in F, there exists a transition (p', m, q') in F' such that (q, q') is in $\widehat{\beta}$ and *vice versa*.

(a) Show that there is a bisimulation between $\mathcal{B}$ and $\mathcal{B}'$ if and only if their minimal quotients are isomorphic.

(b) Deduce from this the complexity of an algorithm for deciding if there is a bisimulation between two given finite automata.

3.3 Automata coverings

We shall now study another class of morphisms, more constrained than Out-surjective morphisms, and which are the basis of numerous constructions in the rest of what follows.

3.3.1 From local to global

Definition 3.3 An automata morphism φ from $\mathcal{B} = \langle R, M, F, J, U \rangle$ to $\mathcal{A} = \langle Q, M, E, I, T \rangle$ is a *covering* if it is both Out-bijective and globally surjective. The Out-bijectivity implies in particular that:

(i) for every i in I, there exists a *unique* j in J such that $j\varphi = i$;

(ii) for every t in T, $t\varphi^{-1}$ is contained in U: that is, $T\varphi^{-1} = U$. □

And dually:

Definition 3.4 An automata morphism φ from $\mathcal{B}$ to $\mathcal{A}$ is a *co-covering*[27] if it is both In-bijective and globally surjective. The In-bijectivity implies in particular that:

(i) for every i in I, $i\varphi^{-1}$ is contained in J; that is, $I\varphi^{-1} = J$;

(ii) for every t in T, there exists a *unique* s in U such that $s\varphi = t$. □

Example 3.1 (continued) The morphism in Figure 3.1 (a) is not Out-injective, while that in Figure 3.1 (b) is not Out-surjective. That in Figure 3.1 (c), reproduced in Figure 3.4 (a), is a covering, and that in Figure 3.4 (b) is a co-covering. □

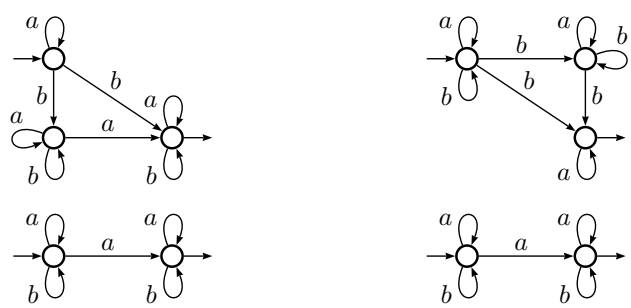


Figure 3.4: A covering and a co-covering

Definitions 3.3 and 3.4 are designed to obtain the following proposition, which makes the link between the *local* (equality of outgoing bouquets) and the *global* (equality of successful computations):

Proposition 3.6 *A covering (resp. a co-covering) $\varphi\colon \mathcal{B} \to \mathcal{A}$ induces a bijection between the successful computations of $\mathcal{A}$ and those of $\mathcal{B}$.*

Proof. Since φ is a morphism, the image $c = d\varphi$ of a successful computation d of $\mathcal{B}$ is a successful computation of $\mathcal{A}$; since it is Out-surjective, every successful computation c of $\mathcal{A}$ is the image of *at least one computation* d of $\mathcal{B}$. First, we suppose just that *the source of c is an initial state* and we will show that there exists a

[27] *Translator's note:* Duality's co-equal conception coerces co-operation with this cacophonous concoction. (In French, the term 'co-revêtement' is easier on the ears.)

unique computation d of $\mathcal{B}$ *whose source is an initial state* and such that $d\varphi = c$, by induction on the length of c and by repeating almost verbatim the proof of Proposition 3.1.

cf. p. 261

The property is true if c is a computation of length 0 since for every i in I there exists a *unique* j in J such that $j\varphi = i$ (condition 3.3 (i)). Let

$$c := i \xrightarrow[\mathcal{A}]{f} q \xrightarrow[\mathcal{A}]{m} r$$

be a computation (in $\mathcal{A}$), where (q, m, r) is a transition. By the induction hypothesis, there exists a *unique computation* $e := j \xrightarrow[\mathcal{B}]{f} t$ such that $e\varphi = i \xrightarrow[\mathcal{A}]{f} q$ (hence $t\varphi = q$) and that j is an initial state (of $\mathcal{B}$). Since φ is a covering, there exists *a unique transition* $t \xrightarrow[\mathcal{B}]{m} u$ with source t and whose image under φ is $q \xrightarrow[\mathcal{A}]{m} r$. The computation $d := j \xrightarrow[\mathcal{B}]{f} t \xrightarrow[\mathcal{B}]{m} u$ is uniquely determined and is what we require.

If now c is a successful computation – that is, r is a final state – then u, which belongs to $r\varphi^{-1}$, is also a final state (condition 3.3 (ii)) and d is indeed a successful computation.

The proof in the case where φ is a co-covering is analogous (and dual, or according to taste, follows from the proof above applied to transposed automata). ■

It will be helpful to say that $\mathcal{B}$ is a *covering of* $\mathcal{A}$ (resp. a *co-covering of* $\mathcal{A}$) if there exists a morphism $\varphi\colon \mathcal{B} \to \mathcal{A}$ which has this property. With this terminology, we deduce from Proposition 3.6:

Corollary 3.7 *A covering (resp. a co-covering) of an automaton is equivalent to that automaton.* ■

Corollary 3.8 *A covering (resp. a co-covering) of an unambiguous automaton is an unambiguous automaton.* ■

For the constructions required by the proofs to come we shall often use the accessible subset of an automaton rather than the automaton itself; but taking the accessible part of an automaton never leads to the suppression of an element of an outgoing bouquet of a state which belongs to this accessible part. Hence the following lemma (the dual argument gives the other part of the property).

Lemma 3.9 *Let $\varphi\colon \mathcal{B} \to \mathcal{A}$ be a covering (resp. a co-covering) and let $\mathcal{C}$ be the* accessible *(resp.* co-accessible*) subset of $\mathcal{B}$. The restriction of φ to $\mathcal{C}$ is a covering (resp. a co-covering) a $\mathcal{C}$ over $\mathcal{A}$.* ■

Definition 3.5 An *immersion* (resp. *co-immersion*) is a sub-automaton of a covering (resp. of a co-covering); more precisely and directly, this means that a morphism $\varphi\colon \mathcal{B} \to \mathcal{A}$ is an immersion if it is Out-injective, which implies among other things that for each i in I there exists *at most one* j in J such that $j\varphi = i$. Likewise, a morphism $\varphi\colon \mathcal{B} \to \mathcal{A}$ is a co-immersion if it is In-injective, which implies that for each t in T there exists *at most one* s in U such that $s\varphi = t$. □

If $\varphi\colon \mathcal{B} \to \mathcal{A}$ is an immersion (or a co-immersion), not only is $|\mathcal{B}| \subseteq |\mathcal{A}|$ – which holds when there exists a morphism from $\mathcal{B}$ to $\mathcal{A}$ – but φ is also *an injection* from the set of successful computations of $\mathcal{B}$ to those of $\mathcal{A}$. As for coverings, we shall say that $\mathcal{B}$ is an *immersion in* $\mathcal{A}$ if there exists a morphism $\varphi\colon \mathcal{B} \to \mathcal{A}$ which has this property.

The two following paragraphs present two 'standard' coverings, two constructions which are used in a wide range of situations and which it is useful to formalise now to avoid having to prove their properties every time we refer to them.

3.3.2 Product of an automaton with an action

In a general way the notions of covering and immersion allow us to give an interpretation *in automata terms* of properties whose classical statement and proofs owe more to the world of rational expressions or algebraic manipulations. The coverings in question appear most often thanks to a common construction: *the product of an automaton and an action.*

We defined the (cartesian) product of two automata $\mathcal{A}'$ and $\mathcal{A}''$ over A^*. In this case, to which we can always reduce since these are automata over A^*, all the transitions of $\mathcal{A}'$ and $\mathcal{A}''$ are labelled by *letters* of A. Each transition of $\mathcal{A}'\times\mathcal{A}''$ is obtained by taking a transition in $\mathcal{A}'$ and a transition in $\mathcal{A}''$ that has the same label. It would indeed be impossible to perform the same construction for two automata over an arbitrary monoid: we can very well imagine in that case two equivalent automata whose transitions have no label in common. This is why the rational subsets of a monoid are not in general closed under intersection. The situation is different if we are dealing with an automaton *over* M on one hand and an action *of* M *on a set* on the other.

cf. § *I.1.2.2, p. 61*

cf. also Exer. 1.2, p. 220

cf. § *1.2.5, p. 224*

Definition 3.6 Let $\mathcal{A} = \langle Q, M, E, I, T \rangle$ be an automaton over M, δ an action of M on a set S, with s_0 as the distinguished element, and $\mathcal{D}_\delta$ the automaton of the action δ. We write $\mathcal{A}\times\mathcal{D}_\delta$ for the automaton over M:

$$\mathcal{A}\times\mathcal{D}_\delta = \langle Q\times S, M, F, I\times\{s_0\}, T\times S \rangle \ ,$$

whose transitions are defined by

$$F = \{((p,s), m, (q, s\cdot m)) \mid s \in S\,,\ (p,m,q) \in E\} \ .$$

We call the *product of the automaton* $\mathcal{A}$ *and the action* δ the *accessible subset* of $\mathcal{A}\times\mathcal{D}_\delta$, and write it $\mathcal{A}\times\delta$. □

Since, for all s in S, $s\cdot m$ is uniquely determined by m (and since the definition is appropriate for the initial and final states), we have:

Property 3.7 *The projection* $\pi_{\mathcal{A}}$ *is a covering from* $\mathcal{A}\times\mathcal{D}_\delta$ *to* $\mathcal{A}$*, and hence from* $\mathcal{A}\times\delta$ *to* $\mathcal{A}$*.* ■

By induction on the length of computations, we have:

Property 3.8 $\forall (p,s) \in Q \times S\,,\ \forall m \in M$

$$(p,s) \xrightarrow[\mathcal{A}\times\mathcal{D}_\delta]{m} (q,t) \quad \Longrightarrow \quad p \xrightarrow[\mathcal{A}]{m} q \quad \textit{and} \quad t = s \cdot m \qquad \blacksquare$$

cf. p. 239

The following proposition, which makes more precise Proposition 2.6 (*the intersection of a rational subset and a recognisable subset is rational*), is an example of reformulation in automata terms of a property whose standard statement makes no reference to automata.

Proposition 3.10 *Let P be a rational subset of M, accepted by an automaton $\mathcal{A}$. The intersection of P and a recognisable subset of M is accepted by an automaton which is an immersion in $\mathcal{A}$.*

Proof. Let $\mathcal{A} = \langle Q, M, E, I, T \rangle$ be a (finite) automaton over M that accepts the rational subset P and δ be an action of M on S (finite) that recognises the subset $R = \{m \mid s_0 \cdot m \in K\}$.

The product $\mathcal{A} \times \delta = \langle Q \times S, M, F, I \times \{s_0\}, T \times S \rangle$ is a finite automaton and a covering of $\mathcal{A}$. The automaton

$$\mathcal{B} = \langle Q \times S, M, F, I \times \{s_0\}, T \times K \rangle$$

is a sub-automaton of $\mathcal{A}\times\delta$ and hence an immersion in $\mathcal{A}$. By Property 3.8, it accepts the subset

$$|\mathcal{B}| = \{m \mid i \xrightarrow[\mathcal{A}]{m} t\,,\ i \in I\,,\ t \in T \quad \text{and} \quad s_0 \cdot m \in K\} = P \cap R\ . \qquad \blacksquare$$

Remark 3.7 A crucial point of Definition 3.6 is that it is possible for $\mathcal{A}\times\delta$ to be a *finite automaton*, even if δ is an action of M on an *infinite set* S. □

In anticipation of Sec. IV.6.2, p. 602

Example 3.3 We take a monoid which is a direct product of free monoids: $M = A^* \times B^*$ (an automaton over $A^* \times B^*$ will be called a *transducer*) and for the action, the *'difference of lengths' action*, an action of $A^* \times B^*$ on $\mathbb{Z}$ defined by

$$\forall z \in \mathbb{Z}\,,\ \forall (f,g) \in A^* \times B^* \qquad z \cdot (f,g) = z + |f| - |g|\ .$$

cf. Exam. IV.6.1, p. 605

Figure 3.5 shows a transducer $\mathcal{F}_2$ and its product with the difference of lengths action. This product is finite: we shall say that the transducer $\mathcal{F}_2$ is 'of bounded lag'. □

We end by defining the *value* of a state (p,s) of $\mathcal{A}\times\delta$ as the element s of S and we shall say that the action δ is a *valuation* of $\mathcal{A}$ if $\pi_{\mathcal{A}}$ is a *bijection* from $\mathcal{A}\times\delta$ to $\mathcal{A}$. This new definition also makes it apparent that the construction corresponding to the product of an automaton by a given action does not cause anything to be repeated. More precisely, we have the following property.

Property 3.9 *Let $\mathcal{A}$ be an automaton over M and δ an action of M on a set S. The action δ is a valuation of $\mathcal{A}\times\delta$ and the value of a state of $(\mathcal{A}\times\delta)\times\delta$ is equal to the value of the corresponding state of $\mathcal{A}\times\delta$: that is, if $((p,s),t)$ is a state of $(\mathcal{A}\times\delta)\times\delta$, then $s = t$.* $\blacksquare$

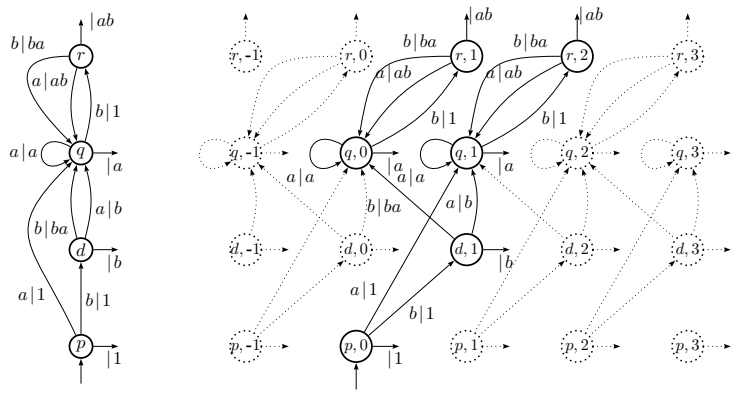


Figure 3.5: The transducer $\mathcal{F}_2$ and its product with the difference of lengths action. The accessible subset (which is finite) is drawn with solid lines; part of the rest is also shown, drawn with dotted lines.

3.3.3 The Coloured Transition Lemma

The notion of covering is also used in the following result that we propose to call the *Coloured Transition Lemma*.

Lemma 3.11 *Let $\mathcal{A} = \langle Q, M, E, I, T \rangle$ be a (finite) automaton over M. Suppose that the transitions of $\mathcal{A}$ (the elements of E) are* coloured*: either* red *or* blue*. Then the set of states of M that are the label of a successful computation of $\mathcal{A}$ which contains* at least one *red transition is a rational subset of M.*

Proof. We construct a covering $\mathcal{B}'$ of $\mathcal{A}$:

$$\mathcal{B}' = \langle Q \times \{0,1\}, M, F, I \times \{0\}, T \times \{0,1\} \rangle \ ,$$

in the following manner (the morphism from $\mathcal{B}'$ to $\mathcal{A}$ is clearly the projection on the first component):

$$\begin{aligned} F = \{((p,0), m, (q,0)) \mid (p,m,q) \in E \quad \text{and} \quad (p,m,q) \text{ is blue}\} \\ \cup \{((p,0), m, (q,1)) \mid (p,m,q) \in E \quad \text{and} \quad (p,m,q) \text{ is red}\} \\ \cup \{((p,1), m, (q,1)) \mid (p,m,q) \in E\} \ . \end{aligned}$$

A successful computation of $\mathcal{B}'$ which ends in a state $(t,0)$ contains only blue transitions, by construction. Hence the behaviour of $\mathcal{B} = \langle Q \times \{0,1\}, M, F, I \times \{0\}, T \times 1 \rangle$ is the rational set we seek. This construction is shown in Figure 3.6. ∎

Exercises

3.11 Show that every action (taken as an automaton) is a covering of the equivalent minimal action. *cf. Exam. 2.2, p. 233*

In the following exercises, we consider automata over a free monoid A^*.

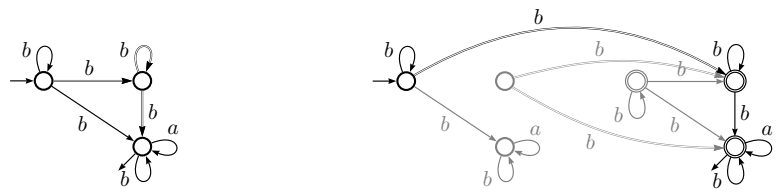
- 3.12 Show that a covering (resp. a co-covering) of a deterministic (resp. co-deterministic) automaton is a deterministic (resp. co-deterministic) automaton.

  3.13 Show that every deterministic automaton is a covering of its minimal automaton.

- 3.14 Prove Corollary 3.8.

- 3.15 Show that a covering (resp. a co-covering) of a co-deterministic (resp. deterministic) automaton is an unambiguous automaton.

- 3.16 Prove Property 3.9.



(a) The automaton $\mathcal{A}$ (with the red transitions drawn doubled)

(b) The automaton $\mathcal{B}'$ (with the states of $Q \times 1$ drawn double, and the non-useful part in grey)

Figure 3.6: The Coloured Transition Lemma

3.4 The Schützenberger covering

Automata over a free monoid are susceptible to determinisation by the subset method, and in this case a canonical construction enables us to associate each automaton with a special covering to which we shall give a crucial role in the coming chapters.

Theorem 3.2 *Let $\mathcal{A}$ be an accessible automaton over A and $\mathcal{A}_{\mathsf{det}}$ its determinisation. The* Schützenberger covering *or* S-covering *of $\mathcal{A}$ is the accessible subset $\mathcal{S}$ of $\mathcal{A}_{\mathsf{det}} \times \mathcal{A}$. The automaton $\mathcal{S}$ satisfies:*

(i) *$\pi_{\mathcal{A}}$ is a covering from $\mathcal{S}$ to $\mathcal{A}$;*

(ii) *$\pi_{\mathcal{A}_{\mathsf{det}}}$ is an In-surjective morphism from $\mathcal{S}$ to $\mathcal{A}_{\mathsf{det}}$.*

Example 3.4 (Continuation of Example I.1.1) Figure 3.7 shows the Schützenberger covering of the automaton $\mathcal{A}_1$, our first running example. □

We first show a more general proposition that proves as a special case condition (i) of Theorem 3.2.

Proposition 3.12 *Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an accessible automaton, $\mathcal{B} = \langle R, A, \eta, \{j\}, U \rangle$ a complete deterministic automaton equivalent to $\mathcal{A}$, and $\mathcal{E}$ the accessible subset of $\mathcal{B} \times \mathcal{A}$. Then $\pi_{\mathcal{A}}$, the projection of $\mathcal{B} \times \mathcal{A}$ onto $\mathcal{A}$, is a covering from $\mathcal{E}$ to $\mathcal{A}$.*

cf. Propty 3.7 *Proof.* If we abstract its final states, we can view $\mathcal{B}$ as *an action* of A^* on R and $\pi_{\mathcal{A}} \colon \mathcal{B} \times \mathcal{A} \to \mathcal{A}$ is an Out-bijective *labelled graph morphism.*

cf. Rem. 3.5

The automaton $\mathcal{B}$ has a single initial state, hence condition (i) of Definition 3.3 is satisfied. Let (s,t) be a state of $\mathcal{E}$; there exists f in A^* and i in I such that $i \xrightarrow[\mathcal{A}]{f} t$ and $j \xrightarrow[\mathcal{B}]{f} s$. If t is final, f belongs to $L(\mathcal{A})$; as s is uniquely determined by f and $\mathcal{B}$ is equivalent to $\mathcal{A}$, then s is final in $\mathcal{B}$ and (s,t) is final in $\mathcal{B}\times\mathcal{A}$ hence in $\mathcal{E}$: condition (ii) of Definition 3.3 is satisfied and $\pi_{\mathcal{A}}$ is an Out-surjective *automata morphism*. Finally, $\pi_{\mathcal{A}}$ is surjective from $\mathcal{E}$ to $\mathcal{A}$ since $\mathcal{A}$ is assumed to be accessible. ■

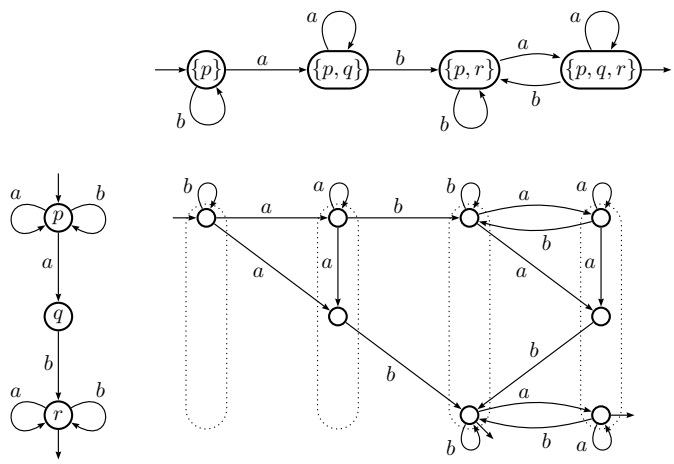


Figure 3.7: The S-covering of $\mathcal{A}_1$

Proof of Theorem 3.2. Since $\mathcal{A}_{\mathsf{det}}$ is a complete deterministic automaton equivalent to $\mathcal{A}$, it remains to prove condition (ii) of the theorem. From the definition of transitions in $\mathcal{A}_{\mathsf{det}} = \langle\, \mathfrak{P}(Q), A, F, \{I\}, U \,\rangle$, namely, *cf. Rem. I.3.1, p. 107*

$$P \xrightarrow[\mathcal{A}_{\mathsf{det}}]{a} S \quad \Longleftrightarrow \quad S = \{q \mid \exists p \in P \quad p \xrightarrow[\mathcal{A}]{a} q\} \ , \tag{3.4}$$

we first deduce:

Property 3.10 *The states of $\mathcal{S}$ are the pairs (P,p) where P is a state of $\mathcal{A}_{\mathsf{det}}$ and p is in P.*

Proof. Let P be a state of $\mathcal{A}_{\mathsf{det}}$: that is, there exists f in A^* such that

$$P = \{p \mid \exists i \in I \quad i \xrightarrow[\mathcal{A}]{f} p\} \ .$$

Thus (P,p) is a state of $\mathcal{S}$; that is, it is accessible in $\mathcal{A}_{\mathsf{det}}\times\mathcal{A}$ for all p in P. Conversely, if (P,q) is a state of $\mathcal{S}$, there exists f in A^* and i in I such that both $\{I\} \xrightarrow[\mathcal{A}_{\mathsf{det}}]{f} P$ and $i \xrightarrow[\mathcal{A}]{f} q$, and hence q is in P. ■

We next deduce by (3.4) that

$$\begin{aligned} \forall P, S \subseteq Q\,,\ \forall q \in S\,,\ \forall a \in A \qquad P \xrightarrow[\mathcal{A}_{\mathsf{det}}]{a} S \quad &\Longrightarrow \quad \exists p \in P \qquad p \xrightarrow[\mathcal{A}]{a} q \\ &\Longrightarrow \quad \exists p \in P \qquad (P,p) \xrightarrow[\mathcal{A}_{\mathsf{det}}\times\mathcal{A}]{a} (S,q) \end{aligned}$$

since (P,p) is a state of $\mathcal{S}$, which indeed means that $\pi_{\mathcal{A}_{\mathsf{det}}}\colon \mathcal{S} \to \mathcal{A}_{\mathsf{det}}$ is an In-surjective labelled graph morphism.

If $P \subseteq Q$ is final in $\mathcal{A}_{\mathsf{det}}$ there exists at least one t in P which is final in $\mathcal{A}$, hence a state (P,t) which is final in $\mathcal{S}$; every p in I, the unique initial state of $\mathcal{A}_{\mathsf{det}}$, is initial in $\mathcal{A}$ and every state (I,p) is initial in $\mathcal{S}$: hence, $\pi_{\mathcal{A}_{\mathsf{det}}}$ is an In-surjective automata morphism. ■

The duals of Theorem 3.2 and Proposition 3.12 are proved analogously.

Theorem 3.3 *Let $\mathcal{A}$ be a co-accessible automaton and $\mathcal{A}_{\mathsf{cod}}$ its co-determinisation. The* Schützenberger co-covering *or* SS-co-covering *of $\mathcal{A}$ is the co-accessible subautomaton $\mathcal{S}'$ of $\mathcal{A}_{\mathsf{cod}} \times \mathcal{A}$. The automaton $\mathcal{S}'$ satisfies:*

(i) *$\pi_{\mathcal{A}}$ is a co-covering from $\mathcal{S}'$ to $\mathcal{A}$;*

(ii) *$\pi_{\mathcal{A}_{\mathsf{cod}}}$ is an Out-surjective morphism from $\mathcal{S}'$ to $\mathcal{A}_{\mathsf{cod}}$.* ■

Proposition 3.13 *Let $\mathcal{A}$ be a co-accessible automaton, $\mathcal{B}$ a co-complete co-deterministic automaton equivalent to $\mathcal{A}$, and $\mathcal{S}'$ the co-accessible subset of $\mathcal{B}\times\mathcal{A}$. Then $\pi_{\mathcal{A}}$ is a co-covering from $\mathcal{S}'$ to $\mathcal{A}$.* ■

One of the important properties of the Schützenberger covering (and of the Schützenberger co-covering) is that some remarkable sub-automata can be extracted from it. In fact, we deduce from Theorem 3.2 (and Theorem 3.3):

Corollary 3.14 *Every automaton (over a free monoid) has an equivalent and unambiguous immersion (resp. co-immersion).*

Proof. Let $\mathcal{S}$ be the S-covering of an automaton $\mathcal{A}$. Since $\pi_{\mathcal{A}_{\mathsf{det}}}$ is In-surjective from $\mathcal{S}$ to $\mathcal{A}_{\mathsf{det}}$, we can, by deleting some transitions in $\mathcal{S}$ if $\pi_{\mathcal{A}_{\mathsf{det}}}$ is not In-injective, and possibly making certain states non-final, construct a *sub-automaton* $\mathcal{T}$ of $\mathcal{S}$ which is *a co-covering* of $\mathcal{A}_{\mathsf{det}}$ and hence unambiguous, equivalent to $\mathcal{A}_{\mathsf{det}}$ and hence to $\mathcal{A}$. A sub-automaton of $\mathcal{S}$, a covering of $\mathcal{A}$, is an immersion in $\mathcal{A}$. ■

Of course, it is not news that an automaton $\mathcal{A}$ over A^* should be equivalent to an unambiguous automaton – since one of the first results of the theory was the construction of a deterministic (resp. co-deterministic) automaton equivalent to $\mathcal{A}$. What gives Corollary 3.14 its substance, and what we shall use later, is that this unambiguous automaton is an *immersion* in $\mathcal{A}$ and that we can also find a correspondence between the computations of this automaton and those of $\mathcal{A}$. We formalise the construction of the automaton $\mathcal{T}$ above with the following definition.

cf. Exer. 3.5, p. 260

Definition 3.7 An *S-immersion* (resp. *S-co-immersion*) in $\mathcal{A}$ is any sub-automaton of the S-covering (resp. of the S-co-covering) of $\mathcal{A}$ which is a co-covering of $\mathcal{A}_{\mathsf{det}}$ (resp. a covering of $\mathcal{A}_{\mathsf{cod}}$). □

Example 3.4 (continued) In the case of the S-covering of $\mathcal{A}_1$, there is only one state, $(\{p,r\},r)$, where $\pi_{\mathcal{A}_{\mathsf{det}}}$ is not In-bijective, and in that state $\pi_{\mathcal{A}_{\mathsf{det}}}$ sends two

transitions to a single transition of $\mathcal{A}_{1\mathsf{det}}$. There are thus two possible S-immersions (*cf.* Figure 3.8). □

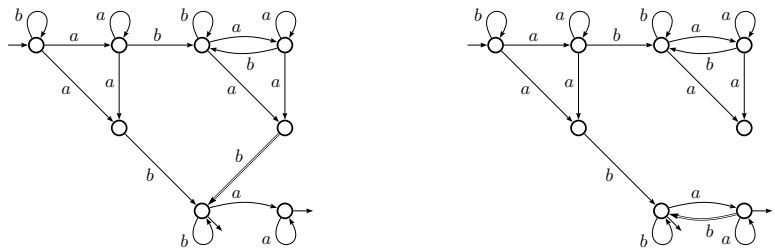
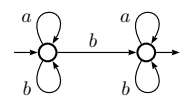


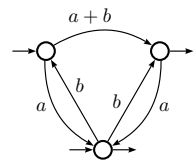
Figure 3.8: The two S-immersions of $\mathcal{A}_1$. The conserved transition is drawn doubled.

All these constructions, the S-covering, the S-co-covering, the S-immersion and the S-co-immersion, can equally be made on the matrix representations of automata. *cf. Sec. V.2.2, p. 667* We leave this to Chapter V where they will be used.

Exercises



- 3.17 Construct the S-covering and the S-co-covering of the adjacent automaton.
- 3.18 Construct the S-co-covering and the S-co-immersions of the automaton $\mathcal{A}_1$.
- 3.19 Show that the S-covering of a deterministic automaton is isomorphic to that automaton.
- 3.20 Show, by considering for example the product of $\mathcal{A}_1$ with its equivalent minimal deterministic automaton, that condition (ii) of Theorem 3.2 is not satisfied in general if the product of an automaton $\mathcal{A}$ is taken with an equivalent deterministic automaton other than $\mathcal{A}_{\mathsf{det}}$.



3.21 Construct the S-covering and the S-co-covering of the adjacent automaton.

How many S-immersions and S-co-immersions are there in this automaton?

4 Universal automaton

Let $\mathcal{A} = \langle Q, M, E, I, T \rangle$ be an automaton over M. For each state q of $\mathcal{A}$, the *past of q (in $\mathcal{A}$)* is the set of labels of computation which go from an initial state of $\mathcal{A}$ to q, and we write it $\mathsf{Past}_{\mathcal{A}}(q)$; that is,

$$\mathsf{Past}_{\mathcal{A}}(q) = \{m \in M \mid \exists i \in I \quad i \xrightarrow[\mathcal{A}]{m} q\} = |\langle Q, M, E, I, \{q\} \rangle| \ .$$

Dually,[28] the *future of q (in $\mathcal{A}$)* is the set of labels of computations that go from q to a final state of $\mathcal{A}$ and we write it $\mathsf{Fut}_{\mathcal{A}}(q)$ – that is:

$$\mathsf{Fut}_{\mathcal{A}}(q) = \{m \in M \mid \exists t \in T \quad q \xrightarrow[\mathcal{A}]{m} t\} = |\langle Q, M, E, \{q\}, T \rangle| \ .$$

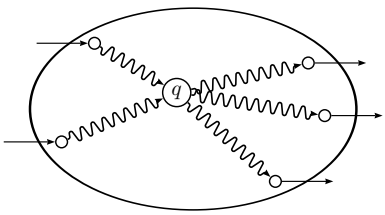


Fig. 4.1: Representation of the past and future of q in $\mathcal{A}$.

For each q in Q, we clearly have

$$[\mathsf{Past}_{\mathcal{A}}(q)] \ [\mathsf{Fut}_{\mathcal{A}}(q)] \subseteq |\mathcal{A}| \ . \qquad (*)$$

If we also write $\mathsf{Trans}_{\mathcal{A}}(p, q)$ for the set of labels of computations that go from p to q, we also have:

$$[\mathsf{Past}_{\mathcal{A}}(p)] \ [\mathsf{Trans}_{\mathcal{A}}(p, q)] \ [\mathsf{Fut}_{\mathcal{A}}(q)] \subseteq |\mathcal{A}| \ . \qquad (**)$$

Thus, each automaton, and in each automaton, each state, induces a set of 'factorisations' – which is the name we give to equations of the type $(*)$ or $(**)$ – of the subset which it accepts. The main idea of this section is to prove the converse of this observation, namely that we can construct from the set of factorisations of a subset K of M an automaton which accepts K. From this idea springs an array of original results, essentially due to J. H. Conway,[29] of which the simplest, but not the least notable, is that the automaton thus constructed contains a quotient of *every* automaton that accepts K and is minimal for this universal property.

This universal automaton itself appears as a special case of a general construction which enables us to define 'the best rational approximation' of a subset P by a family of subsets of M.

In what follows M is a monoid, *a priori* arbitrary; for the most part the results that we shall obtain apply *to recognisable subsets* of M. However, we should not lose sight of the fact that we shall thereby prove properties of the recognisable – and hence *rational* – subsets of the free monoid.

The monoid M is implicitly (and partially) ordered by inclusion; the product $M \times M$ and, more generally, any product $M \times \cdots \times M$, is ordered by the *product order*: that is, $(F_1, F_2, \ldots, F_n)$ is greater than $(G_1, G_2, \ldots, G_n)$ (we also say that $(F_1, F_2, \ldots, F_n)$ *dominates* $(G_1, G_2, \ldots, G_n)$) if $G_i \subseteq F_i$ for all i.

To start with, and following Conway in this, we shall define some objects using only the order structure on $M \times M$, and on M^n; we shall reuse the same definitions in a future subsection but for quotients (by the elements of M). The two points of view are complementary.

[28] The term 'dual' is here slightly improper and really applies only if $\mathcal{A}$ is an automaton over A^*. In this case $\mathsf{Fut}_{\mathcal{A}}(q) = [\mathsf{Past}_{\mathcal{A}^{\mathsf{t}}}(q)]^{\mathsf{t}}$.

[29] And presented in chapter 6 of his book *Regular Algebra and Finite Machines* [66]. Some of them were subsequently rediscovered in several places. The other sources are detailed in the notes section.

4.1 Factorisations

4.1.1 2-factorisations

Definition 4.1 Let K be a subset of M. A 2-*factorisation* of K is a pair (L, R) of subsets of M such that

$$L\,R \subseteq K \quad ,$$

and which is *maximal* for this property. We shall say that this L is a *left factor* of K and that R is a *right factor* of K. □

For the rest of the section, K is a subset of M.

Property 4.1 *For each subset P of M, there exists a unique* maximal subset R *such that $P\,R$ is contained in K and R is a* right factor *of K. Symmetrically, there exists a unique maximal subset L such that $L\,P$ is included in K and this L is a* left factor *of K.*

Proof. If $P\,D$ and $P\,D'$ are included in K, then so is $P\,(D \cup D')$ and the union of all the D such that $P\,D$ is included in K is the *unique maximal subset* R (a maximum subset) such that $P\,R$ is included in K.

If the pair (P, R) thus computed is dominated by a factorisation (L', R'), the series $P\,R \subseteq P\,R' \subseteq L'\,R' \subseteq K$ implies that $R = R'$ since R is maximal. ■

From this we deduce the following property, since every left factor L of K corresponds to a unique maximum subset R which is a right factor to which, symmetrically, a unique left factor corresponds, which is necessarily L.

Property 4.2 *The property '(L, R) is a factorisation of K' induces a bijection between the left and right factors of K.* ■

Remark 4.1 Let $\psi_M \colon M \times M \to M$ be the map defined by $(u, v)\psi_M = u\,v$. (This map is not a morphism unless M is commutative.) It can be seen as the *multiplication table* of M: in a matrix T of size $M \times M$, each element m appears as the entry (u, v) of T, for all (u, v) in $m\psi_M^{-1}$.

A 2-factorisation of K appears as a maximal rectangle in the subset $K\psi_M^{-1}$ and this point of view is without doubt the simplest when M is finite. □

Example 4.1 (Continuation of Example 2.3) The table of the monoid M_1, cleverly laid out (we have inverted the order of the elements x and y by row and column) is shown in the adjacent figure. The 2-factorisations of the subset $\{z\}$ are clearly visible: they are the rectangles

$$(\{z\}, M_1) \ , \quad (M_1, \{z\}) \quad \text{and} \quad (\{x, t, z\}, \{y, t, z\}) \, . \quad \square$$

1_M	x	y	t	z
y	t	y	t	z
x	x	z	z	z
t	t	z	z	z
z	z	z	z	z

Example 4.2 We set $M = \mathbb{Z}/6\mathbb{Z}$ and $K = \{1, 3, 4, 5\}$. We then verify that $(\{1\}, \{0, 2, 3, 4\})$ and $(\{0, 2, 4\}, \{1, 3, 5\})$ are 2-factorisations of K, and that $\{0, 2\}$ and $\{0, 1, 2, 3\}$ are not left (or right) factors. □

Notation We write Q_K for the *set of 2-factorisations* of K or, to be more precise, a set of indices in bijection with the set of 2-factorisations of K. For each q in Q_K there is a corresponding unique 2-factorisation (L_q, R_q) ; all the L_q and all the R_q are distinct.

4.1.2 Sub-factorisations and factorisations

What we have just defined and proved for the simple case of the product of two subsets generalises in fact to products of any number of subsets.[30]

Definition 4.2 Again, let K be a subset of M. A *sub-factorisation* of K is an n-tuple $(F_1, F_2, \ldots, F_n)$ of subsets of M such that

$$F_1 F_2 \cdots F_n \subseteq K \quad .$$

Each F_i is a *term*. A *factorisation* is a maximal sub-factorisation, and a *factor* is a term of a factorisation. □

Property 4.1 generalises easily:

Property 4.3 *Every sub-factorisation is dominated by a factorisation. Every factorisation which dominates a sub-factorisation coincides in the maximal terms of the sub-factorisation.*

Proof. Let $(F_1, \ldots, F_i, \ldots, F_n)$ be a sub-factorisation. For every i, the union of all the D such that $(F_1, \ldots, F_{i-1}, D, F_{i+1}, \ldots, F_n)$ is a sub-factorisation of K is the unique maximal subset F'_i (maximum subset) such that $(F_1, \ldots, F'_i, \ldots, F_n)$ is a sub-factorisation of K. By iteration we obtain a factorisation.

If $(F'_1, \ldots, F'_i, \ldots, F'_n)$ is a factorisation that dominates $(F_1, \ldots, F_i, \ldots, F_n)$, then $(F_1, .., F_{i-1}, F'_i, F_{i+1}, .., F_n)$ is a sub-factorisation of K. If F_i is maximal, $F_i = F'_i$. ■

The factorisation thus computed from a given sub-factorisation is not unique. In fact, all factorisations dominate $(\emptyset, \ldots, \emptyset)$ (but not all are obtained from this empty sub-factorisation by the above process). It is unique if the sub-factorisation under consideration has only *one* non-maximal term.

Property 4.4 *Every factor is the central factor of a* 3*-factorisation.*

Proof. Let $(F_1, \ldots, F_i, \ldots, F_n)$ be a factorisation K. Then F_i is maximal in the sub-factorisation $\big((F_1 \cdots F_{i-1}), F_i, (F_{i+1} \cdots F_n)\big)$ and we apply the above property. Likewise, F_1 is maximal in the sub-factorisation $\big(1_M, F_1, (F_2 \cdots F_n)\big)$, and F_n is maximal in $\big((F_1 \cdots F_{n-1}), F_n, 1_M\big)$. ■

Consistency with the previous subsubsection is ensured by the following property.

Property 4.5 *Every factor which appears as the leftmost (resp. rightmost) term of a factorisation is a left (resp. right) factor.*

[30] This subsubsection follows J. H. Conway, *op. cit.*, almost word for word, even though we are working in an arbitrary monoid, not in a free monoid.

Proof. As above, if $(F_1, \ldots, F_n)$ is a factorisation, then F_1 is maximal in the 2-term sub-factorisation $\big(F_1, (F_2 \cdots F_n)\big)$, and F_n is maximal in the sub-factorisation $\big((F_1 \cdots F_{n-1}), F_n\big)$, and we apply Property 4.3. ∎

4.1.3 Morphisms and factorisations

The map $\psi_M \colon M \times M \to M$ is not a morphism, but it does *commute with morphisms* in the sense that the adjacent diagram commutes, since $\alpha \colon M \to N$ is a morphism. It follows that factorisations are 'syntactic'[31] objects, or more precisely:

$$\begin{tikzcd} M \times M \arrow[r, "\psi_M"] \arrow[d, "\alpha \times \alpha"'] & M \arrow[d, "\alpha"] \\ N \times N \arrow[r, "\psi_N"] & N \end{tikzcd}$$

Proposition 4.1 *Let $\alpha \colon M \to N$ be a* surjective *monoid morphism and K a subset of M* recognised by α. *Then:*

(i) *the factors of K are recognised by α;*

(ii) *α induces a bijection between the factorisations of K and those of $K\alpha$.*

Proof. If $(F_1, \ldots, F_n)$ is a sub-factorisation of K, then $(F_1\alpha, \ldots, F_n\alpha)$ is a sub-factorisation of $K\alpha$, and $(F_1\alpha\alpha^{-1}, \ldots, F_n\alpha\alpha^{-1})$ is a sub-factorisation of $K\alpha\alpha^{-1} = K$ that dominates $(F_1, \ldots, F_n)$. If therefore $(F_1, \ldots, F_n)$ is a factorisation of K, then $(F_1, \ldots, F_n) = (F_1\alpha\alpha^{-1}, \ldots, F_n\alpha\alpha^{-1})$, which proves (i).

(ii) Let $(F_1, \ldots, F_n)$ be a factorisation of K and $(F'_1, \ldots, F'_n)$ be a factorisation of $K\alpha$ that dominates $(F_1\alpha, \ldots, F_n\alpha)$, a sub-factorisation of $K\alpha$. Then $(F'_1\alpha^{-1}, \ldots, F'_n\alpha^{-1})$ is a sub-factorisation of K which dominates

$$(F_1\alpha\alpha^{-1}, \ldots, F_n\alpha\alpha^{-1}) = (F_1, \ldots, F_n)$$

and thus is equal to it (since $(F_1, \ldots, F_n)$ is a factorisation). Hence

$$(F_1\alpha, \ldots, F_n\alpha) = (F'_1\alpha^{-1}\alpha, \ldots, F'_n\alpha^{-1}\alpha) = (F'_1, \ldots, F'_n)$$

(since α is surjective) is a factorisation of $K\alpha$.

Conversely, if $(F'_1, \ldots, F'_n)$ is a sub-factorisation of H, a subset of N, then $(F'_1\alpha^{-1}, \ldots, F'_n\alpha^{-1})$ is a sub-factorisation of $H\alpha^{-1}$ since

$$H\alpha^{-1} \supseteq (F'_1 \cdots F'_n)\alpha^{-1} \supseteq F'_1\alpha^{-1} \cdots F'_n\alpha^{-1} \quad .$$

Now let $(F'_1, \ldots, F'_n)$ be a factorisation of $K\alpha$ and $(F_1, \ldots, F_n)$ be a factorisation of $K = K\alpha\alpha^{-1}$ that dominates the sub-factorisation $(F'_1\alpha^{-1}, \ldots, F'_n\alpha^{-1})$. The foregoing implies that $(F_1\alpha, \ldots, F_n\alpha) = (F'_1, \ldots, F'_n)$, from which

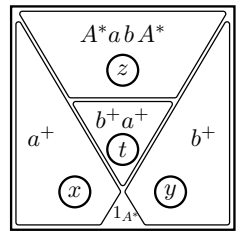
$$(F'_1\alpha^{-1}, \ldots, F'_n\alpha^{-1}) = (F_1\alpha\alpha^{-1}, \ldots, F_n\alpha\alpha^{-1}) = (F_1, \ldots, F_n)$$

is a factorisation of K. ∎

Corollary 4.2 *Let $\alpha \colon M \to N$ be a* surjective *morphism and R a subset of N. The inverse images under α of the factors of R are the factors of $R\alpha^{-1}$.* ∎

[31] *Cf.* Rem. 2.6, p. 250

Example 4.1 (continued) We first compute (see adjacent figure)



$$\begin{aligned}\{x,t,z\}\check{\delta}_1^{\,-1} &= a^+ \cup b^+a^+ \cup A^*abA^* \\ &= b^*a^+ \cup (A^* \setminus b^*a^*) = A^* \setminus b^* = A^*aA^* \\ \{y,t,z\}\check{\delta}_1^{\,-1} &= b^+ \cup b^+a^+ \cup A^*abA^* = A^*bA^* \ .\end{aligned}$$

The three 2-factorisations of $L_1 = A^*abA^*$ are thus

$$(L_1, A^*) \ , \qquad (A^*, L_1) \ , \qquad \text{and} \qquad (A^*aA^*, A^*bA^*) \ . \qquad \square$$

Corollary 4.3 *A subset K of M is recognisable if and only if it has a finite number of factors.*

Proof. If K is recognisable, it is recognised by a finite monoid in which its image has a necessarily finite number of 3-factorisations, hence of factors.

For each m in M, $Km^{-1} = \{x \mid xm \in K\}$ and hence $(Km^{-1}, \{m\})$ is a sub-factorisation of K in which Km^{-1} is maximal: it is a left factor of K (Property 4.2 (i)). If there is only a finite number of left factors, there is only a finite number of quotients Km^{-1}: K is recognisable (Theorem 2.3). ■

cf. p. 247

In the course of this proof, we can characterise left (and right) factors with the help of right (and left) quotients.

Proposition 4.4 *Let K be a subset of M. Every intersection of left quotients of K (by elements of M) is a right factor of K and conversely every right factor of K is an intersection of left quotients of K.*

Proof. Let P be an arbitrary subset of M; in the sub-factorisation $\left(P, \bigcap_{m\in P} m^{-1}K\right)$ the second term is maximal and is hence a right factor. Conversely, if (L, R) is a factorisation of K, then $R = \bigcap_{m\in L} m^{-1}K$. ■

If K is recognisable, the preceding intersections reduce to finite intersections since there is only a finite number of distinct quotients (left or right) of K, and Corollary 4.3 is a direct consequence of Proposition 4.4. We shall return in Subsection 4.3 to the definition of factors in terms of quotients.

Exercises

- 4.1 Compute the 2-factorisations of the language aA^*.
- 4.2 Compute the 2-factorisation of the subset $K_2 = \{x, x^2, xy\}$ in the monoid M_2, the quotient of $\{x,y\}^*$ by the relations $x^3 = x$, $y^2 = y$, $yx = x$ and $x^2y = y$.
- 4.3 How many 2-factorisations are there of the subset $K = \{1,3,4,5\}$ in $\mathbb{Z}/6\mathbb{Z}$?
- 4.4 Give a necessary and sufficient condition on the subset K for $(\emptyset, M)$ (resp. $(M, \emptyset)$) to be a 2-factorisation of K.

- 4.5 Use the decomposition of a recognisable subset into a finite number of 2-factorisations to prove Proposition 2.19. *cf. p. 253*

- 4.6 Let $\mathsf{Conj}\colon A^* \to A^*$ be the relation which associates a word f with the set of its conjugates: $\mathsf{Conj}(f) = \{v\,u \mid u, v \in A^* \quad u\,v = f\}$. Show that $\mathsf{Conj}(K)$ is rational when K is. *cf. Exer. I.2.12, p. 92*

4.2 Universal automata of a subset

The richness of the concept of factors of a subset K comes, among other things, from the fact that they can been seen as the labels of an automaton – a finite automaton if K is recognisable. From now on, K is a fixed subset of M.

4.2.1 Definitions and examples

Notation Let Q_K be the set of (indices of) 2-factorisations of K. For all p and q in Q_K, we write $F_{p,q}$ for the maximum subset of M such that $L_p\,F_{p,q}\,R_q$ is contained in K:

$$F_{p,q} = \{m \in M \mid L_p\,m\,R_q \subseteq K\} \ .$$

We obtain in this way sub-factorisations $(L_p, F_{p,q}, R_q)$ of K in each of which $F_{p,q}$ is maximal.

Property 4.6 *Each $F_{p,q}$ is a factor and every factor is an $F_{p,q}$.*

Proof. Since $F_{p,q}$ is maximal in the sub-factorisation $(L_p, F_{p,q}, R_q)$, $F_{p,q}$ is a factor (Property 4.3). Conversely, every factor G is the central factor of a 3-factorisation (Property 4.4) whose left factor is necessarily an L_p and right factor an R_q (Property 4.5), and hence $G = F_{p,q}$ for a pair of indices (p, q) in $Q_K \times Q_K$, and this pair is not necessarily unique. ■

Definition 4.3 Let K be a subset of M. The *universal automaton* of K, written $\mathcal{U}_K$, is the following automaton over M:

$$\mathcal{U}_K = \langle\, Q_K, M, F_K, I_K, T_K \,\rangle \ ,$$

whose transitions are defined by

$$F_K = \{\big((L_p, R_p), F_{p,q}, (L_q, R_q)\big) \mid (L_p, R_p), (L_q, R_q) \in Q_K\}$$

and where the sets I_K and T_K are defined by

$$I_K = \{(L_p, R_p) \in Q_K \mid 1_M \in L_p\} \quad \text{and} \quad T_K = \{(L_q, R_q) \in Q_K \mid 1_M \in R_q\} \ . \quad \square$$

Example 4.1 (continued) Set $Q_1 = \{u, v, w\}$, the set of factorisations of $\{z\}$ where $(L_u, R_u) = (M_1, \{z\})$, $(L_v, R_v) = (\{x, t, z\}, \{y, t, z\})$ and $(L_w, R_w) = (\{z\}, M_1)$. We then easily determine the universal automaton of $\{z\}$. We then deduce mechanically, by Corollary 4.2, the universal automaton of L_1; the factorisations are written $(L_u, R_u) = (A^*, A^*a\,b\,A^*)$, $(L_v, R_v) = (A^*a\,A^*, A^*b\,A^*)$ and $(L_w, R_w) = (A^*a\,b\,A^*, A^*)$. Figure 4.2 shows these two universal automata. $\square$ *cf. pp. 275 and 277*

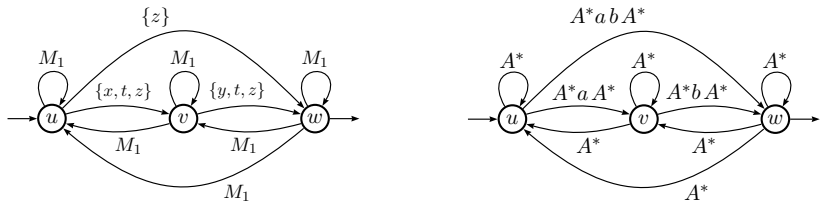
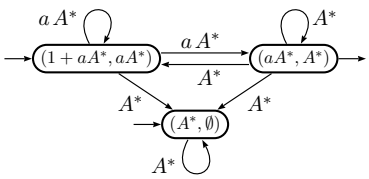


Figure 4.2: The universal automata of $\{z\}$ and L_1



Example 4.3 The adjacent figure shows the universal automaton of the language $a\,A^*$, without going via the syntactic monoid. This example allows us to see that a universal automaton is not necessarily trim: a factorisation $(\emptyset, M)$ (resp. $(M, \emptyset)$), if it exists, corresponds to a non-accessible state (resp. non-co-accessible). □

The set of transitions F_K can be naturally organised as a matrix which Conway calls the *factor matrix*[32] of K.

4.2.2 Properties

The least that one expects of the universal automaton of K is that it should accept K, which is written, briefly:

Theorem 4.1 *For all K in M, $|\mathcal{U}_K| = K$.*

We shall not prove this result in the most direct way, but via a series[33] of properties each of which is interesting in its own right.

Properties 4.7 *For all n-tuples $p_1, p_2, \ldots, p_n$ of factorisations of Q_K, we have:*

(i) $L_{p_1}\, F_{p_1,p_2}\, F_{p_2,p_3} \cdots F_{p_{n-1},p_n} \subseteq L_{p_n}$;

(ii) $F_{p_1,p_2}\, F_{p_2,p_3} \cdots F_{p_{n-1},p_n}\, R_{p_n} \subseteq R_{p_1}$;

(iii) $F_{p_1,p_2}\, F_{p_2,p_3} \cdots F_{p_{n-1},p_n} \subseteq F_{p_1,p_n}$;

(iv) $1_M \in F_{p_1,p_1}$.

Proof. From $L_{p_1}\, F_{p_1,p_2}\, R_{p_2} \subseteq K$, we deduce $L_{p_1}\, F_{p_1,p_2} \subseteq L_{p_2}$ since L_{p_2} is a maximum such that $L_{p_2}\, R_{p_2} \subseteq K$.

By iterating the reasoning we establish (i) and by symmetry we obtain (ii). One or the other gives $L_{p_1}\, F_{p_1,p_2}\, F_{p_2,p_3} \cdots F_{p_{n-1},p_n}\, R_{p_n} \subseteq K$ and hence (iii), since F_{p_1,p_n} is a maximum such that $L_{p_1}\, F_{p_1,p_n}\, R_{p_n} \subseteq K$.

The same argument gives (iv) since $L_{p_1}\, R_{p_1} \subseteq K$. ■

[32] In fact, Conway, *op. cit.*, does not define the universal automaton but only the factor matrix (but it is only one step from a transition matrix to an automaton). Conway writes this matrix $\lceil K \rceil$ and $K_{i,j}$ for its coefficients the factors. I have opted for a hybrid solution, without explicit reference to K in the factors, but which simplifies the notation in the equations.

[33] Again following Conway, *op. cit.*, almost verbatim.

We said that F_K can be seen as a matrix whose coefficients are the factors of K. Property 4.7(iv) implies that $F_K \subseteq F_K^2$ and (iii) implies the reverse inclusion. As $F_K^0 \subseteq F_K$, again by (iv), we have just proved:

Property 4.8 $F_K^* = F_K$. ■

Properties 4.3 and 4.6 imply a sort of inverse of Property 4.7 (iii):

Property 4.9 *If H_1, H_2,..., are n subsets of M such that*

$$H_1 H_2 \cdots H_n \subseteq F_{p,q} \ ,$$

then there exists an $n+1$-tuple $p = p_0$, p_1,..., $p_n = q$ of factorisations of Q_K such that

$$H_1 \subseteq F_{p,p_1} \ , \ H_2 \subseteq F_{p_1,p_2} \ , \ \ldots \ , \ H_n \subseteq F_{p_{n-1},q} \ .$$ ■

Property 4.10 *There exist two unique indices l and r in Q_K, such that*

$$K = L_r = R_l = F_{l,r} \ ,$$

and, for all q in Q_K, $L_q = F_{l,q}$ and $R_q = F_{q,r}$.

Proof. Since K is maximal in the sub-factorisation $(1_M, K)$, there exists a unique l in Q_K such that

$$L_l K \subseteq K \ ; \tag{4.1}$$

that is, $K = R_l$ by Property 4.2. Symmetrically, there exists a unique r such that

$$K R_r \subseteq K \ , \tag{4.2}$$

hence $K = L_r$. By (4.1) and (4.2) it follows that (L_l, K, R_r) is a sub-factorisation in which each term is maximal, hence in particular $K = F_{l,r}$.

For each q, $(L_l, F_{l,q}, R_q)$ is a sub-factorisation in which $F_{l,q}$ is maximal; hence $(F_{l,q}, R_q)$ is a sub-factorisation of $R_l = K$ in which $F_{l,q}$ is maximal, so $F_{l,q} = L_q$. Symmetrically, $(L_q, F_{q,r}, R_r)$ is a sub-factorisation in which $F_{q,r}$ is maximal; hence $(L_q, F_{q,r})$ is a sub-factorisation of $L_r = K$ in which $F_{q,r}$ is maximal, so $F_{q,r} = R_q$. ■

This organisation is shown in Figure 4.3. The factorisation (L_l, R_l) is called the *initial factorisation* and is indeed an initial state of $\mathcal{U}_K$; (L_r, R_r) is called the *final factorisation* and is a final state of $\mathcal{U}_K$.

Property 4.11 *For all q in Q_K, $\mathsf{Past}_{\mathcal{U}_K}(q) = L_q$ and $\mathsf{Fut}_{\mathcal{U}_K}(q) = R_q$.*

Proof. By definition of $\mathcal{U}_K$ we have

$$\mathsf{Past}_{\mathcal{U}_K}(q) = \bigcup_{1_M \in L_p} F_{p,q} \ .$$

As 1_M belongs to L_p, we conclude $F_{p,q} \subseteq L_p F_{p,q} \subseteq L_q$; moreover, we have just shown that $F_{l,q} = L_q$ and the first equality is proved. A symmetric computation gives the second. ■

Proof of Theorem 4.1. The proof is analogous to the previous one. By definition, $|\mathcal{U}_K| = I \cdot F_K^* \cdot T$ and hence (Property 4.8):

$$|\mathcal{U}_K| = I_K \cdot F_K \cdot T_K = \bigcup_{\substack{1_M \in L_p, \\ 1_M \in R_q}} F_{p,q} \ .$$

To start with, $1_M \in L_p$ and $1_M \in R_q$ imply that $F_{p,q} \subseteq L_p F_{p,q} R_q \subseteq K$ so $|\mathcal{U}_K| \subseteq K$. We obtain the inverse inclusion by taking $p = l$ and $q = r$; that is, taking the initial factorisation as the initial state, and the final factorisation as the final state,

$$K = F_{l,r} \subseteq |\mathcal{U}_K| \ . \qquad \blacksquare$$

$$\begin{array}{rcc|ccccccc}
 & R & \rightarrow & & R_l = K & & R_q & & R_r & \\
L & Q_K & \rightarrow & .. & l & & q & .. & r & .. \\
\downarrow & \downarrow & & & & & & & & \\
 & \vdots & & & \vdots & & \vdots & & \vdots & \\
L_l & l & & .. & L_l & .... & L_q & ..... & K & ..... \\
 & \vdots & & & \vdots & & \vdots & & \vdots & \\
L_q & q & & ..... & F_{q,l} & .... & F_{q,q} & .... & R_q & ..... \\
 & \vdots & & & \vdots & & \vdots & & \vdots & \\
L_r = K & r & & ..... & F_{r,l} & .... & F_{r,q} & .... & R_r & ..... \\
 & \vdots & & & \vdots & & \vdots & & \vdots &
\end{array}$$

Figure 4.3: Organisation of the matrix F_K

Exercises

4.7 Show that the indices l and r defined in Property 4.10 are equal if and only if K is a submonoid of M (that is, $K = (K')^*$ for some K').

4.8 Show that the factors of factors of K are themselves factors of K.

4.2.3 Universal automaton relative to a generating set

It is never easy, and sometimes tiresome, to work with automata whose transitions are labelled by subsets, possibly infinite, of M. As we did for the automaton associated with an action, we now consider the 'trace' of a generating set of M on the universal automaton.

cf. § *2.2.4, p. 241*

Definition 4.4 Let G be a generating set of M. Let K be a subset of M, $\mathcal{U}_K$ its universal automaton and F_K its factor matrix. Write $E_K^{(G)}$ for the 'trace' of G over F_K; that is, writing $E_{p,q}$ for the entries of $E_K^{(G)}$:

$$\forall p, q \in Q_K \qquad E_{p,q} = F_{p,q} \cap G \ .$$

The automaton $\mathcal{U}_K^{(G)} = \langle Q_K, M, E_K^{(G)}, I_K, T_K \rangle$ is called the *universal automaton of K relative to G*, or the *G-universal automaton* of K. $\square$

Theorem 4.2 *For every generating set G of M and every subset K of M, the behaviour of $\mathcal{U}_K^{(G)}$ is equal to K; that is, $|\mathcal{U}_K^{(G)}| = K$.*

Proof. We write $F_K \cap G^+$ for the matrix whose entries are the intersections $F_{p,q} \cap G^+$ and $\mathsf{c}\,(F_K)$ for the matrix whose entries are the $\mathsf{c}\,(F_{p,q})$. (Note that $F_{p,q} \cap G^+$ and $\mathsf{c}\,(F_{p,q})$ are not necessarily disjoint if M has divisors of the identity.) Obviously (and by Theorem 4.1), we have

$$|\mathcal{U}_K| = I_K \cdot F_K \cdot T_K = I_K \cdot \mathsf{c}\,(F_K) \cdot T_K + I_K \cdot \big(F_K \cap G^+\big) \cdot T_K$$

and

$$|\mathcal{U}_K^{(G)}| = I_K \cdot \big(E_K^{(G)}\big)^* \cdot T_K = I_K \cdot T_K + I_K \cdot \big(E_K^{(G)}\big)^+ \cdot T_K \ .$$

We shall show successively

$$\big(E_K^{(G)}\big)^+ = F_K \cap G^+ \qquad \text{and} \qquad I_K \cdot T_K = I_K \cdot \mathsf{c}\,(F_K) \cdot T_K \ . \tag{4.3}$$

First, $E_K^{(G)} \subseteq F_K$ implies $\big(E_K^{(G)}\big)^+ \subseteq F_K^+ = F_K$. As each entry of $E_K^{(G)}$ is in G, each entry of $\big(E_K^{(G)}\big)^+$ is in G^+ and $\big(E_K^{(G)}\big)^+ \subseteq F_K \cap G^+$.

Conversely, let $m = g_1\, g_2 \cdots g_n$ be an entry of $F_{p,q} \cap G^n$. If $n = 1$, m belongs to $F_{p,q} \cap G = E_{p,q}$. If $n \geqslant 2$, Property 4.9 ensures that there exists a sequence $r_1, \ldots, r_{n-1}$ in Q_K such that

$$g_1 \in F_{p,r_1} \ , \quad g_2 \in F_{r_1,r_2} \ , \quad \ldots \ , \quad g_n \in F_{r_{n-1},q} \ .$$

Hence $g_1 \in E_{p,r_1}$, $g_2 \in E_{r_1,r_2}$, $\ldots$, $g_n \in E_{r_{n-1},q}$, and m belongs to the entry (p,q) of $\big(E_K^{(G)}\big)^n$ hence of $\big(E_K^{(G)}\big)^+$.

Property 4.7 (iv) implies that $I_K \cdot T_K \subseteq I_K \cdot \mathsf{c}\,(F_K) \cdot T_K$. Conversely, if 1_M belongs to $I_K \cdot \mathsf{c}\,(F_K) \cdot T_K$, there exist p and q in Q_P such that 1_M belong to L_p, $F_{p,q}$ and R_q. From $F_{p,q}\, R_q \subseteq R_p$, we conclude that 1_M belongs to $L_p\, R_p \subseteq I_K \cdot T_K$. ■

The notation of Theorem 4.2 is cumbersome; in fact, and except for a few special cases, we will study one or the other of these two universal automata – the generating set G being most often canonical and implicit – and we will simply write $\mathcal{U}_K = \langle\, Q_K, M, E_K, I_K, T_K \,\rangle$, using context to determine whether we mean the 'full' universal automaton or that relative to G. Figure 4.4 shows the $\{a,b\}$-universal automata of $L_1 = A^* a\, b\, A^*$ and $a\, A^*$ respectively.

cf. Fig. 4.2 and Exam. 4.3

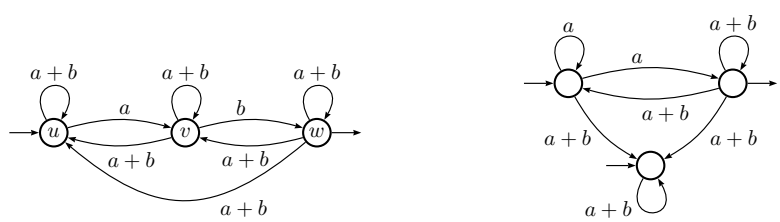


Figure 4.4: The $\{a,b\}$-universal automata of L_1 and $a\, A^*$

The proof of Theorem 4.2 also proves:

Property 4.12 *For all q in Q_K, $\mathsf{Past}_{\mathcal{U}_K^{(G)}}(q) = L_q$ and $\mathsf{Fut}_{\mathcal{U}_K^{(G)}}(q) = R_q$.* ■

4.2.4 Universality of universal automata

The automata $\mathcal{U}_K$ and $\mathcal{U}_K^{(G)}$ are universal in the sense that they 'contain' every automaton that accepts K. More precisely:

Proposition 4.5 *Let $\mathcal{A} = \langle R, M, H, J, U \rangle$ be an automaton over M whose behaviour is contained in K. Then there exists an automata morphism φ from $\mathcal{A}$ to $\mathcal{U}_K$. If in addition all the elements of H are in the generating set G, then φ sends $\mathcal{A}$ into $\mathcal{U}_K^{(G)}$.*

Proof. For each u in R, we have $\mathsf{Past}_{\mathcal{A}}(u)\,\mathsf{Fut}_{\mathcal{A}}(u) \subseteq |\mathcal{A}| \subseteq K$ and we note that if i is in J, then $\mathsf{Past}_{\mathcal{A}}(i)$ contains 1_M, and that if t is in U, then $\mathsf{Fut}_{\mathcal{A}}(t)$ contains 1_M.

Let P be *the maximal subset* such that $P\,\mathsf{Fut}_{\mathcal{A}}(u) \subseteq K$; it is a left factor L_p of K, for some p in Q_K. Set

$$u\varphi = p \ .$$

Then we have $\mathsf{Past}_{\mathcal{A}}(u) \subseteq L_{u\varphi}$ and $\mathsf{Fut}_{\mathcal{A}}(u) \subseteq R_{u\varphi}$. The preceding remark already ensures that $J\varphi \subseteq I_K$ and $U\varphi \subseteq T_K$.

It remains to verify that this map φ is indeed a labelled graph morphism; that is, for all u and v in R we indeed have

$$H_{u,v} \subseteq F_{u\varphi,v\varphi} \ . \tag{4.4}$$

We have: $H_{u,v}\,\mathsf{Fut}_{\mathcal{A}}(v) \subseteq \mathsf{Fut}_{\mathcal{A}}(u)$, as in every automaton,

hence $L_{u\varphi}\,H_{u,v}\,\mathsf{Fut}_{\mathcal{A}}(v) \subseteq L_{u\varphi}\,\mathsf{Fut}_{\mathcal{A}}(u) \subseteq K$, by definition of $u\varphi$,

hence $L_{u\varphi}\,H_{u,v} \subseteq L_{v\varphi}$, by definition of $v\varphi$,

hence $L_{u\varphi}\,H_{u,v}\,R_{v\varphi} \subseteq L_{v\varphi}\,R_{v\varphi} \subseteq K$ from which (4.4), by definition of $F_{u\varphi,v\varphi}$.

If $H_{u,v} \subseteq G$ then $H_{u,v} \subseteq F_{u\varphi,v\varphi} \cap G$ and φ is a morphism from $\mathcal{A}$ to $\mathcal{U}_K^{(G)}$. ■

Remark 4.2 In contrast to all the material we have developed up to now (in this section) which is entirely symmetric between left and right, between past and future, this proof is oriented (since we defined $u\varphi$ seeking the maximal subset such that $P\,\mathsf{Fut}_{\mathcal{A}}(u) \subseteq K$). We could equally have sought the maximal subset such that $\mathsf{Past}_{\mathcal{A}}(u)\,P \subseteq K$, which would have defined *another* morphism, equally oriented. *A priori*, these two morphisms are not the only possible morphisms from $\mathcal{A}$ to $\mathcal{U}_K$. □

The automata $\mathcal{U}_K$ and $\mathcal{U}_K^{(G)}$ are minimal *for this property of universality*; that is, every automaton $\mathcal{V}$ which satisfies Proposition 4.6 (for which there exists a morphism ψ from every automaton $\mathcal{A}$ that accepts K to $\mathcal{V}$) is isomorphic to $\mathcal{U}_K$ and every automaton $\mathcal{W}$ whose transitions have labels included in G and which satisfies Proposition 4.6 is isomorphic to $\mathcal{U}_K^{(G)}$. This results in $\mathcal{U}_K$ and $\mathcal{U}_K^{(G)}$ themselves accepting K, and in the following property.

Proposition 4.6 *Every morphism from $\mathcal{U}_K$ (or $\mathcal{U}_K^{(G)}$) to an automaton that accepts K is (globally) injective.*

Proof. Let ω be a morphism of $\mathcal{U}_K$ (or $\mathcal{U}_K^{(G)}$) in $\mathcal{V}$, an automaton whose behaviour is equal to K (or contained in K). If two distinct states p and q of $\mathcal{U}_K$ have the same image under ω, then

$$L_p \cup L_q \subseteq \mathsf{Past}_{\mathcal{V}}(p\omega) \qquad \text{and} \qquad \mathsf{Past}_{\mathcal{V}}(p\omega)\,\mathsf{Fut}_{\mathcal{V}}(p\omega) \subseteq K \ ,$$

and $\mathsf{Past}_{\mathcal{V}}(p\omega)$ is dominated by a left factor of K. This is a contradiction (with the fact that as L_p and L_q are distinct, their union strictly dominates each of them): the morphism ω must be injective. ■

We are therefore led to define a notion of minimality for automata (over M), as distinct from that of the minimal action of a subset of M; hence, for automata over A^*, from that of the *minimal automaton* of a language. For these last it amounts to a notion of minimality for *non-deterministic* automata.

Definition 4.5 Let $\mathcal{A}$ be an automaton over M and K the subset that it accepts. The automaton $\mathcal{A}$ is called *minimal* if:

(i) the behaviour of every *strict* sub-automaton of the *trim part* of $\mathcal{A}$ is *strictly* contained in K;

(ii) every morphism from $\mathcal{A}$ to an automaton that accepts K is *injective.* □

The definition implies in particular that a minimal automaton contains at most *one* non-accessible state and *one* non-co-accessible state. The automaton $\mathcal{A}$ is minimal if each of its *useful* states is 'necessary' in the sense that if one is removed then some elements (of $|\mathcal{A}|$) are no longer accepted; and that if two are merged, certain elements (of M) are accepted that were not before. The definition is consistent with that of the *minimal automaton* of a language, which can easily have a non-co-accessible state. We deduce by Proposition 4.5: *cf. p. 284*

Corollary 4.7 *Let K be a subset of M and $\mathcal{U}_K$ its universal automaton. Every minimal automaton which accepts K is a sub-automaton of $\mathcal{U}_K$. Every minimal automaton which accepts K and whose labels are in G is a sub-automaton of $\mathcal{U}_K^{(G)}$.* ■

For example, we see in the A-universal automaton of L_1 (Figure 4.4) the two minimal deterministic and minimal co-deterministic automata of L_1, as well as $\mathcal{A}_1$, another minimal automaton (Figure 4.5).

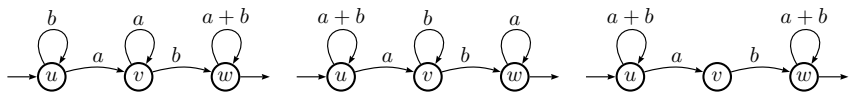
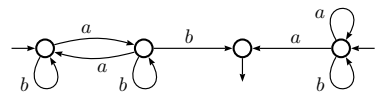


Figure 4.5: Three minimal sub-automata of the A-universal automaton of L_1

Exercises

- 4.9 Compute the $\{1\}$-universal automaton of the subset $\{1,2\}$ of $\mathbb{Z}/3\mathbb{Z}$ (additive notation).
- 4.10 Compute the universal automaton of the language recognised by the adjacent automaton.



4.11 In the case where $M = A^*$, show that the determinisation of the A-universal automaton of a language L is the minimal automaton of L.

4.3 Construction of the universal automaton

The ideas developed above show that, under reasonable assumptions, the universal automaton of a subset K of M is computable. If K is an effectively described recognisable subset, we can compute the (finite) syntactic monoid M_K of K, the image K' of K in M_K, the factorisations, and finally the factors of K', all finite subsets which we can thus obtain at least by simple inspection. The aim of this subsection is to describe a method of direct computation[34] of $\mathcal{U}_K$ in the sense that it avoids the detour via the syntactic monoid.

In the rest of this subsection, we examine the case where M is a free monoid A^*, and K a rational subset of A^*, recognised by a (complete) deterministic automaton $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$. Recall that the action δ of A^* on Q, written $\cdot$, extends to an action of A^* on $\mathfrak{P}(Q)$:

$$\forall f \in A^*\,,\ \forall P \subseteq Q \qquad P \cdot f = \bigcup_{p \in P} p \cdot f \ .$$

4.3.1 Expansion of a deterministic automaton

Definition 4.6 Let $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ be a complete deterministic automaton over A. We give the name *subset expansion*[35] *of* $\mathcal{A}$ to the automaton denoted by $\mathcal{A}_{\mathrm{CB}} = \langle \mathfrak{P}(Q), A, D, J, U \rangle$ and defined as follows:

$$D = \{(P, a, S) \mid P, S \subseteq Q \quad P \cdot a \subseteq S\} \ ,$$

$$J = \{P \subseteq Q \mid i \in P\} \qquad \text{and} \qquad U = \{P \subseteq Q \mid P \subseteq T\} \ . \qquad \square$$

Remark 4.3 A method for constructing $\mathcal{A}_{\mathrm{CB}}$ is to begin by constructing the transitions corresponding to the extension of the action on $\mathfrak{P}(Q)$, adding all the spontaneous transitions $(P, 1_{A^*}, P')$ where $P \subseteq P'$, then eliminating these spontaneous transitions by backward closure. This is done in Figure 4.6 for the complete deterministic automaton that recognises the language $a\,A^*$. $\square$

Property 4.13 $\quad \forall P \subseteq Q \qquad \mathsf{Fut}_{\mathcal{A}_{CB}}(P) = \bigcap_{p \in P} \mathsf{Fut}_{\mathcal{A}}(p)\ .$

[34] Due to Sylvain Lombardy [152].

[35] This construction is due to R. Cohen and J. Brzozowski [64], hence the notation $\mathcal{A}_{\mathrm{CB}}$, even if the connection to the universal automaton was made more recently.

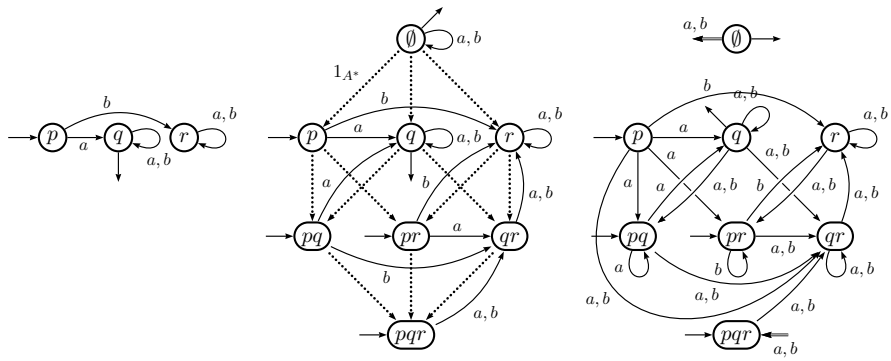


Figure 4.6: Construction of a subset expansion using spontaneous transitions.
To avoid cluttering the right-hand figure, we use a double arrow leaving the state $\emptyset$ (resp. arriving at pqr) to show the fact that there is a transition labelled a, b that leaves $\emptyset$ and arrives at every state in the automaton, including $\emptyset$ itself (resp. that leaves every state, including pqr and arrives at pqr). Similarly, in the middle figure only a minimal set of spontaneous transitions is shown; its transitive closure must be taken to find them all.

Proof. By definition, and for all f in A^*, we have that $P \cdot f$ is final in $\mathcal{A}_{\mathrm{CB}}$ if and only if $p \cdot f$ is final in $\mathcal{A}$ for all p in P, hence the inclusion $\mathsf{Fut}_{\mathcal{A}_{CB}}(P) \supseteq \bigcap_{p \in P} \mathsf{Fut}_{\mathcal{A}}(p)$.

Conversely, if $P \xrightarrow[\mathcal{A}_{CB}]{f} S$, we can easily show by induction on the length of f that $P \cdot f \subseteq S$ and if S is final in $\mathcal{A}_{\mathrm{CB}}$, then $P \cdot f$ is too, hence the reverse inclusion. ■

Note that Property 4.13 holds in particular if $P = \emptyset$ and in this case it reads $\bigcap_{p \in \emptyset} \mathsf{Fut}_{\mathcal{A}}(p) = A^*$. We then conclude:

Property 4.14 *$\mathcal{A}_{CB}$ is equivalent to $\mathcal{A}$.*

Proof. Since $\mathcal{A}$ is a sub-automaton of $\mathcal{A}_{\mathrm{CB}}$, we have $|\mathcal{A}| \subseteq |\mathcal{A}_{\mathrm{CB}}|$. The fact that the future of every initial state of $\mathcal{A}_{\mathrm{CB}}$ is contained in $|\mathcal{A}|$ gives the reverse inclusion. ■

Theorem 4.3 *Let K be a language of A^* recognised by a complete deterministic automaton $\mathcal{A}$. The A-universal automaton $\mathcal{U}_K^{(A)}$ is the minimum quotient of $\mathcal{A}_{CB}$, the expansion of $\mathcal{A}$ (and is therefore computable by Moore's algorithm).*

Proof. For simplicity we shall write $\mathcal{U}_K$ for $\mathcal{U}_K^{(A)}$ from now on. By virtue of Proposition 4.6 it suffices to show that $\mathcal{U}_K$ is a quotient of $\mathcal{A}_{\mathrm{CB}}$: that is, that among the morphisms from $\mathcal{A}_{\mathrm{CB}}$ to $\mathcal{U}_K$ (we know that one exists by Proposition 4.5) we know how to find one that is *totally surjective*.

We define the map $\varphi \colon \mathcal{A}_{\mathrm{CB}} \to \mathcal{U}_K$ by

$$\forall P \subseteq Q \qquad P\varphi = (L, R) \qquad \text{where} \qquad R = \mathsf{Fut}_{\mathcal{A}_{CB}}(P) = \bigcap_{p \in P} \mathsf{Fut}_{\mathcal{A}}(p) \ . \tag{4.5}$$

Since $\mathcal{A}$ is deterministic, $\mathsf{Fut}_{\mathcal{A}}(p) = f^{-1}K$ for all f such that $i \cdot f = p$, Proposition 4.4 and Property 4.13 ensure both that $P\varphi$ is indeed a right factor of K and

that φ is *globally* surjective. It remains to show that φ is a morphism and that it is *Out-surjective*.

If $P \xrightarrow[\mathcal{A}_{CB}]{a} P'$ is a transition of $\mathcal{A}_{\mathrm{CB}}$, it follows that $a\,\mathsf{Fut}_{\mathcal{A}_{CB}}(P') \subseteq \mathsf{Fut}_{\mathcal{A}_{CB}}(P)$ – that is, $a\,R' \subseteq R$ with $R = \mathsf{Fut}_{\mathcal{A}_{CB}}(P)$ and $R' = \mathsf{Fut}_{\mathcal{A}_{CB}}(P')$ – and hence

$$(L,R) \xrightarrow[\mathcal{U}_K]{a} (L',R') \quad \text{is a transition of} \quad \mathcal{U}_K \ .$$

If P is initial in $\mathcal{A}_{\mathrm{CB}}$, then $i \in P$ hence $\mathsf{Fut}_{\mathcal{A}_{CB}}(P) \subseteq K$, hence $1_{A^*} \in L$ if $P\varphi = (L,R)$ and (L,R) is initial in $\mathcal{U}_K$. If P is final in $\mathcal{A}_{\mathrm{CB}}$, then $1_{A^*} \in \mathsf{Fut}_{\mathcal{A}_{CB}}(P) = R$ and $P\varphi = (L,R)$ is final in $\mathcal{U}_K$. Thus φ is a morphism. To show that it is Out-surjective, we first derive another useful property.

Property 4.15 *If (L,R) is a factorisation of K, then $i \cdot L$ is the* maximum *element (for inclusion) of the set of subsets P of Q such that*

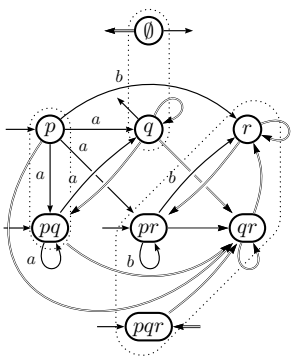
$$\mathsf{Fut}_{\mathcal{A}_{CB}}(P) = R \ . \tag{4.6}$$

Proof. From $L\,R \subseteq K$, we deduce first that $\mathsf{Fut}_{\mathcal{A}_{CB}}(i \cdot L) \supseteq R$. The maximality of R in $L\,R \subseteq K$ ensures the reverse inclusion; the maximality of L (and the accessibility of $\mathcal{A}$) ensures that P satisfies (4.6); we have $P \subseteq i \cdot L$. ∎

Let $(L,R) \xrightarrow[\mathcal{U}_K]{a} (L',R')$ be a transition of $\mathcal{U}_K$, and P included in Q such that $\mathsf{Fut}_{\mathcal{A}_{CB}}(P) = R$. Set $P' = i \cdot L'$; then by Property 4.15 $\mathsf{Fut}_{\mathcal{A}_{CB}}(P') = R'$ hence $P'\varphi = (L',R')$. The hypothesis implies that $L\,a \subseteq L'$ hence $P \cdot a \subseteq (i \cdot L) \cdot a = i \cdot (L\,a) \subseteq i \cdot L' = P'$ and

$$P \xrightarrow[\mathcal{A}_{CB}]{a} P' \quad \text{is a transition of} \quad \mathcal{A}_{\mathrm{CB}} \ .$$

If (L,R) is initial in $\mathcal{U}_K$, then 1_{A^*} belongs to L, $i \cdot L$ contains i and is thus an initial state of $\mathcal{A}_{\mathrm{CB}}$ which is sent by φ to (L,R). Finally, the very definition of φ implies that every state of $\mathcal{A}_{\mathrm{CB}}$ which is sent by φ to a final state of $\mathcal{U}_K$ is final in $\mathcal{A}_{\mathrm{CB}}$: thus, φ is locally surjective. ∎



The adjacent figure shows the effect of the construction on the expansion given in Figure 4.6: the states that are identified in the quotient are outlined in dots (to simplify the figure all transitions labelled $\{a,b\}$ are shown as unlabelled double arrows).

This result is already itself a method for constructing the (A-)universal automaton of a language L without using its syntactic monoid; it implies however the construction of an object whose size is exponential in that of the minimal automaton of L, independent of the final size of the universal automaton. We shall now free ourselves from this detour.

4.3.2 Extraction of the universal automaton

Proposition 4.8 *Let $\mathcal{A}$ be a complete accessible deterministic automaton and $\mathcal{U}_K$ the universal automaton of the language K that it recognises. The application $\psi\colon \mathcal{U}_K \to \mathcal{A}_{CB}$ defined by $(L,R)\psi = i \cdot L$ is an automata morphism, the inverse of the morphism φ defined by (4.5), and the image of $\mathcal{U}_K$ under ψ is a sub-automaton of $\mathcal{A}_{CB}$ isomorphic to $\mathcal{U}_K$.*

Proof. Property 4.15 ensures that $\psi\varphi$ is the identity on the set of states of $\mathcal{U}_K$ and that ψ is (globally) injective.

Let $(L,R) \xrightarrow[\mathcal{U}_K]{a} (L',R')$ and $P = i \cdot L$; there exists, since φ is Out-surjective, a transition $P \xrightarrow[\mathcal{A}_{CB}]{a} S$ such that $S\varphi = (L',R')$: that is, such that $\mathsf{Fut}_{\mathcal{A}_{CB}}(S) = R'$; hence we deduce, again by Property 4.15, that $S \subseteq i \cdot L' = (L',R')\psi = P'$ and $P \xrightarrow[\mathcal{A}_{CB}]{a} P'$ by definition of $\mathcal{A}_{\mathrm{CB}}$. From this we see that ψ is a morphism such that $\psi\varphi$ is the identity on the set of transitions of $\mathcal{U}_K$: thus, ψ is Out-injective and $\mathcal{U}_K$ is a sub-automaton[36] of $\mathcal{A}_{\mathrm{CB}}$. ■

Using this proposition we can identify each state of $\mathcal{U}_K$ with a subset of Q and Q_K with a subset of $\mathfrak{P}(Q)$.

Theorem 4.4 *Let $\mathcal{A}$ be a complete accessible deterministic automaton. Let X be the set of states of $\mathcal{A}_{\mathsf{cod}}$, the co-determinisation of $\mathcal{A}$: X is a subset of $\mathfrak{P}(Q)$. Let $\mathcal{U}_K$ be the universal automaton of the language K recognised by $\mathcal{A}$, and Q_K the set of factorisations of K – that is, of states of $\mathcal{U}_K$ – identified with a subset of $\mathfrak{P}(Q)$ by Proposition 4.8.*

Then, Q_K is the closure under intersection of X.

Proof. The main argument of the proof is the same as that which showed that the determinisation of a co-accessible co-deterministic automaton is the minimal automaton of a language. *cf. Prop. I.3.13, p. 116*

There is a bijection between the *right quotients* of K by the words of A^* and the states of $\mathcal{A}_{\mathsf{cod}}$. More precisely, set $\mathcal{A}_{\mathsf{cod}} = \langle X, A, \eta, V, \{T\} \rangle$, where η is a *left action* of A^* on X, written $\underset{\mathcal{A}_{\mathsf{cod}}}{\cdot}$, and $\{T\}$ is the unique final state. It follows that P is a subset of Q that belongs to X if and only if there exists f in A^* such that $P = f \underset{\mathcal{A}_{\mathsf{cod}}}{\cdot} T$; this allows us to write

$$\begin{aligned} P = f \underset{\mathcal{A}_{\mathsf{cod}}}{\cdot} T &= \{p \in Q \mid f \in \mathsf{Fut}_{\mathcal{A}}(p)\} = \{p \in Q \mid (\mathsf{Past}_{\mathcal{A}}(p))\, f \subseteq K\} \\ &= \{p \in Q \mid (\mathsf{Past}_{\mathcal{A}}(p)) \subseteq K\, f^{-1}\} = i \cdot K\, f^{-1} \end{aligned}$$

since $\mathcal{A}$ is complete accessible deterministic. By Proposition 4.4, (L,R) is a state of $\mathcal{U}_K$ if and only if L is an intersection of *right quotients*:

$$L = K\, f_1^{-1} \cap K\, f_2^{-1} \cap \cdots \cap K\, f_n^{-1} :$$

[36] Here is a case where the terminology is not precise enough: we have just proved that $\mathcal{U}_K$ is isomorphic to the sub-automaton of $\mathcal{A}_{\mathrm{CB}}$ obtained by keeping the states that are the image of Q_K under ψ and *all* the transitions which are adjacent to them.

that is, if and only if

$$(L, R)\psi = P = i \cdot L = \bigcap_{j \in [n]} P_j \ ,$$

where $P_j = K\, f_j^{-1}$ and each of the P_j is a state of $\mathcal{A}_{\mathsf{cod}}$. ■

Remark 4.4 In the above computation it is important to remember the intersection of an empty set of elements of X which gives $P = Q$ even if this set is not a state of $\mathcal{A}_{\mathsf{cod}}$. It corresponds to the factorisation $(A^*, \emptyset)$. □

Example 4.3 (continued) Figure 4.7 shows the construction applied to the deterministic automaton that recognises $a\,A^*$; we see the automaton (given in Figure 4.6), its co-determinisation and the sub-automaton of the expansion isomorphic to the universal automaton of $a\,A^*$ (states in grey). □

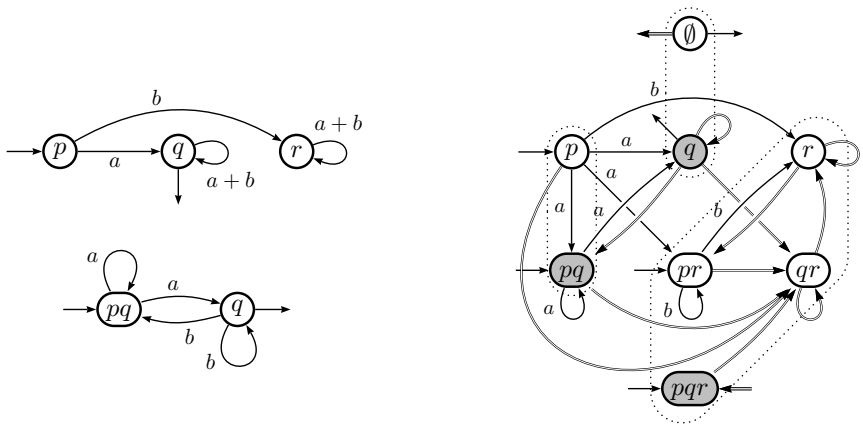
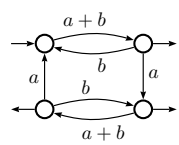


Figure 4.7: Construction of the universal automaton of $a\,A^*$

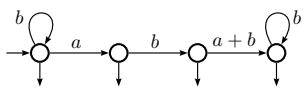
Exercises

4.12 Compute the expansion of the deterministic automaton shown opposite, then the universal automaton of the language that it recognises.

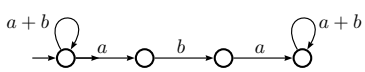


4.13 Compute, by the method of Theorem 4.4, the A-universal automaton of the languages recognised by the following automata.

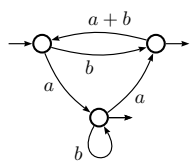
(a)



(b)



(c)



4.4 Language approximations

The construction of the automaton $\mathcal{U}_K^{(G)}$, the G-universal automaton of the subset K, can be seen as a special case of an approximation problem: the reasoning that proves that $\mathcal{U}_K^{(G)}$ accepts K can be generalised to other families of subsets than the generating sets of M, with remarkable results.[37]

Let K be a subset of M and H_1, H_2, …, H_n be a family of n subsets of M. We set $A = \{a_1, a_2, \ldots, a_n\}$ an n-letter alphabet and $\sigma\colon A^* \to M$ to the *substitution* defined by

$$\forall i \in [n] \qquad a_i\sigma = H_i \ .$$

We already know that the language L over A^*, *cf. Cor. 2.18, p. 253*

$$L = \{f \in A^* \mid f\sigma \subseteq K\} \ ,$$

is rational if K is recognisable, but we shall now prove it and give it a more precise interpretation using the universal automaton.

To simplify the propositions, with H_i and hence σ being fixed, we write $\breve{\sigma}$ for the map from $\mathfrak{P}(M)$ to $\mathfrak{P}(A^*)$ defined by

$$\forall K \in \mathfrak{P}(M) \qquad K\breve{\sigma} = \{f \in A^* \mid f\sigma \subseteq K\} \ :$$

that is, $K\breve{\sigma}$ is equal to the language L defined above. The map $\breve{\sigma}$ acts as the *inverse* of the substitution σ but retains only those words whose image under σ is *contained* in K. In other words, $(K\breve{\sigma})\sigma$ is the *best possible approximation* (by default) to K as a *sum of products* of subsets H_i and $K\breve{\sigma}$ describes how this approximation is constructed. With a proof identical to that of Theorem 4.2, we obtain:

Theorem 4.5 *Let $\sigma\colon A^* \to M$ be a substitution. Let K be a subset of M and $\mathcal{U}_K$ its universal automaton. The language $K\breve{\sigma}$ is recognised by the automaton over A^* obtained from $\mathcal{U}_K$ by replacing each label $F_{p,q}$ with the set of letters of A whose image under σ is contained in $F_{p,q}$.* ■

As every subset K is, for every generating set of M, a sum of products of generators, Theorem 4.2 is indeed a particularly interesting special case[38] of Theorem 4.5. If K is recognisable, $\mathcal{U}_K$ has a finite number of states and $K\breve{\sigma}$ is rational: we are back at Corollary 2.18.

The statement of Theorem 4.5 is itself a procedure for computing the 'best approximation' to K, on condition that we know how to compute effectively the factors of K and the inclusion of H_i in these factors. These conditions are fulfilled in particular when considering the rational sets of a free monoid. We then deduce:

Corollary 4.9 *Given a rational language L and a* finite *family $\mathcal{H}$ of rational languages over A^*, we can decide whether L belongs to* $\mathrm{Rat}\,\mathcal{H}$*, the rational closure of $\mathcal{H}$.*

[37] These results were developed by Conway, *op. cit.*, whom once again we follow.

[38] Which justifies giving it a separate proposition.

Proof. We compute the best approximation to L by the n languages of the family $\mathcal{H}$ and then decide whether this approximation is equal to L. ■

cf. Th. I.3.1, p. 110

The problem of approximation is susceptible to a 'dual' approach.[39] The (recognisable) subset K of M having been fixed, instead of choosing the subsets H_i – that is, the substitution $\sigma\colon A^* \to M$ – and trying to compute the language $K\breve{\sigma}$ over A^*, we can choose a language L (not even necessarily rational) over a free monoid A^* and seek a substitution $\sigma\colon A^* \to M$ such that $L\sigma \subseteq K$, which will be called a *sub-solution* of the problem (K, L). The sub-solutions are naturally (and partially) ordered by inclusion of the images of the letters of A, and the interesting sub-solutions are the maximal ones.

Proposition 4.10 *Let K be a subset of M, L a language of A^* and $\sigma\colon A^* \to M$ a maximal sub-solution of the problem (K, L). Then for each a in A, $a\sigma$ is an intersection of factors of K.*

If K is recognisable, these maximal sub-solutions are k-tuples ($k = \|A\|$) of recognisable subsets of M. If in addition M is a free monoid and L is rational, we can effectively compute all the maximal sub-solutions of the problem (K, L).

Example 4.4 A recognisable subset K of M being chosen, let us find all the subsets X such that $X^2 \subseteq K$ and X is maximal for this property (that is, find the maximal sub-solutions of the problem (K, a^2)).

If $X^2 \subseteq K$, (X, X) is a sub-2-factorisation of K, it is dominated by (at least) one factorisation (L_q, R_q), and $X \subseteq L_q \cap R_q$. The maximal sub-solutions are thus all the $L_q \cap R_q$ when q varies over Q_K. □

Proof of Proposition 4.10. Let $f = a_1 a_2 \cdots a_n$ be a word of L. If σ is a solution of (K, L), $a_1\sigma\, a_2\sigma \cdots a_n\sigma \subseteq K$ and by Property 4.9 (and 4.10 as well) there exist $p_0 = l, p_1, \ldots, p_{n-1}, p_n = r$ in Q_K such that

cf. p. 281

$$a_i\sigma \subseteq F_{p_{i-1},p_i}$$

for each i in $[n]$. As these inclusions are verified for each f in L, each $a_i\sigma$ is contained in an intersection of factors and such an intersection is a maximal component in a sub-solution of the problem.

If K is recognisable, there is only a finite number of factors that are all recognisable and their intersections are finite in number and recognisable. There is only a finite number of k-tuples of intersections among which all the maximal sub-solutions are found. If M is free and L is rational we can effectively find all the k-tuples which are sub-solutions and keep only the maximal ones. ■

We should record, to finish, a unsolved problem whose statement is related to finding maximal sub-solutions of (K, L).

[39] This is not the left–right duality of automata, but rather a vector–linear form duality.

Problem 4.1 [**Conway**] *Let K be a subset of M and $\mathcal{N}(K)$ the* normaliser *of K defined as the* largest *subset H such that $KH = HK$. Is this set $\mathcal{N}(K)$ recognisable when K is?*[40]

Exercises

- 4.14 Let K be a recognisable subset of M and A a fixed alphabet. Show that for all substitutions $\sigma\colon A^* \to M$, there is only a finite number of possible languages $K\check{\sigma}$.
- 4.15 Find the best approximation to the language $C_1 = (a^*b + bb^*a)^*$ by the languages $H_1 = a^2$, $H_2 = b$ and $H_3 = ba$.
- 4.16 Give a version of the universal automaton and Theorem 4.5 in terms of rational expressions.
- 4.17 Find all the languages whose square is contained in the language $(a(a+b)+b)^*a$.
- 4.18 Given a recognisable subset K of M, find all the subsets S such that $S^3 \subseteq K$ and that are maximal for this property.
- 4.19 What is the relationship between the maximal solutions of $X^2 \subseteq K$ and the *square root* $\mathcal{R}_2(K)$ computed in Exercise I.8.1 and in § 2.5.1? *cf. p. 175 and p. 251*

5 The importance of being well ordered

A relation α on a set E is an *order relation* if it is reflexive, anti-symmetric and transitive; it is a *quasi-ordering* if it is only reflexive and transitive. In that case, $(x\,\alpha\,y \text{ and } y\,\alpha\,x)$ is an equivalence relation ε_α on E and the image of α in the quotient E/ε_α is an order. In general we write a (quasi-)ordering as $\leqslant$.

An ordering $\leqslant$ on E is an *ordering*, or *total order*, if $x \leqslant y$ or $y \leqslant x$ for every pair (x, y) of elements of E; Otherwise it is a *partial ordering*. The identity relation on E is a special (partial) ordering. An equivalence relation is a quasi-ordering; the order on the corresponding quotient is the identity.

5.1 Well quasi-orderings

Let E be a set quasi-ordered by $\leqslant$. A *chain* of E is a subset all of whose elements are pairwise comparable.

An *antichain* of E is a non-empty subset none of whose elements are pairwise comparable. If $\leqslant$ is an order and S is included in E, we write $\min S$ for the set of minimal elements of S. For every subset S of E, $\min S$ is an antichain if it is non-empty.

[40] *Note added in translation:* A negative answer has been given to this problem by Michal Kunc who showed that, contrary to the intuition of many, $\mathcal{N}(K)$ may well be non recursive, even for *finite* K [143]. For the presentation of this result as a 'problem', see Note IV.28, p. 615.

An *ideal* of E (for $\leqslant$) is a non-empty subset I of E which is closed 'above'; that is, such that every element of E which is greater than an element of I also belongs to I. The ideal generated[41] by a subset B of E is written $\mathsf{I}(B)$:

$$\mathsf{I}(B) = \{x \in E \mid \exists y \in B \quad y \leqslant x\} \ .$$

We say that E has the *finite basis property* if for every subset X of E, there exists a *finite subset* Y such that $Y \subseteq X \subseteq \mathsf{I}(Y)$.

An *ideal chain* is an increasing sequence of ideals of E

$$I_1 \subseteq I_2 \subseteq \cdots \subseteq I_n \subseteq \cdots \ . \tag{5.1}$$

As above, the word 'chain' refers to an order, but not that on E, rather that of inclusion in $\mathfrak{P}(E)$. We say that the ideals of a set E (ideals for the order, like ideals for a monoid structure) satisfy the *ascending chain condition* if every ideal chain is eventually stationary: that is, in the sequence (5.1), there exists an r such that $I_r = I_{r+1} = \cdots = I_{r+s} = \cdots$ for every integer s.

All these notions are integrated in the definition of a *well quasi-ordering.*

Theorem 5.1 [**Higman**] *Let E be a quasi-ordered set. The following properties are equivalent:*

(i) *every ideal of E is finitely generated;*

(ii) *the ideals of E satisfy the ascending chain condition;*

(iii) *E has the finite basis property;*

(iv) *every infinite sequence of elements of E contains an infinite increasing subsequence;*

(v) *every infinite sequence of elements of E contains an increasing subsequence of length* 2*;*

(vi) *every strictly decreasing chain of elements of E is finite and every antichain of E is finite.*

A quasi-ordering that satisfies one of these properties (and hence all) is called a well quasi-ordering*. If it is a partial ordering, it is called a* well partial ordering*; if it is a linear ordering, then it is called a* well ordering.[42] *In the latter case, property (vi) above becomes:*

(vi′) *Every non-empty subset of E contains a minimum element.*

Proof [Sketch]. The equivalence of conditions (i), (ii) and (iii) is a well-known phenomenon[43] which occurs when a closure satisfies, as well as the usual properties, the

[41] Generating an ideal is a *closure operation* in $\mathfrak{P}(E)$; that is, a monotonic operation, isotone and idempotent, and ideals are often called the *closed subsets* of E (*e.g.*, in [118]). This explains the notation of a generated ideal by a bar; but we shall need this notation (in Section IV.4.3) in the context of the free group, where the bar has another meaning. The notation $\mathsf{I}(\cdot)$ is cumbersome and will only be used in a few propositions.

[42] *Well partial ordering* and *well quasi-ordering* are often abbreviated *wpo* and *wqo*.

[43] *Cf.* [34] for example. The most well-known instance is the equivalence of the ascending chain

condition that $x \in \widehat{X}$ implies $x \in \widehat{Y}$ for a finite subset Y of X (possibly dependent on x). The equivalence of conditions (iv), (v) and (vi) is equally well known.[44]

Assume (iii) and let X be the set of elements of the infinite sequence $x_1, x_2, \ldots$. For every finite set $Y \subseteq X$ there exists an index j such that the index of every element of Y is less than j. If $X \subseteq \widehat{Y}$ there exists an x_i in Y such that $x_i \leqslant x_j$ and thus (v) is satisfied. Conversely, let $I_1, I_2, \ldots$ be an infinite strictly increasing chain of ideals. By choosing for each i an a_i in $I_{i+1} \setminus I_i$ we obtain a sequence that contradicts (v); thus (v) entails (ii). ■

One of the fundamental results in the theory of well quasi-orderings rests on the notion of *division*. Let X be an arbitrary set; we will write, for the rest of this section, $\mathrm{V}(X)$ for the free monoid generated by X: that is, the set of finite sequences of elements of X equipped with concatenation.[45] If X is a subset of a monoid M, we have a canonical morphism $\alpha\colon \mathrm{V}(X) \to M$ trivially defined by:

$$(u_1, u_2, \ldots, u_n)\alpha = u_1 \cdot u_2 \cdot \ \cdots \ \cdot u_n \ ,$$

in which the formation of sequences is shown by parentheses and commas and multiplication in M by a dot.

Definition 5.1 Let X be a set equipped with a quasi-ordering $\leqslant$. The relation of *division* $f \trianglelefteq g$ is defined on the set $\mathrm{V}(X)$ in the following manner. For each integer n, X^n is ordered as a direct product:

$$\forall f, g \in X^n \qquad f \leqslant g \qquad \text{if and only if} \qquad \forall i \in [n] \quad f_i \leqslant g_i \ .$$

For two elements f and g of $\mathrm{V}(X)$, $f \trianglelefteq g$ if and only if there exists a subsequence of g which is greater than f (in $X^{|f|}$). □

It is immediate that the division relation is a quasi-ordering on $\mathrm{V}(X)$. Like every relation, a quasi-ordering $\leqslant$ on a monoid M is called *regular* if it is consistent with the product of M: that is, $m \leqslant n$ and $m' \leqslant n'$ imply $m\,m' \leqslant n\,n'$. Division is a regular quasi-ordering on $\mathrm{V}(X)$ (equipped with concatenation). Note that if A is an alphabet equipped with the identity relation (as a quasi-ordering on A) and without any other structure, we can identify A^* and $\mathrm{V}(A^*)$ and obtain: $f \trianglelefteq g$ if and only if f is a *subword* of g. We will also say that f *divides* g.

Theorem 5.2 [**Higman**] *If X is a set equipped with a well quasi-ordering, then division is a well quasi-ordering on* $\mathrm{V}(X)$.

---

condition and the finite basis property for the ideals of a ring, which is found in every treatise on algebra, *cf.* [105, 145].

[44] *Cf.* [118, 140].

[45] This is the notation suggested by Higman; 'V' perhaps stands for *verba*, 'words'. This notation, which could seem redundant given X^*, will become necessary later, when shall use as our basis set X a free monoid A^* or one of its subsets, and we want to distinguish $\mathrm{V}(X)$ from the submonoid generated by X in A^*. Accordingly, the elements of $\mathrm{V}(X)$ will be written explicitly as sequences, not omitting the commas and parentheses as we did from p. 19 onwards (*cf.* Note 0.10, p. 18).

Proof [(adapted from) Conway]. We first state two obvious properties of the division ordering:

$$\forall a, b \in X\,,\ a \not\leqslant b\,,\ \forall u, v \in \mathrm{V}(X) \qquad \begin{cases} \text{(i)} & u \not\trianglelefteq b\,v & \Longrightarrow & u \not\trianglelefteq v \\ \text{(ii)} & a\,u \not\trianglelefteq b\,v & \Longrightarrow & u \not\trianglelefteq v \end{cases}$$

We use condition (v) of Theorem 5.1. Suppose that the set E of infinite sequences in $\mathrm{V}(X)$ with no increasing subsequences of length 2 is non-empty. We construct a 'minimal' element of E in the following manner.

Let u_1 be in $\mathrm{V}(X)$, *of minimal length* such that there exists at least one infinite sequence without increasing subsequences of length 2 which starts with u_1; let u_2 be in $\mathrm{V}(X)$, of minimal length such that there exists at least one infinite sequence without increasing subsequences of length 2 which starts with u_1, u_2; let u_3 be in $\mathrm{V}(X)$, of minimal length such that there exists at least one infinite sequence without increasing subsequences of length 2 which start with u_1, u_2, u_3; *etc.*

Gradually we will therefore have constructed an infinite sequence $\{u_n\}_{n\in\mathbb{N}}$ without increasing subsequences of length 2. We write each u_i in the form $u_i = (a_i, u'_i)$ with a_i in X. As X is partially well ordered, we can extract from the infinite sequence of a_i an infinite increasing subsequence $a_{i_1}, a_{i_2}, \ldots$. It follows from the above properties that the infinite sequence

$$u_1, u_2, \ldots, u_{i_1-1}, u'_{i_1}, u'_{i_2}, \ldots$$

also has no increasing subsequences of length 2, which contradicts the hypothesis that u_{i_1} is of minimal length such that there exists at least one infinite sequence without increasing subsequences of length 2 which start with $u_1, u_2, \ldots, u_{i_1-1}, u_{i_1}$. ■

Corollary 5.1 *Let A be a finite alphabet. Division is a regular partial well ordering on A^*.* ■

cf. Th. 2.3, p. 247

These definitions and results lead first, for our purposes, to a generalisation[46] of the Myhill–Nerode Theorem.

Theorem 5.3 *A subset P of M is recognisable if and only if it is an ideal for a regular well quasi-ordering.*

Proof. By condition (vi) above, it follows that a congruence (that is, a regular equivalence) which is a quasi-ordering is of finite index if and only if it is a well quasi-ordering. The condition is therefore necessary by Theorem 2.3.

Let P be an ideal for a regular well quasi-ordering of M and suppose that ρ_P does not have a finite index. We can then find an infinite set of elements which are not equivalent modulo ρ_P from which we can extract an increasing sequence $\{m_i\}_{i\in\mathbb{N}}$.

For all n and all i, j, $i \leqslant j$, we have $m_i n \leqslant m_j n$. Since P is an ideal for $\leqslant$, $m_i n \in P$ implies $m_i n \in P$ and hence the sequence $\{\mathsf{RC}_P(m_i)\}_{i\in\mathbb{N}}$ forms an increasing

[46] Due to A. Ehrenfeucht, D. Haussler and G. Rozenberg [78].

chain; indeed, it is strictly increasing since m_i and m_j are not equivalent modulo ρ_P. Furthermore, again since P is an ideal, then, if $n \leqslant n'$, $m_i n \in P$ implies $m_i n' \in P$ which means that each of the $\mathsf{RC}_P(m_i)$ is an ideal. This is a contradiction: ρ_P has a finite index and P is recognisable. ■

Corollary 5.2 *Let A be a finite alphabet. Every ideal of A^* for division is a rational language.* ■

Exercise

• 5.1 Prove Corollary 5.2 without using Theorem 5.3 (but basing your proof on Corollary 5.1).

5.2 Derivations

Let A be an arbitrary (finite) alphabet and Π a non-empty set[47] of words of A^*; this set will allow us to define a *rewriting system*.

We write $\underset{\Pi}{\vdash\!\!-}$ for the regular closure of the set $(\Pi, 1_{A^*})$: that is, for f and g in A^*, we have

$$f \underset{\Pi}{\vdash\!\!-} g \quad \text{if and only if} \quad \exists u, v \in A^*,\ \exists x \in \Pi \quad f = u\,x\,v \quad \text{and} \quad g = u\,v\ .$$

Logically, we write $\overset{\star}{\underset{\Pi}{\vdash\!\!-}}$ for the reflexive and transitive closure of $\underset{\Pi}{\vdash\!\!-}$: that is $f \overset{\star}{\underset{\Pi}{\vdash\!\!-}} g$ if there exists a *chain of derivations* from f to g, a sequence $f_0, f_1, \ldots, f_n$ *cf. Prop. 0.1.2, p. 13* of words in A^* such that $f = f_0$, $g = f_n$, and if, for all i, $0 \leqslant i < n$, $f_i \underset{\Pi}{\vdash\!\!-} f_{i+1}$. Since each derivation 'reduces' the length of the words, such a chain is necessarily finite.

Property 5.1 $\forall f, g, h, k \in A^* \quad f \overset{\star}{\underset{\Pi}{\vdash\!\!-}} g \ \textit{and}\ h \overset{\star}{\underset{\Pi}{\vdash\!\!-}} k \implies f\,h \overset{\star}{\underset{\Pi}{\vdash\!\!-}} g\,k$. ■

We write θ_Π for the relation from A^* to itself which associates with each word f the set of words which are obtained by derivation from f:

$$\forall f \in A^* \qquad f\theta_\Pi = \{g \in A^* \mid f \overset{\star}{\underset{\Pi}{\vdash\!\!-}} g\}\ .$$

The relation $\leqslant_\Pi$, defined by $g \leqslant_\Pi f$ if and only if $g \in f\theta_\Pi$, is a regular ordering on A^*. If $\Pi = A$, the ordering $\leqslant_\Pi$ coincides with division on A^*.

Definition 5.2 A subset Π of A^* is called *inevitable* if there exists only a finite number of words in A^* which do not contain factors in Π: that is, if $A^* \setminus A^*\Pi A^*$ is finite. If Π is inevitable, then the *index*[48] of Π is the smallest integer k such that every word of length greater than k contains at least one factor in Π: that is, such that $A^* \setminus A^*\Pi A^* \subseteq A^{\leqslant k}$. □

We shall show the following generalisation of Corollary 5.1.

[47] We use a Greek letter since the same procedure will allow us to define a congruence in Section 6.

[48] This is not an expressive term, but we shall only use it in this section.

Theorem 5.4 [**Ehrenfeucht, Haussler & Rozenberg**] *Let A be a finite alphabet. The ordering $\leqslant_\Pi$ associated with a finite subset Π of A^* is a partial well ordering if and only if Π is an inevitable set of A^*.*

5.2.1 Preparations

We start by showing a series of properties of well quasi-orderings, which will also be useful afterwards and which for that reason are stated in more general terms than required just to prove Theorem 5.4. We shall later consider not only ordered sets but also maps between these ordered sets, compatible with this structure.

Definition 5.3 If E and F are two sets equipped with a quasi-ordering, a map $\alpha\colon E \to F$ is a *morphism* (of quasi-ordered sets) if $a \leqslant b \Rightarrow a\alpha \leqslant b\alpha$. □

Condition (v) of Theorem 5.1 immediately gives us:

Properties 5.2 *Let E be a set equipped with a well quasi-ordering $\leqslant$.*
(i) *The trace of $\leqslant$ on every subset of E is a well quasi-ordering.*
(ii) *Every image of E under a morphism is equipped with a well quasi-ordering.* ■

If E and F are equipped with a well quasi-ordering, the *direct product* $E \times F$ is canonically equipped with a well quasi-ordering defined by $(x_1, y_1) \leqslant (x_2, y_2)$ if and only if $x_1 \leqslant x_2$ *and* $y_1 \leqslant y_2$.

Property 5.3 *The quasi-ordering on $E \times F$ is a well quasi-ordering if the quasi-orderings on E and F are.*

Proof. From an infinite sequence (x_i, y_i), $i = 1, 2, \ldots$, we first extract an infinite subsequence in which the x_i are increasing; from that we extract an infinite subsequence in which the y_i are increasing; we thus obtain an infinite increasing subsequence from the (x_i, y_i). ■

Property 5.4 *Let $\leqslant$ be a regular quasi-ordering on A^*. The map $\pi\colon A^* \times A^* \to A^*$ defined by $(f, g)\pi = f\,g$ is a morphism (of quasi-ordered sets).* ■

From these last two properties, we deduce:

Lemma 5.3 *Let $\leqslant$ be a regular quasi-ordering on A^*, and R and S, be contained in A^*, on which the trace of $\leqslant$ is a well quasi-ordering. Then, $\leqslant$ is a well quasi-ordering on RS.* ■

Lemma 5.4 *Let $\leqslant$ be a regular quasi-ordering on A^*. Let S be contained in A^* on which the trace of $\leqslant$ is a well quasi-ordering, and such that $1_{A^*} \leqslant s$ for all s in S. Then, $\leqslant$ is a well quasi-ordering on S^*.*

Proof. We shall show that the canonical morphism $\alpha\colon \mathrm{V}(S) \to A^*$ is a morphism of quasi-ordered sets (when $\mathrm{V}(S)$ is ordered by division relative to $\leqslant$). The conclusion follows from $(\mathrm{V}(S))\,\alpha = S^*$ and Theorem 5.2.

Suppose $(s_1, s_2, \ldots, s_p) \trianglelefteq (t_1, t_2, \ldots, t_n)$; there exists therefore an increasing sequence with indices $i_1, i_2, \ldots, i_p$ such that for all j, $s_j \leqslant t_{i_j}$. We then construct the sequence $(s'_1, s'_2, \ldots, s'_n)$ with

$$s'_l = \begin{cases} s_j & \text{if} \quad \exists j \in [p] \quad l = i_j\,; \\ 1_{A^*} & \text{otherwise}\ . \end{cases}$$

By hypothesis, $s'_l \leqslant t_l$ for all l in $[n]$ and we have

$$(s_1, s_2, \ldots, s_p)\alpha = (s'_1, s'_2, \ldots, s'_n)\alpha \leqslant (t_1, t_2, \ldots, t_n)\alpha\ ,$$

since $\leqslant$ is regular. ■

5.2.2 Demonstration of Theorem 5.4

The composition of a chain of derivations (modulo θ_Π) hides operations of a very different nature: 'parallel derivations' and 'nested derivations', which better notation will allow us to disentangle.

We shall write $\underset{\Pi}{\models\!\!=}$ for parallel derivation: that is, the relation whose graph is $[\widehat{\iota}(\Pi, 1_{A^*})\widehat{\iota}]^*$, or, more explicitly,

$$\forall f, g \in A^* \qquad f \underset{\Pi}{\models\!\!=} g \qquad \text{if and only if}$$
$$f = u_0 x_1 u_1 x_2 \cdots u_{n-1} x_n u_n\ , \qquad g = u_0 u_1 \cdots u_{n-1} u_n\ ,$$
$$\text{with} \quad u_i \in A^* \quad \text{and} \quad x_i \in \Pi\ .$$

From $\underset{\Pi}{\vdash} \subseteq \underset{\Pi}{\models\!\!=} \subseteq \overset{\star}{\underset{\Pi}{\vdash}}$ we deduce that $\overset{\star}{\underset{\Pi}{\models\!\!=}} = \overset{\star}{\underset{\Pi}{\vdash}}$. Set, for all n in $\mathbb{N}$,

$$\Psi_\Pi^{(n)} = \{f \in A^* \mid f \overset{n}{\underset{\Pi}{\models\!\!=}} 1_{A^*}\} \qquad \text{and} \qquad \Omega_\Pi^{(n)} = \{f \in A^* \mid \exists g \in A^{\leqslant n} \quad f \overset{n}{\underset{\Pi}{\models\!\!=}} g\}\ .$$

We easily obtain:

Property 5.5

$$\Psi_\Pi^{(1)} = \Pi^*\ ,$$
$$\forall n \in \mathbb{N} \qquad \Psi_\Pi^{(n+1)} = \left(\left[\bigcup_{\substack{a_1 a_2 \cdots a_p \in \Pi \\ p \leqslant n}} \Psi_\Pi^{(n)} a_1 \Psi_\Pi^{(n)} a_2 \Psi_\Pi^{(n)} \cdots \Psi_\Pi^{(n)} a_p \Psi_\Pi^{(n)}\right] \cup \Psi_\Pi^{(n)}\right)^*\ ,$$
$$\Omega_\Pi^{(n)} = \bigcup_{\substack{a_1, a_2, \ldots, a_p \in A \\ p \leqslant n}} \Psi_\Pi^{(n)} a_1 \Psi_\Pi^{(n)} a_2 \Psi_\Pi^{(n)} \cdots \Psi_\Pi^{(n)} a_p \Psi_\Pi^{(n)}\ .$$ ■

Some combinatorial properties of the sets $\Psi_\Pi^{(n)}$ and $\Omega_\Pi^{(n)}$ are then established.

Property 5.6 *For every finite subset $\Pi \subseteq A^*$ and every integer $n \geqslant 1$:*
(i) *if $f\,g \in \Psi_\Pi^{(n)}$ and $x \in \Pi$, then $f\,x\,g \in \Psi_\Pi^{(n+1)}$;*
(ii) *if $f\,g \in \Psi_\Pi^{(n)}$ and $x \in \Pi$ and if $|f| < n$, then $f\,x\,g \in \Psi_\Pi^{(n)}$.*

Proof. If $f\,g \overset{n}{\underset{\Pi}{=\!=}} 1_{A^*}$, then $f\,x\,g \overset{n+1}{\underset{\Pi}{=\!=}} 1_{A^*}$, which proves (i).

We prove (ii) by induction on n. The case $n = 1$ is obvious: $f = 1_{A^*}$, $\Pi \subseteq \Psi_\Pi^{(n)}$ and $\Psi_\Pi^{(n)}$ is a submonoid. If $f\,g \in \Psi_\Pi^{(n+1)}$, there exist p and j such that

$$f = w_0 a_1 w_1 a_2 w_2 \cdots w_{j-1} a_j w'_j \qquad \text{and} \qquad g = w''_j a_{j+1} w_{j+1} \cdots w_{p-1} a_p w_p$$

where we have set

$$w_j = w'_j w''_j \ , \quad a_1 a_2 \cdots a_p \in \Pi \qquad \text{and} \quad \forall i\,,\ 0 \leqslant i \leqslant p\,, \quad w_i \in \Psi_\Pi^{(n)} \ .$$

If $j = 0$, we have $w'_0\,x\,w''_0 \in \Psi_\Pi^{(n+1)}$ by (i), and (ii) is proved since

$$a_1 w_1 a_2 w_2 \cdots w_{p-1} a_p w_p \in \Psi_\Pi^{(n+1)} \ .$$

If $j \geqslant 1$, $|w'_j| < n$ and by induction it follows that $w'_j\,x\,w''_j \in \Psi_\Pi^{(n)}$, hence (ii). ■

From Lemmas 5.3 and 5.4 and from Property 5.5 it follows that:

Property 5.7 *For every finite subset $\Pi \subseteq A^*$ and every integer $n \geqslant 1$, the order $\leqslant_\Pi$ is a partial well ordering on $\Psi_\Pi^{(n)}$ and on $\Omega_\Pi^{(n)}$.* ■

The heart of the proof of Theorem 5.4 rests in the following proposition.

Proposition 5.5 *Let Π be an inevitable finite set of A^* and k its index. Then $\Omega_\Pi^{(k)} = A^*$.*

Proof. Suppose that $A^* \setminus \Omega_\Pi^{(k)}$ is non-empty and choose a word h of minimal length; h is longer than k and the prefix of h of length $k+1$ contains a factor in Π, namely, $h = f\,x\,g$ where $x \in \Pi$ and $|f| \leqslant k$.

The minimality of h entails that $f\,g$ belongs to $\Omega_\Pi^{(k)}$ and hence there exist p and j, $p \leqslant k$ and $0 \leqslant j \leqslant p$, such that we can write

$$f = w_0 a_1 w_1 a_2 w_2 \cdots w_{j-1} a_j w'_j \qquad \text{and} \qquad g = w''_j a_{j+1} w_{j+1} \cdots w_{p-1} a_p w_p$$

where we have, for every i, $0 < i \leqslant p$, $a_i \in A$ and $w_i \in \Psi_\Pi^{(k)}$ (with the convention that $w_j = w'_j w''_j$). It follows that $|w'_j| \leqslant |f| \leqslant k$ from which, by Property 5.6 (ii), $w'_j x\, w''_j$ belongs to $\Psi_\Pi^{(k)}$ and h to $\Omega_\Pi^{(k)}$: this is a contradiction. ■

Proof of Theorem 5.4. If Π is not an inevitable set of A^*, we can find an infinite set X of words none of which contains a factor in Π: thus, X is an antichain for $\leqslant_\Pi$ which is not a partial well ordering. If Π is an inevitable set of A^* with index k, then $\leqslant_\Pi$ is a partial well ordering on $\Omega_\Pi^{(k)} = A^*$. ■

Corollary 5.6 *Let Π be a finite inevitable set of A^* and K an arbitrary subset of A^*. The language $K\theta_\Pi^{-1} = \{f \mid \exists g \in K \quad g \in f\theta_\Pi\}$ is rational.* ■

DEEPER & BROADER

6 Rationals in the free group

The study of rational subsets of the free group is justified, indeed, compelled, as much for the coherence of our programme as by the further developments to which it leads.

These notions, which we have studied for general monoids, should, if they are meaningful, assume a special form in structures as fundamental as groups, and relate to already familiar concepts. Our hope shall not be in vain.

A group is a group because each of its elements possesses an *inverse*, another element connected with it (one might say, *that recognises it*) and whose presence can annihilate it. This twofold nature of the inverse perfectly models several mechanisms of computation and communication. Thus, the transitions of a *pushdown automaton* are described by the product of subsets of the free group. In this section we lay the foundation of a truly mathematical theory of *context-free* languages.

This section contains four subsections. We begin by considering recognisable sets and rational sets in general groups. We turn next to the description of the free group, seen as a set of words over a symmetric alphabet. The rational subsets of the free group are characterised in the third subsection, and this is the main result of the section (Theorem 6.1). To finish, we use techniques we have developed to give a description of the *regular canonical systems* introduced by J. Büchi as a definition[49] equivalent to finite automata.

6.1 Recognisable and rational in groups

Essentially, this subsection consists of three propositions that show how the concepts defined in terms of automata in arbitrary monoids correspond to classic notions of group theory. The first says that we can 'translate' *be of finite index* by *recognisable*, the second that being *finitely generated* is the same as being *rational* for subgroups. The third finally states a *Fatou property* for subgroups.[50]

6.1.1 Recognisable subsets

In the following, G is a group and H a subgroup of G. A (right) *coset* of H is a subset of the form Hg for some g in G. The group structure gives the elementary but fundamental property:

[49] In [45].

[50] We explain what such a property is below and justify the terminology in Chapter III (*added in translation:* at least with a note, see p. 520).

Property 6.1 *For all g' and g'' in G, the cosets Hg' and Hg'' are either equal or disjoint.*

Proof. If $k = h'g' = h''g''$, $g' = h'^{-1}h''g''$, then $g'' = h''^{-1}h'g'$ and $Hg' = Hg''$. ■

Hence we deduce that (right) multiplication defines a (right) action of G on the set of cosets of H

$$Hk \cdot g = Hkg \qquad \text{for } g \text{ and } k \text{ in } G \ ,$$

in the sense of Def. 2.1, p. 234

and that the set of g that send H, the distinguished coset, to a fixed coset Hk is exactly Hk:

$$g \in Hk \quad \Longrightarrow \quad Hg = Hk \quad \Longrightarrow \quad g \in Hk \ .$$

A *subgroup* H of G is *of finite index* if it has only a finite number of distinct cosets. A *subset* of G is called *of finite index* if it is the union of cosets of a subgroup of finite index.

Proposition 6.1 *A subset of a group G is recognisable if and only if it is of finite index.*

Proof. Let P be a union of cosets of H, a subgroup of finite index of G. Then, $P = \{g \in G \mid Hg \subseteq P\}$; that is, P is recognised by the action of G on the cosets of H, where we obviously take H as the distinguished coset and the set of cosets that form P as the final subset.

Conversely, let δ be an action of G over a finite set S and let s be an arbitrary element of S. The set of elements of G that 'fix' s by δ: that is, $H = \{g \mid s \cdot g = s\}$, is a subgroup of G; the equalities

$$s \cdot (g\,g') = (s \cdot g) \cdot g' = s \cdot g' = s \qquad \text{and} \qquad s \cdot g^{-1} = (s \cdot g) \cdot g^{-1} = s \cdot 1_G = s$$

show that H is closed under product and contains the inverse of each of its elements.

For all t in S, the set of g in G that 'sends' s to t is a coset of H: $s \cdot g = t$ and $s \cdot k = t \Longrightarrow s \cdot k\,g^{-1} = s \Longrightarrow k = (k\,g^{-1})g \in H$, and every coset of H sends s to the same element of S.

Thus, every subset recognised by δ is a union of cosets for the subgroup that fixes the distinguished element of S, and this subgroup is of finite index. ■

6.1.2 Rational subgroups

Proposition 6.2 *A subgroup of a group G is a rational subset of G if and only if it is finitely generated.*

Proof. If H is the subgroup generated by the subset F of G, then H is the *submonoid* generated by $F \cup F^{-1}$: that is, $H = \langle F \cup F^{-1} \rangle = (F \cup F^{-1})^*$ and H is rational when F is finite.

cf. Th. 1.1, p. 225

Conversely, suppose H is rational and let $\mathcal{A} = \langle Q, G, E, I, T \rangle$ be an automaton over G that accepts H.

We shall say, in imitation of graph theory, that a computation in $\mathcal{A}$ is *simple* if it corresponds to a simple path: that is, if it does not go through the same state twice (except perhaps the source, which can be equal to the destination). The length of a simple computation of $\mathcal{A}$ is bounded by $\|Q\|$. We shall say that a computation c is *almost simple* if it is not simple and if, writing

$$c := p \xrightarrow{u} q \xrightarrow{v} q \xrightarrow{w} r \ , \qquad (*)$$

the two computations $p \xrightarrow{u} q \xrightarrow{w} r$ and $q \xrightarrow{v} q$ are simple.[51] The length of an almost simple computation of $\mathcal{A}$ is bounded by $2\,\|Q\|$.

Let S be the set of labels of successful simple computations of $\mathcal{A}$ and V the set of labels of successful almost simple calculations of $\mathcal{A}$. We show that H is generated, as a subgroup of G, by $S \cup V$. (In fact, that H is generated, as a submonoid of G, by $S \cup S^{-1} \cup V$.) Proceeding by *reductio ad absurdum*, let c be a successful computation of $\mathcal{A}$ of minimal length such that its label h does not belong to $\langle S \cup S^{-1} \cup V \rangle$. We can always write the computation in the form

$$c := i \xrightarrow{g} q \xrightarrow{v} q \xrightarrow{w} t \ ,$$

with i in I, t in T and the computation $q \xrightarrow{v} q \xrightarrow{w} t$ such that there exists an almost simple computation $j \xrightarrow{u} q \xrightarrow{v} q \xrightarrow{w} t$ with j in I. It then follows that

$$h = g\,v\,w = g\,w\,(u\,w)^{-1} u\,v\,w$$

is in $\langle S \cup S^{-1} \cup V \rangle$ since $g\,w$ is, by the minimality hypothesis on the computation c, and that uvw and $(uw)^{-1}$ are too by definition: contradiction. ■

If G is finitely generated, we have $\operatorname{Rec} G \subseteq \operatorname{Rat} G$ (Proposition 2.5). As a corollary of Propositions 6.1 and 6.2, we then obtain the classic property: *p. 239*

Proposition 6.3 *A subgroup of finite index of a finitely generated group is finitely generated.* ■

6.1.3 A Fatou property

A rational subset, not only of a group G but more generally of an arbitrary monoid M, can be seen as the *result* of a computation – a sequence of rational operations – on a finite set of operands taken from M. We shall say that a submonoid N of M has *the Fatou property* if every rational subset of M contained in N is in fact a rational subset of N: that is to say, if we are in a situation where a *property of the result* of a computation (here, being in N) gives us information on *the way in which the calculation can be performed* (here, we can take the operands in N). We know, for example, that a submonoid of a free monoid does not necessarily have the Fatou property. The group structure gives the remarkable property that every subgroup has the Fatou property. *cf. Exer. 1.5, p. 225*

[51] Note that the decomposition of c into the form $(*)$ is not necessarily unique.

Proposition 6.4 *Let G be a group and H a subgroup of G. A rational subset of G contained in H is a rational subset of H: that is,*

$$P \in \operatorname{Rat} G \quad \textit{and} \quad P \subseteq H \quad \Longrightarrow \quad P \in \operatorname{Rat} H \ .$$

cf. Sec. I.6, p. 157 *Proof.* By induction on the star height of a rational expression that describes P. The property is true for height 0 since P is then a finite, hence rational, subset of H.

Suppose that P is of height h. It can be written as the finite union of monomials R, contained in H, in the form

$$R = g_0 K_1^* g_1 K_2^* g_2 \cdots g_{n-1} K_n^* g_n \ ,$$

where the g_i are in G and the K_i are (rational subsets) of height at most $h-1$. It is enough to prove the property for each R.

Taking the power 0 in each star, we have that $g_0 g_1 \cdots g_n$ is in H, hence so is the element $k = g_n^{-1} g_{n-1}^{-1} \cdots g_0^{-1}$. Taking the power 0 in all the stars except the first where we take the power 1, and multiplying the result on the right by k, it follows that

$$L_1 = g_0 K_1 g_0^{-1}$$

is included in H; it is a set of star height less than or equal to $h-1$ to which we can apply the induction hypothesis: L_1 is a rational subset of H. Likewise, the sets

$$L_2 = g_0 g_1 K_2 (g_0 g_1)^{-1} \ , \quad \cdots \ , \quad L_n = (g_0 g_1 \cdots g_{n-1}) K_n (g_0 g_1 \cdots g_{n-1})^{-1} \ ,$$

are in H and of star height less than or equal to $h-1$: they are in $\operatorname{Rat} H$.

The computation in G implies that $L_1^* = g_0 K_1^* g_0^{-1}$, $L_2^* = g_0 g_1 K_2^* (g_0 g_1)^{-1}$, *etc.*, that $L_1^* L_2^* = g_0 K_1^* g_1 K_2^* (g_0 g_1)^{-1}$, *etc.*, and it remains only to write R in the form

$$R = L_1^* L_2^* \cdots L_n^* (g_0 g_1 \cdots g_n)$$

to show that R is also a rational subset of H. ■

Exercises

• 6.1 What is the syntactic monoid of a subset P of G? Show that a singleton of an infinite group is disjunctive. State Proposition 2.1 for groups. *p. 235*

6.2 Describe the recognisable subsets of $\mathbb{Z}$.

• 6.3 Show that $\operatorname{Rec} G$ is closed under product.

6.4 Let G be a group (generated by A). The Cayley graph of G (associated with A) in which we choose 1_G as the initial and final state is the minimal automaton $\mathcal{A}$ over G that recognises 1_G. Show that the (A-)universal automaton associated with the subset $\{1_G\}$ is isomorphic to $\mathcal{A}$.

• 6.5 Show that the loop complexity of the Cayley graph of the additive group $\mathbb{Z}/2^q\mathbb{Z}$ with $\{1, -1\}$ as the generating set is q.

6.2 Description of the free group

For many moons mathematicians have been studying certain properties of groups, and particularly those of the free group, based on the properties of rewriting operations on words representing the elements of the group, these rewritings therefore representing the product operation in the group under consideration.[52] It seems to me that this pair of a set of words over a free monoid, seen as representing the elements of the monoid and of an operation on the words of this free monoid that implements multiplication, could be called a *description*[53] of the monoid, or more specifically of a group.

With this point of view, we see that what will interest us in the multiplicative structure of the free group – as much because of the properties that we can deduce for its rational subsets as for its use as a model which we shall take up later – is founded on a combinatorial property of certain rewriting systems which we shall call *simplifications*. We will thus be able to deal at the same time with two related but distinct structures: the free group and the *involutive monoid* which a more natural (and simpler) model for the computation of pushdown automata.

6.2.1 Dyck congruence and Dyck words

Let A be a finite alphabet. We shall write $\overline{A}$ for a *copy* of A: an alphabet disjoint from A and in bijection with A; for each letter a of A there is a corresponding letter $\overline{a}$. We write

$$\widetilde{A} = A \cup \overline{A} \ .$$

The 'bar' operation is extended to an *involution* first over $\widetilde{A}$: that is, $\overline{\overline{a}} = a$ for all a in A; then over all of $\widetilde{A}^*$, by

$$f = a_1 a_2 \cdots a_n \quad \Longrightarrow \quad \overline{f} = \overline{a}_n \overline{a}_{n-1} \cdots \overline{a}_1 \ .$$

Example 6.1 $A = \{a, b\}$, $\widetilde{A} = \{a, b, \overline{a}, \overline{b}\}$, and $\overline{a\,\overline{b}\,\overline{a}\,b\,b} = \overline{b}\,\overline{b}\,a\,b\,\overline{a}$. □

It follows that:

Property 6.2 $\forall f, g \in \widetilde{A} \qquad \overline{f\,g} = \overline{g}\,\overline{f}$. ■

A *symmetric alphabet* is an alphabet such that $\widetilde{A}$ is implicitly equipped with the 'bar' involution. We write Δ_A for the set of the following pairs of words of $\widetilde{A}^*$:

$$\Delta_A = \{(z\,\overline{z}, 1_{\widetilde{A}^*}) \mid z \in \widetilde{A}\} \ .$$

We write $[\Delta_A]$ for the congruence on $\widetilde{A}^*$ generated by Δ_A, $F(A)$ for the quotient of $\widetilde{A}^*$ by $[\Delta_A]$, and γ_A for the canonical morphism from $\widetilde{A}^*$ to $F(A)$:

$$\gamma_A \colon \widetilde{A}^* \longrightarrow F(A) \cong \widetilde{A}^*/[\Delta_A] \ .$$

[52] *Cf.* the treatises of Magnus, Karrass and Solitar [162] and of Lyndon and Schupp [161].

[53] Since *representation* and *presentation* have already been given precise mathematical senses.

The set of words equivalent to $1_{\widetilde{A}^*}$ modulo Δ_A is a submonoid of $\widetilde{A}^*$ (the inverse image of a submonoid). We write D_A for its minimum generating set:

$$D_A^* = (1_{F(A)})\gamma_A^{-1} = \{f \mid f \equiv 1_{\widetilde{A}^*} \bmod [\Delta_A]\} \ .$$

We call $[\Delta_A]$ the *Dyck congruence*, and D_A^* the set of *Dyck words* over A (or over $\widetilde{A}$).

Property 6.3 *For all f in $\widetilde{A}^*$, $f\,\overline{f} \equiv \overline{f}\,f \equiv 1_{\widetilde{A}^*} \bmod [\Delta_A]$.*

Proof. By induction on the length of words: $1_{\widetilde{A}^*}\overline{1_{\widetilde{A}^*}} = 1_{\widetilde{A}^*}1_{\widetilde{A}^*} = 1_{\widetilde{A}^*}$ and

$$\forall f \in \widetilde{A}^*\,,\ \forall z \in \widetilde{A} \qquad f\,z\,\overline{f\,z} = f\,z\,\overline{z}\,\overline{f} \equiv f\,\overline{f} \equiv 1_{\widetilde{A}^*} \bmod [\Delta_A] \ . \qquad \blacksquare$$

From this we conclude that $F(A)$ *is a group* generated by A. Indeed, it is the *free group* generated by A. Before defining generalised simplifications, we present another congruence that is often used in automata theory and which uses the same techniques as the Dyck congruence.

cf. Exer. 6.8

6.2.2 Shamir congruence and parenthetic words

We again consider a symmetric alphabet $\widetilde{A}$. In the Dyck congruence, each letter a simplifies to right *and* left with the letter $\overline{a}$: $a\,\overline{a} \equiv 1_{\widetilde{A}^*}$ *and* $\overline{a}\,a \equiv 1_{\widetilde{A}^*}$. In the Shamir congruence, which we shall now define, the same simplification only takes place on the right, as if a were a sort of opening parenthesis, and $\overline{a}$ the corresponding closing parenthesis.

We write Σ_A for the set of following pairs of words of $\widetilde{A}^*$:

$$\Sigma_A = \{(a\,\overline{a}, 1_{\widetilde{A}^*}) \mid a \in A\} \ .$$

We write $[\Sigma_A]$ for the congruence on $\widetilde{A}^*$ generated by Σ_A, $I(A)$ for the quotient of $\widetilde{A}^*$ by $[\Sigma_A]$, and β_A for the canonical morphism from $\widetilde{A}^*$ to $I(A)$:

$$\beta_A \colon \widetilde{A}^* \longrightarrow I(A) \cong \widetilde{A}^*/[\Sigma_A] \ .$$

The set of words equivalent to $1_{\widetilde{A}^*}$ modulo $[\Sigma_A]$ is a submonoid of $\widetilde{A}^*$, and we write P_A for its minimum generating set:

$$P_A^* = (1_{I(A)})\beta_A^{-1} = \{f \mid f \equiv 1_{\widetilde{A}^*} \bmod [\Sigma_A]\} \ .$$

We call $[\Sigma_A]$ the *Shamir congruence*, and P_A^* the set of *parenthetic words* over A (or over $\widetilde{A}$).

Property 6.4 *For all f in A^*, $f\,\overline{f} \equiv 1_{\widetilde{A}^*} \bmod [\Sigma_A]$.*

Proof. By induction on the length of words, $1_{\widetilde{A}^*}\overline{1_{\widetilde{A}^*}} = 1_{\widetilde{A}^*}1_{\widetilde{A}^*} = 1_{\widetilde{A}^*}$ and

$$\forall f \in A^*\,,\ \forall a \in A \qquad f\,a\,\overline{f\,a} = f\,a\,\overline{a}\,\overline{f} \equiv f\,\overline{f} \equiv 1_{\widetilde{A}^*} \bmod [\Sigma_A] \ . \qquad \blacksquare$$

The quotient $I(A)$ is not a group; we call it the *involutive monoid*[54] (generated by A).

[54] Term proposed by M. Fliess [89]. The original name given by E. Shamir was *free half group* [231].

6.2.3 Simplifications

Let A be an arbitrary (finite) alphabet and let Π be a subset of A^2. We say that Π is a *simplification* (over A) if the following property is satisfied:

$$\forall a, b, c \in A \qquad a\,b \in \Pi \quad \text{and} \quad b\,c \in \Pi \quad \Longrightarrow \quad a = c \ . \tag{6.1}$$

We shall also use the term 'simplification', and write $[\Pi]$ – a slight misuse – for the congruence on A^* generated by the set of pairs $\{(a\,b, 1_{A^*}) \mid a\,b \in \Pi\}$. It is clear that $[\Delta_A]$ and $[\Sigma_A]$ are two simplifications over $\widetilde{A}^*$.

We write T_Π for the quotient of A by $[\Pi]$, and α_Π for the canonical morphism from A to T_Π:

$$\alpha_\Pi \colon A^* \longrightarrow T_\Pi \cong A^*/[\Pi] \ .$$

The set of words equivalents to 1_{A^*} modulo $[\Pi]$ is a submonoid of A^* and we write U_Π for its minimum generating set:

$$U_\Pi^* = (1_{T_\Pi})\alpha_\Pi^{-1} = \{f \mid f \equiv 1_{A^*} \bmod [\Pi]\} \ .$$

The expected properties of multiplication in T_Π will be shown using the properties of the *rewriting system* generated by Π that we shall now define.

We write $\underset{\Pi}{\vdash\!\!-}$ for the regular closure of the set $(\Pi, 1_{A^*})$: that is, for f and g in A^*, we have *cf. Sec. 5.2, p. 297*

$$f \underset{\Pi}{\vdash\!\!-} g \quad \text{if and only if} \quad \exists u, v \in A^*\,,\ \exists a\,b \in \Pi \quad f = u\,a\,b\,v \quad \text{and} \quad g = u\,v \ .$$

Logically we write $\overset{\star}{\underset{\Pi}{\vdash\!\!-}}$ for the reflexive and transitive closure of $\underset{\Pi}{\vdash\!\!-}$; that is, $f \overset{\star}{\underset{\Pi}{\vdash\!\!-}} g$ if there exists a *chain of derivations* from f to g (that is, a sequence f_0, f_1, …, f_n of words in A^* such that $f = f_0$, $g = f_n$, and, for all i, $0 \leqslant i < n$, $f_i \underset{\Pi}{\vdash\!\!-} f_{i+1}$). Since each derivation reduces the length by 2, such a chain is necessarily finite. The following properties are obvious: *cf. Prop. 0.1.2, p. 13*

Properties 6.5

$$\forall f, g \in A^* \qquad f \overset{\star}{\underset{\Pi}{\vdash\!\!-}} g \quad \Longrightarrow \quad \begin{cases} \text{(i)} & f \equiv g \bmod [\Pi] \\ \text{(ii)} & |f| \geqslant |g| \\ \text{(iii)} & |f| = |g| \quad \Rightarrow \quad f = g \end{cases} \qquad \blacksquare$$

Property 6.6 *cf. Propty 5.1, p. 297*

$$\forall f, g, h, k \in A^* \qquad f \overset{\star}{\underset{\Pi}{\vdash\!\!-}} g \quad \textit{and} \quad h \overset{\star}{\underset{\Pi}{\vdash\!\!-}} k \quad \Longrightarrow \quad f\,h \overset{\star}{\underset{\Pi}{\vdash\!\!-}} g\,k \ . \qquad \blacksquare$$

Property 6.7 $\qquad \forall f \in U_\Pi^*\,,\ \forall a\,b \in \Pi \qquad a\,f\,b \in U_\Pi^* \ .$ $\blacksquare$

The definition of the congruence generated by a set of pairs of words implies that $f \equiv g \bmod [\Pi]$ if and only if there exists a *chain of derivations and inverse derivations* from f to g; that is, a sequence f_0, f_1, …, f_n of words in A^* such that $f = f_0$, $g = f_n$, and, for all i, $0 \leqslant i < n$, $f_i \underset{\Pi}{\vdash\!\!-} f_{i+1}$ or $f_{i+1} \underset{\Pi}{\vdash\!\!-} f_i$. If we write $f \underset{\Pi}{\dashv\!\!-} g$ for $g \underset{\Pi}{\vdash\!\!-} f$ and $\underset{\Pi}{\vdash\!\!\dashv}$ for $(\underset{\Pi}{\vdash\!\!-} \vee \underset{\Pi}{\dashv\!\!-})$, we obtain:

Property 6.8 *For all f, g in A^*,* $f \equiv g \bmod [\Pi] \iff f \overset{\star}{\underset{\Pi}{\vdash\!\dashv}} g$. ■

We write θ_Π for the relation from A^* to itself which associates with each word f the set of words that can be derived from f:

$$\forall f \in A^* \qquad f\theta_\Pi = \{g \in A^* \mid f \overset{\star}{\underset{\Pi}{\vdash}} g\} \ .$$

A word in A^* is called *reduced* (for Π) if it contains no factor in Π and we write S_Π for the set of reduced words:

$$S_\Pi = A^* \setminus A^* \Pi A^* \ .$$

The language S_Π is thus a rational of A^*, closed under factors. Every word that is not reduced can be reduced and we thus have

$$\forall f \in A^* \qquad f\theta_\Pi \cap S_\Pi \neq \emptyset \qquad \text{and} \qquad S_\Pi = \left\{h \in A^* \mid h\theta_\Pi = \{h\}\right\} \ .$$

In the case where $\Pi = \Delta_A$, we write $\delta_A = \theta_{\Delta_A}$ and $K_A = S_{\Delta_A}$:

$$\forall f \in \widetilde{A}^* \qquad f\delta_A = \{g \in \widetilde{A}^* \mid f \overset{\star}{\underset{\Delta_A}{\vdash}} g\}$$

and

$$K_A = \widetilde{A}^* \setminus \widetilde{A}^*\{z\bar{z} \mid z \in \widetilde{A}\}\widetilde{A}^* = \left\{h \in \widetilde{A}^* \mid h\delta_A = \{h\}\right\} \ .$$

In the case where $\Pi = \Sigma_A$, we write $\sigma_A = \theta_{\Sigma_A}$ and $J_A = S_{\Sigma_A}$:

$$\forall f \in \widetilde{A}^* \qquad f\sigma_A = \{g \in \widetilde{A}^* \mid f \overset{\star}{\underset{\Sigma_A}{\vdash}} g\} \ ,$$

and

$$J_A = \widetilde{A}^* \setminus \widetilde{A}^*\{a\bar{a} \mid a \in A\}\widetilde{A}^* = \left\{h \in \widetilde{A}^* \mid h\sigma_A = \{h\}\right\} \ .$$

We shall repeat[55] below the main results proved for generalised simplifications on the free group and involutive monoid using the above notation. After all these definitions and notations we can now move on to the properties of simplifications.

6.2.4 Reduction associated with a simplification

The definition of simplifications translates first into a *confluence* property of the derivation $\underset{\Pi}{\vdash}$:

Lemma 6.5 *Let f be in A^*. If there exist g and h such that $f \underset{\Pi}{\vdash} g$ and $f \underset{\Pi}{\vdash} h$, then there exists k such that $g \underset{\Pi}{\vdash} k$ and $h \underset{\Pi}{\vdash} k$.*

In other words, the existence of the upper part of the diagram in Figure 6.1 implies the existence of the lower part of the same diagram.

[55] Risking boredom, but allowing the reader in a rush to follow more easily.

Proof. The hypotheses imply that there exist u, v, x and y in A^*, ab and cd in Π such that

$$f = uabv = xcdy \ , \qquad g = uv \qquad \text{and} \qquad h = xy \ .$$

If $|u| = |x|$, $g = h = k$. Exchanging g and h if necessary, we can assume that $|u| > |x|$.

There are two cases:

(a) $|u| \geq |x| + 2$. Then,

$$f = xcdwabv \ , \qquad g = xcdwv \underset{\Pi}{\longmapsto} xwv \qquad \text{and} \qquad h = xwabv \underset{\Pi}{\longmapsto} xwv \ .$$

(b) $|u| = |x| + 1$. Then, $u = xc$, $y = bv$, $d = a$ and $b = c$ since Π is a simplification. Thus, $g = xcv = xbv = h = k$. ■

$$\begin{tikzcd}
 & f & \\
g \arrow[ur, mapsto, "\Pi"] & & h \arrow[ul, mapsto, "\Pi"'] \\
 & k \arrow[ul, mapsto, "\Pi"] \arrow[ur, mapsto, "\Pi"'] &
\end{tikzcd}$$

Fig. 6.1: Confluence of Π

Hence we deduce:

Lemma 6.6 *Two distinct words in S_Π are not congruent modulo Π.*

Proof. Suppose that x and y in S_Π are congruent modulo Π: there exists (Property 6.6) at least one sequence $x_0, x_1, \ldots, x_n$ of words in A^* such that $x = x_0$, $y = x_n$, and, for all i, $0 \leqslant i < n$, $x_i \underset{\Pi}{\longmapsto} x_{i+1}$.

A *chain* from x to y is such a sequence of words; the *height* of a chain is the maximum length of the words which constitutes it. We will *order* all the chains from x to y by height and, among the chains of the same height, by the number of words of maximum height. We choose a chain $C = \{x_0, x_1, \ldots, x_n\}$ from x to y that is minimal for this order.

If $x \neq y$, then C is of non-zero length. Since x and y are in S_Π, we have $x = x_0 \underset{\Pi}{\longmapsto} x_1$ and $x_{n-1} \underset{\Pi}{\longmapsto} x_n = y$. A word of maximum length in C, x_i, is necessarily a 'peak': that is, C can be written

$$x \overset{\star}{\underset{\Pi}{\longmapsto}} x_{i-1} \underset{\Pi}{\longmapsto} x_i \underset{\Pi}{\longmapsto} x_{i+1} \overset{\star}{\underset{\Pi}{\longmapsto}} y \ .$$

By Lemma 6.5, there exists z such that $x_{i-1} \underset{\Pi}{\longmapsto} z$ and $x_{i+1} \underset{\Pi}{\longmapsto} z$. The chain C'

$$x \overset{\star}{\underset{\Pi}{\longmapsto}} x_{i-1} \underset{\Pi}{\longmapsto} z \underset{\Pi}{\longmapsto} x_{i+1} \overset{\star}{\underset{\Pi}{\longmapsto}} y$$

goes from x to y and is smaller than C: if x_i is the unique word of maximum length, the height of C' is less than that of C, and if there exist other words of the same length as x_i, the number of words of maximum length in C' is smaller. Contradiction: the length of C is zero and $x = y$. ■

Proposition 6.7 *Every word f of A^* is congruent modulo $[\Pi]$ to a unique reduced word, written $f\eta_\Pi$, which can be obtained by derivations from f:*

$$f\eta_\Pi = f\theta_\Pi \cap S_\Pi \ .$$

Proof. We have already said $f\theta_\Pi \cap S_\Pi$ is non-empty for any f in A^*. Furthermore, if g and h belong to $f\theta_\Pi$ – i.e., if $f \overset{\star}{\underset{\Pi}{\vdash\!\!\!-}} g$ and $f \overset{\star}{\underset{\Pi}{\vdash\!\!\!-}} h$ – then $g \equiv f \equiv h \bmod [\Pi]$ (Property 6.5 (i)).

Thus, by Lemma 6.6, $f\theta_\Pi \cap S_\Pi$ reduces to a singleton, $f\eta_\Pi$, and f cannot be congruent modulo $[\Pi]$ to another element of S_Π. ■

The map $\eta_\Pi \colon A^* \to A^*$ is called the *reduction associated with* Π. We have $\alpha_\Pi = \eta_\Pi \circ \alpha_\Pi$ and $\mathrm{x}\, S_\Pi = (A^*)\eta_\Pi$ is in bijection with T_Π: we write ι_Π for the injection from T_Π to A^* such that $\eta_\Pi = \alpha_\Pi \circ \iota_\Pi$: the adjacent diagram commutes.

$$\begin{tikzcd} A^* \arrow[rr, "\alpha_\Pi"] \arrow[dr, "\eta_\Pi"'] & & T_\Pi \arrow[dl, shift left, "\iota_\Pi"] \\ & A^* \arrow[ur, shift left, "\alpha_\Pi"] & \end{tikzcd}$$

Fig. 6.2: Reduction η_Π

The following properties follow directly from Proposition 6.7:

Corollary 6.8 $\forall f, g \in A^* \qquad f \equiv g \bmod [\Pi] \iff f\eta_\Pi = g\eta_\Pi$. ■

Corollary 6.9
$\forall f, g \in A^* \qquad (f\,g)\eta_\Pi = ((f)\eta_\Pi\, g)\eta_\Pi = (f\,(g)\eta_\Pi)\eta_\Pi = ((f)\eta_\Pi\,(g)\eta_\Pi)\eta_\Pi$. ■

Corollary 6.10 $U_\Pi^* = \{f \in A^* \mid f \overset{\star}{\underset{\Pi}{\vdash\!\!\!-}} 1_{A^*}\}$. ■

We can restate Proposition 6.7 for the Dyck and Shamir congruences, and redraw the corresponding diagrams:

Proposition 6.7 (ii) *Every word f of $\widetilde{A}^*$ is congruent modulo $[\Delta_A]$ to a unique reduced word, written $f\rho_A$, and $f\rho_A = f\delta_A \cap K_A$. The map $\rho_A \colon \widetilde{A}^* \to \widetilde{A}^*$ is called the* Dyck reduction.
(iii) *Every word f in $\widetilde{A}^*$ is congruent modulo $[\Sigma_A]$ with a unique reduced word, written $f\tau_A$, and $f\tau_A = f\sigma_A \cap J_A$. The application $\tau_A \colon \widetilde{A}^* \to \widetilde{A}^*$ is called the* Shamir reduction.

$$\begin{tikzcd} \widetilde{A}^* \arrow[rr, "\gamma_A"] \arrow[dr, "\rho_A"'] & & F(A) \arrow[dl, shift left, "\iota"] \\ & \widetilde{A}^* \arrow[ur, shift left, "\gamma_A"] & \end{tikzcd} \qquad \begin{tikzcd} \widetilde{A}^* \arrow[rr, "\beta_A"] \arrow[dr, "\tau_A"'] & & I(A) \arrow[dl, shift left, "\iota"] \\ & \widetilde{A}^* \arrow[ur, shift left, "\beta_A"] & \end{tikzcd}$$

Figure 6.3: Dyck and Shamir reductions

The relation θ_Π is characterised by the lemma:

Lemma 6.11 *Let f and g be in A^* with $g = a_1\, a_2 \cdots a_n$ (where the a_i are in A). Then g is in $f\theta_\Pi$ if and only if there exist words w_0, w_1, …, w_n in U_Π^* such that $f = w_0\, a_1\, w_1\, a_2\, w_2 \cdots w_{n-1}\, a_n\, w_n$.*

Proof. The condition is necessary, by induction on the length of the chain of derivations which leads from f to g. The property is true for a chain of length 0; that is, if $g = f$, taking all the w_i equal to 1_{A^*}.

Suppose the lemma holds for $f \overset{\star}{\underset{\Pi}{\longmapsto}} g$ and let h be such that $g \underset{\Pi}{\longmapsto} h$. Then, by definition of $\underset{\Pi}{\longmapsto}$, there exists an i, $1 \leqslant i < n$, such that $h = a_1 a_2 \cdots a_{i-1} a_{i+2} \cdots a_n$ and $a_i a_{i+1} \in \Pi$. As $w' = w_{i-1} a_i w_i a_{i+1} w_{i+1}$ is in U_Π^* (Properties 6.6 and 6.7), the lemma holds for $f \overset{\star}{\underset{\Pi}{\longmapsto}} h$ with the sequence of words $w_0, w_1, \ldots, w_{i-2}, w', w_{i+2}, \ldots, w_n$.

The condition is sufficient by Corollary 6.10 and Property 6.6. ■

For the relations δ_A and σ_A, Lemma 6.11 is stated:

Lemma 6.11 *Let f and g be in $\widetilde{A}^*$ with $g = a_1 a_2 \cdots a_n$.*
(ii) *The word g is in $f\delta_A$ if and only if there exist words $w_0, w_1, \ldots, w_n$ in D_A^* such that $f = w_0 a_1 w_1 a_2 w_2 \cdots w_{n-1} a_n w_n$.*
(iii) *The word g is in $f\sigma_A$ if and only if there exist words $w_0, w_1, \ldots, w_n$ in P_A^* such that $f = w_0 a_1 w_1 a_2 w_2 \cdots w_{n-1} a_n w_n$.*

Remark 6.1 For given f and g, $g \in f\theta_\Pi$, the above factorisation of f is not necessarily unique, as the two following examples show, taken over $\widetilde{A} = \{a, \overline{a}\}$ (the factorisations of f are shown with dots).

(i) $\Pi = \Delta_A = \{a\overline{a}, \overline{a}a\}$.

$$f = a\,\overline{a}\,a \ , \quad f \underset{\Delta_A}{\longmapsto} g = a \qquad \text{and} \qquad f = 1 \cdot a \cdot \overline{a}\,a = a\,\overline{a} \cdot a \cdot 1 \ .$$

(ii) $\Pi = \Sigma_A = \{a\overline{a}\}$.

$$f = a\,\overline{a}\,a\,\overline{a} \ , \quad f \underset{\Sigma_A}{\longmapsto} g = a\,\overline{a} \qquad \text{and} \qquad f = 1 \cdot a \cdot 1 \cdot \overline{a} \cdot a\,\overline{a} = a\,\overline{a} \cdot a \cdot 1 \cdot \overline{a} \cdot 1 \ . \ \square$$

6.2.5 Unambiguous factorisation induced by a reduction

The result of this paragraph is partly a solution to the problem raised by the preceding remark and makes the statement of Lemma 6.11 more precise, ensuring not just the mere existence of a factorisation, but its uniqueness. It will only be used in the following chapters.[56] The case of the Shamir congruence is notably simpler than the general case, and makes it worth treating separately (and first).

Proposition 6.12 *Let f be in $\widetilde{A}^*$ and $f\tau_A = z_1 z_2 \cdots z_n$ (where the z_i are in $\widetilde{A}$). There exists a* unique *$(n+1)$-tuple of words $p_0, p_1, \ldots, p_n$ in P_A^* such that*

$$f = p_0 z_1 p_1 z_2 p_2 \cdots p_{n-1} z_n p_n \ . \tag{6.2}$$

Proof. The existence of the $(n+1)$-tuple is proved by Lemma 6.11. We show its uniqueness by induction on n. If $n = 0$, $f\tau_A = 1_{\widetilde{A}^*}$ and $f \in P_A^*$. Suppose that there exist two $(n+1)$-tuples of words in P_A^* which satisfy (6.2):

$$f = q_0 z_1 q_1 z_2 q_2 \cdots q_{n-1} z_n q_n = r_0 z_1 r_1 z_2 r_2 \cdots r_{n-1} z_n r_n \ .$$

[56] *Note added in translation:* Wishful writing that was erroneously left in the French edition, as these applications have not been included; they will appear in the second edition.

If $|q_n| = |r_n|$, let $q_n = r_n$, then the induction hypothesis, applied to

$$g = q_0\, z_1\, q_1\, z_2\, q_2 \cdots q_{n-2}\, z_{n-1}\, q_{n-1} = r_0\, z_1\, r_1\, z_2\, r_2 \cdots r_{n-2}\, z_{n-1}\, r_{n-1} \quad ,$$

gives the required equality. If $|q_n| \neq |r_n|$, we can assume, up to exchanging the q_i and ri, that $|q_n| < |r_n|$ and we set

$$f = g\, z_n\, q_n = h\, z_n\, r_n = h\, z_n k\, z_n\, q_n \quad .$$

The fact that $(r_n)\tau_A = (q_n)\tau_A = 1_{\widetilde{A}^*}$ implies that $(k\, z_n)\tau_A = (k\tau_A\, z_n)\tau_A = 1_{\widetilde{A}^*}$ whence $k\tau_A = a$ and $z_n = \overline{a}$ for some a in A. We then compute

$$g\tau_A = z_1\, z_2 \cdots z_{n-1} = (h\, z_n\, k)\tau_A = (h\tau_A\, z_n\, k\tau_A)\tau_A = (z_1\, z_2 \cdots z_{n-1}\, \overline{a}\, a)\tau_A \quad ,$$

which contradicts the fact that $z_{n-1}\, \overline{a}$ is a factor of $f\tau_A$, hence $z_{n-1} \neq a$. ■

Example (ii) of Remark 6.1 shows that the factorisation of f is no longer unique if we start with g in $f\sigma_A$ in general (as in Lemma 6.11 (iii)) and not from $g = f\tau_A$.

The proposition corresponding to Proposition 6.12 in the general case needs extra notation.

Notation Let Π be a simplification on A^*. For each a in A, we write F_a for the set of words of U_Π which *do not start* with a letter of A that simplifies on the right with a:

$$F_a = U_\Pi \setminus \left(\bigcup_{b \in A,\, ab \in \Pi} b\, A^* \right) \tag{6.3}$$

Proposition 6.13 *Let f be in A^* and $f\eta_\Pi = a_1\, a_2 \cdots a_n$ (where the a_i are in A). There exists a* unique $(n+1)$*-tuple of words u_0, u_1, …, u_n in U_Π^* such that*

$$f = u_0\, a_1\, u_1\, a_2\, u_2 \cdots u_{n-1}\, a_n\, u_n \qquad \textit{and} \qquad \forall i\,,\ 1 \leqslant i \leqslant n \quad u_i \in F_{a_i}^* \quad . \tag{6.4}$$

Remark 6.2 In the case where $\Pi = \Sigma_A$, we have $P_A \cap \overline{A}\, \widetilde{A}^* = \emptyset$, hence $F_z = P_A$ for all z in $\widetilde{A}$ and we are back at the statement of Proposition 6.12.

In the case where $\Pi = \Delta_A$, we obtain $F_z = D_A \setminus \overline{z}\, \widetilde{A}^*$. □

Proof of Proposition 6.13. Fix the simplification Π on A^*, and to leaven the notation set $V = U_\Pi^*$, and $\eta = \eta_\Pi$. To show the existence of a factorisation (6.4), we first prove a lemma.

Lemma 6.14 *Let w be in V and a in A. There exists t in V and u in F_a^* such that $a\, w = t\, a\, u$.*

Proof. Let u be the shortest word in V such that there exists a t in V such that $a\, w = t\, a\, u$ (such words exist since w itself satisfies the condition). If $u = 1_{A^*}$, then $u \in F_a^*$ and the property is proved.

If not, $|u| \geqslant 2$, $u = bu'$ and suppose *a contrario* that $u \notin F_a^*$, hence that $ab \in \Pi$. From $a = (aw)\eta = (tabu)\eta = (u)\eta$, we obtain $u = u'au''$, with u' and u'' in V. Let $aw = tabu'au''$ and $|u''| < |u|$: we have a contradiction. ■

Let $f = w_0 a_1 w_1 a_2 w_2 \cdots w_{n-1} a_n w_n$ be a factorisation given by Lemma 6.11; we shall deduce, by induction on n, a factorisation that satisfies (6.4). If $n = 0$, there is nothing to prove (Equation (6.4) expresses no condition on u_0). If not, Lemma 6.14 ensures that we can write $a_n w_n$ in the form $t a_n u$ and we apply the induction hypothesis to $f = w_0 a_1 w_1 a_2 w_2 \cdots w_{n-1} t$.

The uniqueness of the factorisation (6.4) rests on another lemma.

Lemma 6.15 *Let f in A^* have two factorisations $f = gas = hat$ where a is in A, and s and t are in F_a^*. If $f\eta = g\eta a = h\eta a$, then $s = t$.*

Proof. Suppose $s \neq t$ and, for example, $|s| > |t|$, so that $s = vat$. We obtain

$$f\eta = g\eta a = (g\eta(av)\eta a)\eta \ .$$

There are two cases to consider, according to whether $(av)\eta$ is equal to 1_{A^*} or not.

(i) If $(av)\eta = 1_{A^*}$ then $v\eta = b$ with $ab \in \Pi$.

Lemma 6.11 ensures that $v = v'bv''$ where v' and v'' are in V. From $s = v'bv''at \in F_a^*$ we obtain $bv''a \in V$ then $bv''a \in F_a^*$, which contradicts $ab \in \Pi$.

(ii) Suppose therefore that $(av)\eta \neq 1_{A^*}$. The hypothesis $f\eta = g\eta a = h\eta a$ implies $g\eta = h\eta$, hence $g\eta = (g\eta(av)\eta)\eta$, and then $g\eta = kc$ and $(av)\eta = dlc$ with c and d in A, $cd \in \Pi$ and $ca \notin \Pi$.

We then have $av = wdr$ with $w \in V$, and even $w \neq 1_{A^*}$ as otherwise $a = d$, which contradicts $cd \in \Pi$. Hence $v = udr$ where $u\eta = b$ and $ab \in \Pi$.

From $va = udra \in V$, we deduce $(udra)\eta = (bdlca)\eta = 1_{A^*}$, which dictates, since dlc is reduced, that $l = 1_{A^*}$, which is reasonable, and that bd and ca be in Π, which is impossible. ■

Two distinct factorisations (6.4)

$$f = u_0 a_1 u_1 a_2 u_2 \cdots u_{n-1} a_n u_n \quad \text{and} \quad f = v_0 a_1 v_1 a_2 v_2 \cdots v_{n-1} a_n v_n$$

lead to the situation of Lemma 6.15 which asserts $u_n = v_n$, and this completes the proof of Proposition 6.13 by induction on n. ■

Exercises

• 6.6 **Enumeration of Dyck words.** We first define the *height* and the *width* of a Dyck word.

Definition 6.1 For every word w in D_A^*, we define by induction two integers: $h(w)$, the *height* of w, and $p(w)$, the *width* of w as

$$h(1_{\widetilde{A}^*}) = 0$$
$$h(w) = \begin{cases} 1 + h(w') & \text{if} \quad w \in D_A \quad \text{and} \quad w = z\,w'\,\overline{z}\ , \quad (z \in \widetilde{A}, w' \in D_A^*) \\ \mathsf{max}_{1 \leqslant i \leqslant n}\{h(w_i)\} & \text{if} \quad w \notin D_A \quad \text{and} \quad w = w_1 \cdots w_n\ , \quad (w_i \in D_A) \end{cases}$$

$$p(1_{\widetilde{A}^*}) = 0$$
$$p(w) = \begin{cases} 1 & \text{if} \quad w = z\overline{z}\ , \quad (z \in \widetilde{A}) \\ p(w') & \text{if} \quad w \in D_A \quad \text{and} \quad w = z\,w'\,\overline{z}\ , \quad (z \in \widetilde{A}, w' \in D_A^+) \\ \mathsf{max}_{1 \leqslant i \leqslant n}(\{p(w_i)\}, n) & \text{if} \quad w \notin D_A \quad \text{and} \quad w = w_1 \cdots w_n\ , \quad (w_i \in D_A) \end{cases}$$

(a) Compute $h(f)$ and $p(f)$ for $f = \overline{a}\,\overline{a}\,a\,\overline{a}\,a\,a$.

(b) Show that there is only a finite number of Dyck words of a given height and width. More precisely, prove the following lemma.

Lemma 6.16 *Let h and p be two positive integers and $N(h,p)$ the integer defined by*

$$N(h,p) = 2\,p\,\left(p^{<h}\right) = 2\,p\,\frac{p^h - 1}{p - 1}\ .$$

If a word w in D_A^* has a length greater than or equal to $N(h,p)$, then either $p(w)$ is greater than or equal to p, or $h(w)$ is greater than or equal to h.

(c) Show that the bound of Lemma 6.16 is tight, even for an alphabet with one letter.

• 6.7 **A combinatorial lemma.** Prove:

Lemma 6.17 *Let $f_1, f_2, \ldots, f_n$, be n words in $\widetilde{A}^*$ whose product is in D_A^* : thus, $(f_1\,f_2 \cdots f_n)\rho_A = 1_{\widetilde{A}^*}$. There exists an index l, $1 \leqslant l < n$, such that*

$$|(f_l\,f_{l+1})\rho_A| \leqslant \mathsf{max}_{i \in [n]} |(f_i)\rho_A|\ .$$

6.8 Show that $\mathrm{F}(A)$ is the free group generated by A; that is, for any group G, every map from A to G extends uniquely to a morphism from $\mathrm{F}(A)$ to G.

• 6.9 Show that every element f of $F(A)$ distinct from $1_{F(A)}$ is contained in a recognisable subset of $F(A)$ which does not contain $1_{F(A)}$ (in other words, *the free group is residually finite*).

6.3 Rationals of the free group

One of the aims of this section, and in particular, a justification of the above computations, is the characterisation of the rational subsets of $F(A)$, given by the following theorem (recall that ι is the map defined in Figure 6.3):

see p. 310

Theorem 6.1 [**Benois**] *A subset X of $F(A)$ is rational if and only if $X\iota$ is a rational subset of $\widetilde{A}^*$.*

However, as above, we treat, under the heading of free group, properties of the quotients of free monoids by simplifications which for simplicity we shall call simplification monoids.

6.3.1 Rationals of simplification monoids

Let Π be a simplification on A^*. We start again from Figure 6.2 which is reproduced in the adjacent figure and whose notations we shall use in what follows. And we restate:

$$\begin{tikzcd} A^* \arrow[rr, "\alpha_\Pi"] \arrow[dr, "\eta_\Pi"'] & & T_\Pi \arrow[dl, shift left, "\iota_\Pi"] \\ & A^* \arrow[ur, shift left, "\alpha_\Pi"] & \end{tikzcd}$$

Theorem 6.1 *A subset X of T_Π is rational if and only if $X\iota_\Pi$ is a rational subset of A^*.*

This result is itself the direct consequence of the following proposition:

Proposition 6.18 *The image under η_Π of a rational of A^* is rational:*

$$\forall R \subseteq A^* \qquad R \in \operatorname{Rat} A^* \quad \Longrightarrow \quad R\eta_\Pi \in \operatorname{Rat} A^* \ .$$

Proof of Theorem 6.1. Since α_Π is surjective, if X is rational in T_Π, there exists (Proposition 1.7) a rational R of A^* such that $R\alpha_\Pi = X$ and $X\iota_\Pi = R\eta_\Pi$. ■ *see p. 223*

The proof of Proposition 6.18 rests on a construction made on automata which can easily be described on their matrix representation.

Definition 6.2 Let (λ, μ, ν) be the matrix representation of an automaton $\mathcal{A}$ over A^*. We write $H_\mathcal{A}$ for the image under μ of the language U_Π^*: *cf. Def. 2.2, p. 240*

$$H_\mathcal{A} = \sum_{w \in U_\Pi^*} w\mu \ ,$$

and $\mathcal{A}\theta_\Pi$ for the automaton whose representation (λ, π, ζ) is given by

$$\forall a \in A \qquad a\pi = H_\mathcal{A} \cdot a\mu \qquad \text{and} \qquad \zeta = H_\mathcal{A} \cdot \nu \ . \qquad \square$$

We then obtain:

Proposition 6.19 *The image under θ_Π of a rational of A^* is rational. More precisely, we have, for every automaton $\mathcal{A}$ over A^**

$$L(\mathcal{A}\theta_\Pi) = (L(\mathcal{A}))\,\theta_\Pi \ .$$

Proof. Let (λ, μ, ν) be the matrix representation of an automaton $\mathcal{A}$ over A^*. By definition, $L(\mathcal{A}) = \{f \in A^* \mid \lambda \cdot f\mu \cdot \nu = 1\}$. From this we obtain

$$(L(\mathcal{A}))\,\theta_\Pi = \{g \in A^* \mid \exists f \qquad g \in f\eta_\Pi \quad \text{and} \quad \lambda \cdot f\mu \cdot \nu = 1\}$$

hence, using the characterisation of Lemma 6.11,

$$\begin{aligned}(L(\mathcal{A}))\,\theta_\Pi = \{a_1\, a_2 \cdots a_n \in A^* \mid \exists w_0, w_1, \ldots, w_n \in U_\Pi^* \\ \lambda \cdot w_0\mu \cdot a_1\mu \cdot w_1\mu \cdot a_2\mu \cdot w_2\mu \cdot \cdots \cdot w_{n-1}\mu \cdot a_n\mu \cdot w_n\mu \cdot \nu = 1\} \ .\end{aligned}$$

Since a sum of booleans is 1 if and only if one of them is 1, we can write

$$(L(\mathcal{A}))\,\theta_\Pi = \{a_1\,a_2\cdots a_n \in A^* \mid \Bigg[\sum_{w_0,w_1,\ldots,w_n\in U_\Pi^*} \lambda\cdot w_0\mu\cdot a_1\mu\cdot w_1\mu\cdot a_2\mu\cdot w_2\mu\cdot\ \cdots\ \cdot w_{n-1}\mu\cdot a_n\mu\cdot w_n\mu\cdot\nu\Bigg] = 1\}\ ,$$

or, using distributivity and the notation $H_\mathcal{A} = \sum_{w\in U_\Pi^*} w\mu$,

$$\begin{aligned}(L(\mathcal{A}))\,\theta_\Pi &= \{a_1\,a_2\cdots a_n \in A^* \mid \lambda\cdot H_\mathcal{A}\cdot a_1\mu\cdot H_\mathcal{A}\cdot a_2\mu\cdot H_\mathcal{A}\cdot\ \cdots H_\mathcal{A}\cdot a_n\mu\cdot H_\mathcal{A}\cdot\nu = 1\}\\ &= \{a_1\,a_2\cdots a_n \in A^* \mid \lambda\cdot a_1\pi\cdot a_2\pi\cdot\ \cdots a_n\pi\cdot\zeta = 1\}\ .\end{aligned}$$

Thus,

$$(L(\mathcal{A}))\,\theta_\Pi = \{g\in A^* \mid \lambda\cdot g\pi\cdot\zeta = 1\} = L(\mathcal{A}\theta_\Pi)\ . \qquad \blacksquare$$

Proof of Proposition 6.18. We have $R\eta_\Pi = R\theta_\Pi\cap S_\Pi$ for $R\subseteq A^*$. As S_Π is rational, so is $R\eta_\Pi$ when R is. ■

The above proof *is not directly constructive* inasmuch as it involves the matrix $H_\mathcal{A}$ which is obtained as an infinite sum. To make Proposition 6.19 and all the results we obtain from it effective, we must therefore show that:

Lemma 6.20 *The matrix $H_\mathcal{A}$ is effectively computable.*

Proof. We shall show that $H_\mathcal{A}$ is the limit of an increasing sequence,[57] hence eventually stationary, of boolean matrices of dimension Q, and that this rank can be attained effectively.

Let $H^{(i)}$ and $\mu^{(i)}$ be the sequence of matrices and representations defined recursively from $\mathcal{A}$ by

$$\begin{aligned} H^{(0)} &= I & a\mu^{(0)} &= a\mu\ , \quad \forall a\in A\\ H^{(i+1)} &= \Bigg[\sum_{ab\in\Pi}(a\,b)\mu^{(i)}\Bigg]^* & a\mu^{(i+1)} &= H^{(i+1)}\cdot a\mu^{(i)}\ , \quad \forall a\in A\end{aligned}$$

Since we have $I\leqslant H^{(i)}$ for each i, the sequence of $a\mu^{(i)}$ is increasing for each a in A and hence so is the sequence of $H^{(i)}$'s. Since there is only a finite number of boolean matrices of the given size, this sequence is necessarily eventually stationary. To show that this rank is effectively reached, it suffices to show that if $H^{(i+1)} = H^{(i)}$ then $H^{(i)} = H_\mathcal{A}$.

Let i be such that $H^{(i+1)} = H^{(i)}$ and suppose that there exists w in U_Π^* such that, for a certain pair (p,q), $w\mu_{p,q} = 1$ and ${H^{(i)}}_{p,q} = 0$; we choose such a w of

[57] Implicitly, $0<1$ and the matrices are partially ordered by the product ordering, coefficient by coefficient.

minimal length. If w is a product of (shorter) words in U_Π^*, we have a contradiction with the fact that $H^{(i)}$, which is the star of a matrix, is idempotent. We thus have $w = a\,w'\,b$ with $a\,b$ in Π and w' in U_Π^*, shorter than w, and hence $w'\mu \leqslant H^{(i)}$. Hence we obtain

$$w\mu = a\mu \cdot w'\mu \cdot b\mu \leqslant a\mu^{(i)} \cdot H^{(i)} \cdot b\mu^{(i)} \leqslant a\mu^{(i+1)} \cdot b\mu^{(i+1)} \leqslant H^{(i+1)} \quad .$$

Impossible, since there exists an entry which is 1 in $w\mu$ and 0 in $H^{(i+1)}$. ■

Remark 6.3 The length of an ascending sequence of boolean matrices of dimension n is bounded by n^2. The above proof gives an algorithm *a priori* in $O(n^5)$. A more careful analysis allows us to design an algorithm which computes $H_{\mathcal{A}}$ and whose complexity is $O(n^3)$. □ *cf. Exer. 6.12*

As above, we restate the result for the free group and the involutive monoid.

Proposition 6.18 *Let R be a rational of $\widetilde{A}^*$.*
(ii) *$(R)\rho_A$ is rational.* (iii) *$(R)\tau_A$ is rational.*

We then deduce directly from Theorem 6.1 the main properties of the family of rationals in a simplification monoid.

Corollary 6.21 $\operatorname{Rat} T_\Pi$ *(resp.* $\operatorname{Rat} F(A)$, $\operatorname{Rat} I(A)$*) is an effective Boolean algebra.*

Proof. If X and Y are two subsets of T_Π, then

$$X \cup Y = (X\iota_\Pi \cup Y\iota_\Pi)\,\alpha_\Pi \qquad \text{and} \qquad \complement_{T_\Pi} X = (S_\Pi \setminus X\iota_\Pi)\,\alpha_\Pi \quad .$$

Hence the conclusion, since $\operatorname{Rat} A^*$ is an effective Boolean algebra. ■

In particular, we shall reduce several decision problems to the proposition that we can decide whether a given element of $F(A)$ (in general the identity) belongs to a given rational of $F(A)$.

Corollary 6.22 *Every rational subset of T_Π (resp. $F(A)$, $I(A)$) is unambiguous.*

Proof. If X, a subset of T_Π, is rational, then $X\iota_\Pi$ is a rational of A^*, hence unambiguous, and $X = X\iota_\Pi\alpha_\Pi$ is unambiguous since α_Π is injective on S_Π and hence on $X\iota_\Pi$. ■ *cf. Prop. 1.12, p. 231*

6.3.2 Return to the free group

Considering that A^* is a subset of $F(A)$, via γ_A which is injective on A^*, we can also see Theorem 6.1 as a Fatou property of the free group:

Corollary 6.23 *If X is a rational subset of $F(A)$ included in A^*, then X is a rational subset rational A^*.* ■

We can also deduce from Theorem 6.1 the classic result on finitely generated subgroups of the free group.

Theorem 6.2 [**Howson**] *The intersection of two finitely generated subgroups of the free group is a finitely generated subgroup.*[58]

Proof. A finitely generated subgroup is a rational. The intersection of two such subgroups is a subgroup, and is a rational by Corollary 6.21. It is a finitely generated subgroup by Proposition 6.2. ■

Remark 6.4 There exist groups G that do not satisfy[59] the above property; that is, such that there exist two finitely generated subgroups H and K whose intersection is a subgroup which is not finitely generated. In this case, H and K are necessarily both of infinite index. Indeed, if H is of finite index (hence recognisable) in G finitely generated, its intersection with K finitely generated (hence rational) is rational, hence finitely generated. □

cf. Prop. 6.2, p. 302

The rational subsets of $F(A)$ are not, in general, recognisable; that is, the classic statement of Kleene's Theorem does not hold for the free group. Nonetheless we have the following result which can be seen as a generalisation to rational subsets of a classic result,[60] modulo Proposition 6.2, for the subgroups of $F(A)$. (Recall that a subset of a monoid is *disjunctive* if it is not saturated by any non-trivial congruence.)

cf. Exer. 2.24, p. 251

Theorem 6.3 [**Sénizergues**] *A rational subset of $F(A)$ is either disjunctive, or recognisable.*

As the proof is effective, we deduce:

Corollary 6.24 *It is decidable whether a rational of $F(A)$ is recognisable.* ■

Exercises

- 6.10 Deduce Proposition 6.18 from Theorem 6.1.
- 6.11 Let R be a rational of A^* and $\varphi\colon A^* \to F(B)$ a morphism. Show that we can decide if the intersection $R \cap (1_{F(B)})\varphi^{-1}$ is empty or not.
- 6.12 Convert the proof of Lemma 6.20 into an algorithm that computes the matrix $H_{\mathcal{A}}$ and whose complexity is $O(\|Q\|^3)$.

6.13 Supposing that Theorem 6.3 is true for rational subgroups (*cf.* Note 60), show that it also holds for rational subsets which are the finite union of cosets for a subgroup.

- 6.14 Give a simple proof (one which does not use Theorem 6.3) of the fact that it is decidable whether a finitely generated subgroup of the free group is of finite index.

---

[58] *cf.* A. Howson, [124].

[59] Lyndon and Schupp [161, p. 18] give an example due to Karrass and Solitar. The remark answers a question which they ask on this subject.

[60] The result on subgroups of $F(A)$ is due to Greenberg, [107] (*cf.* also [161, Proposition 3.11]). I conjectured Theorem 6.3 in [208]. Sénizergues's proof [230] depends crucially on Theorem 5.4. Silva recently suggested a proof [234] which is more directly related to that of Greenberg's Theorem.

6.4 Büchi systems

J. Büchi defined particular one-sided rewriting systems,[61] which we shall describe using Shamir reduction.

Let A be an alphabet. Like a Thue system, a *Büchi system* is a *finite* subset K of $A^* \times A^*$. The change of name is justified by the relation from A^* to itself which we shall define from K.

We write $\underset{\mathcal{B}\ K}{\vdash\!\!\!-\!\!\!-}$ for the *regular left closure* of K, considered as a relation on A^*:

$$\forall f, g \in A^* \qquad f \underset{\mathcal{B}\ K}{\vdash\!\!\!-\!\!\!-} g \quad \text{if and only if}$$
$$\exists h \in A^*\,,\ \exists (u,v) \in K \quad f = hu \quad \text{and} \quad g = hv\ .$$

Naturally we write $\overset{\star}{\underset{\mathcal{B}\ K}{\vdash\!\!\!-\!\!\!-}}$ for the transitive and reflexive closure of $\underset{\mathcal{B}\ K}{\vdash\!\!\!-\!\!\!-}$.

Proposition 6.25 *With every Büchi system K over A^*, we can associate a rational set R_K of $\widetilde{A}^*$, contained in $\overline{A}^* A^*$, such that*

$$\forall f, g \in A^* \qquad f \overset{\star}{\underset{\mathcal{B}\ K}{\vdash\!\!\!-\!\!\!-}} g \qquad \textit{if and only if} \qquad g \in (f R_K)\, \tau_A \cap A^*\ .$$

We shall start by proving some properties of the Shamir reduction.

Lemma 6.26 *For all u and v in $\widetilde{A}^*$ we have*

$$((u\tau_A \cap A^*)\, v)\, \tau_A \cap A^* = (uv)\tau_A \cap A^*\ .$$

Proof. Set $g = u\tau_A$. If $g \in A^*$, $u\tau_A \cap A^* = u\tau_A$ and $((u\tau_A \cap A^*)v)\tau_A = ((u\tau_A)v)\tau_A = (uv)\tau_A$ then

$$((u\tau_A \cap A^*)v)\tau_A \cap A^* = (uv)\tau_A \cap A^*\ .$$

If $g \notin A^*$, then $u\tau_A \cap A^* = \emptyset$ and $((u\tau_A \cap A^*)v)\tau_A \cap A^* = \emptyset$. Since g contains at least one barred letter, we can write $g = k\bar{a}h$, with $a \in A$ and $h \in A^*$. If

$$gv \underset{\Sigma_A}{\vdash\!\!\!-\!\!\!-} w \qquad \text{then} \qquad w = k\bar{a}w' \qquad \text{with} \qquad hv \underset{\Sigma_A}{\vdash\!\!\!-\!\!\!-} w'\ ,$$

since a reduction consists of deleting a factor of the form $c\bar{c}$, $c \in A$ and there is no such factor in g since g is reduced. By induction on the length of the derivation, it follows that if

$$g\,v \overset{\star}{\underset{\Sigma_A}{\vdash\!\!\!-\!\!\!-}} w \qquad \text{then} \qquad w = k\,\bar{a}\,w' \qquad \text{where} \qquad h\,v \overset{\star}{\underset{\Sigma_A}{\vdash\!\!\!-\!\!\!-}} w'\ ,$$

and hence if $w \in (g\,v)\sigma_A$, then $w \notin A^*$ and

$$(u\,v)\tau_A \cap A^* = (g\,v)\tau_A \cap A^* \subset (g\,v)\sigma_A \cap A^* = \emptyset\ ,$$

which proves the lemma. ∎

[61] Calling them *regular canonical systems* [45].

Lemma 6.27 *For all f and h in A^*, h is a suffix of f if and only if $(f\overline{h})\tau_A \in A^*$.*

cf. Propty 6.4, p. 306

Proof. From $(h\overline{h})\tau_A = 1_{A^*}$, we deduce $(g\,h\overline{h})\tau_A = g$ since $g \in A^* \subset J_A$.

Conversely, suppose that $(f\overline{h})\tau_A \in A^*$; we show, by induction on the length of h, that h is a suffix of f. The property is trivial for $|h| = 0$. Set $h = a\,h'$. If $(f\overline{a\,h'})\tau_A = (f\overline{h'}\,\overline{a})\tau_A$ is in A^* then

$$(f\overline{h'}\,\overline{a})\tau_A \cap A^* = (((f\overline{h'})\tau_A \cap A^*)\overline{a})\tau_A \cap A^*$$

is non-empty, hence $(f\overline{h'})\tau_A \cap A^*$ is non-empty, hence $(f\overline{h'})\tau_A \in A^*$, or, by the induction hypothesis, $f = g\,h'$. It thus follows that $(f\overline{h'}\,\overline{a})\tau_A = ((f\overline{h'})\tau_A\overline{a})\tau_A = (g\,\overline{a})\tau_A$ is in A^*. For $g\overline{a}$ to be reducible by σ_A, the only possibility is that $g = g'\,a$ and hence $f = g'\,a\,h'$ has h as a suffix. ■

Proof of Proposition 6.25. Let f, u and v be in A^*; the above assertions allow us to write the series of implications

$$(f\overline{u}v)\tau_A \in A^* \;\Leftrightarrow\; (f\overline{u})\tau_A \in A^* \;\Leftrightarrow\; f = hu \;\Rightarrow\; (f\overline{u}v)\tau_A = hv \;\Rightarrow\; (f\overline{u}v)\tau_A \in A^* \;.$$

Set $T_K = \{\overline{u}\,v \in \bar{A}^*A^* \mid (u,v) \in K\}$. Thus, for all f in A^*,

$$\begin{aligned} f \vdash_{\mathcal{B}\,K} g \quad &\Longleftrightarrow \quad \exists (u,v) \in K \qquad (f\overline{u}v)\tau_A = g \\ &\Longleftrightarrow \quad g \in (fT_K)\tau_A \cap A^* \;. \end{aligned} \tag{6.5}$$

We then show, by induction on n,

$$f \overset{n}{\vdash_{\mathcal{B}\,K}} g \quad \Longleftrightarrow \quad g \in (f(T_K)^n)\tau_A \cap A^* \;.$$

The property is true for $n = 0$ and $n = 1$; we therefore have

$$\begin{aligned} f \overset{n+1}{\vdash_{\mathcal{B}\,K}} g \quad &\Longleftrightarrow \quad \exists h \in A^* \qquad f \overset{n}{\vdash_{\mathcal{B}\,K}} h \quad \text{and} \quad h \vdash_{\mathcal{B}\,K} g \\ &\Longleftrightarrow \quad g \in (((f(T_K)^n)\tau_A \cap A^*)T_K)\tau_A \cap A^* \\ &\qquad\qquad \text{by the induction hypothesis and (6.5)} \\ &\Longleftrightarrow \quad g \in (f(T_K)^{n+1})\tau_A \cap A^* \;. \end{aligned}$$

And hence

$$f \overset{\star}{\vdash_{\mathcal{B}\,K}} g \quad \Longleftrightarrow \quad g \in (fS_K)\,\tau_A \cap A^* \;,$$

with $S_K = (T_K)^*$ and $S_K \in \mathrm{Rat}\,\widetilde{A}^*$ since T_K is finite. It remains to show that

$$(fS_K)\,\tau_A \cap A^* = (fR_K)\,\tau_A \cap A^* \qquad \text{with} \quad R_K = (S_K)\,\tau_A \cap \overline{A}^*A^* \;.$$

First we have $(fS_K)\,\tau_A = (f(S_K)\tau_A)\,\tau_A$ (Corollary 6.9) and we know that $(S_K)\,\tau_A$ is a rational of $\widetilde{A}^*$ (Proposition 6.18).

Next we notice that the complement of $\overline{A}^*A^*$ is $\overline{A}^*A^+\overline{A}\widetilde{A}^*$; that is, that a word z in $(S_K)\,\tau_A \setminus \overline{A}^*A^*$ can be written $z = \overline{s}\,r\,b\,\overline{a}\,t$ with r and s in A^*, t in $\widetilde{A}^*$ and a and b in A, $a \neq b$ (otherwise z would not be in J_A).

The only way of reducing a word fz by Σ_A, with f in A^*, is for a factor $c\overline{c}$ to straddle the factorisation fz; that is, that $f = f'c$ and $z = \overline{c}\,\overline{s}'\,r\,b\,\overline{a}\,t$. Thus, whatever r, s and t are, $b\overline{a}$ is a factor of any reduction of fz and $(fz)\tau_A$ does not belong to $\overline{A}^*A^*$, which completes the proof of the property. ■

Remark 6.5 It is essential to use the Shamir reduction τ_A and not the Dyck reduction ρ_A to describe the relation defined by a Büchi system. In fact, the reduction of a word of the form $f\,\overline{u}\,v$ by Σ_A only involves factors inside the word $f\,\overline{u}$, while the reduction of the same word by Δ_A can involve factors inside $\overline{u}\,v$, which corresponds to a *transformation of the system* K and not a *derivation* in that system.

More precisely, let $K = \{(a\,a, a)\}$ and $f = b\,a$. Then

$$(b\,a\,\overline{a}\,\overline{a}\,a)\rho_A = b \in A^* \qquad \text{while} \qquad (b\,a\,\overline{a}\,\overline{a}\,a)\tau_A = b\,\overline{a}\,a \notin A^* \,. \qquad \square$$

Corollary 6.28 *The equivalence of Büchi systems is decidable.*

Proof. Write $\widehat{K}$ for the graph of the relation realised by the Büchi system K:

$$\widehat{K} = \{(f,g) \in A^* \times A^* \mid f \overset{\star}{\underset{\mathcal{B}\ K}{\vdash}} g\} \ .$$

This operation $\widehat{\ }$ is a *closure*:[62] clearly

$$K \subseteq \widehat{K} \qquad \text{and} \qquad K \subseteq J \implies \widehat{K} \subseteq \widehat{J}$$

and, by induction on the length of the derivations

$$f \overset{\star}{\underset{\mathcal{B}\ \widehat{K}}{\vdash}} g \implies f \overset{\star}{\underset{\mathcal{B}\ K}{\vdash}} g \qquad \text{thus} \qquad \widehat{\widehat{K}} = \widehat{K} \ .$$

From this we deduce that $K \subseteq \widehat{J} \Leftrightarrow \widehat{K} \subseteq \widehat{J}$. Thus, two systems K and J are equivalent if and only if $K \subseteq \widehat{J}$ and $J \subseteq \widehat{K}$, which we can test effectively since K and J are finite and each test $(u,v) \in \widehat{J}$ becomes $v \in (u\,R_J)\,\tau_A \cap A^*$, which is just the membership of a word in an effectively computable rational subset. ■

Let us give the name *Büchi grammar*[63] to a quadruple $\mathcal{G} = \langle\, A, N, K, U \,\rangle$ where A is the alphabet called *terminal*, N is the alphabet called *non-terminal* (we have $A \cap N = \emptyset$ and write $V = A \cup N$), K is a Büchi system (over V^*) and where U, the set of *axioms*, is a *finite* subset of V^*.

Theorem 6.4 *A language of A^* is rational if and only if it is generated by a Büchi grammar.*

---

[62] We could well have used the notation $\overline{K}$ if in the context of this result the 'bar' operation were not already taken.

[63] Büchi himself, *art. cit.*, calls them *finite automata*, which is logical from his point of view, and consistent thanks to Theorem 6.4.

Proof. (i) The language generated by a Büchi grammar $\mathcal{G} = \langle N, A, K, U \rangle$ is rational.

As in Proposition 6.25, but with the rewriting system Σ_V and hence relations σ_V and τ_V defined on $\widetilde{V}^*$, we set

$$T_K = \bigcup_{(u,v)\in K} \overline{u}\, v \qquad \text{and} \qquad R_K = (T_K^*)\, \tau_V \cap \overline{V}^* V^* \ .$$

It then follows that $L(\mathcal{G}) = ((U\, R_K)\, \tau_V \cap V^*) \cap A^* = (U\, R_K)\, \tau_V \cap A^*$ is rational since $U\, R_K$ is rational (since U is finite) and that the image under τ_V of a rational is rational.

(ii) Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be a finite automaton that recognises L. Over the alphabet $V = A \cup Q$, we consider the Büchi system K defined by

$$K = K_1 \cup K_2$$

with

$$K_1 = \{(p, aq) \mid (p, a, q) \in E\} \qquad \text{and} \qquad K_2 = \{(p, 1_{V^*}) \mid p \in T\} \ .$$

By induction on the length of g, we show that, for all p and q in Q,

$$p \xrightarrow[\mathcal{A}]{g} q \qquad \Longleftrightarrow \qquad \forall f \in A^* \qquad f\, p \overset{\star}{\underset{\mathcal{B}\ K_1}{\vdash\!\!\!-}} f\, g\, q \ .$$

The property is true for $|g| = 0$ (since $p \xrightarrow[\mathcal{A}]{1_{A^*}} p$ and $\overset{\star}{\underset{\mathcal{B}\ K_1}{\vdash\!\!\!-}}$ is reflexive) and we obtain

$$p \xrightarrow[\mathcal{A}]{ga} q \qquad \Longleftrightarrow \qquad \exists r \in Q \qquad p \xrightarrow[\mathcal{A}]{g} r \quad \text{and} \quad r \xrightarrow[\mathcal{A}]{a} q$$

whence

$$p \xrightarrow[\mathcal{A}]{ga} q \qquad \Longleftrightarrow \qquad \forall f \in A^* \qquad f\, p \overset{\star}{\underset{\mathcal{B}\ K_1}{\vdash\!\!\!-}} f\, g\, r \tag{6.6}$$

by induction, and by $f\, g\, r \underset{\mathcal{B}\ K_1}{\vdash\!\!\!-} f\, g\, a\, q$ since $(r, a\, q)$ is in K_1. It therefore follows, observing that the words of A^* cannot be reduced by K, that

$$i \xrightarrow[\mathcal{A}]{f} t \ , \quad t \in T \qquad \Longleftrightarrow \qquad i \overset{\star}{\underset{\mathcal{B}\ K_1}{\vdash\!\!\!-}} f\, t \underset{\mathcal{B}\ K_2}{\vdash\!\!\!-} f \qquad \Longleftrightarrow \qquad i \overset{\star}{\underset{\mathcal{B}\ K}{\vdash\!\!\!-}} f \ .$$

Setting $\mathcal{G} = \langle Q, A, K, I \rangle$, we obtain

$$\begin{aligned} L(\mathcal{G}) &= \{f \in A^* \mid \exists i \in I \quad i \overset{\star}{\underset{\mathcal{B}\ K}{\vdash\!\!\!-}} f\} \\ &= \{f \in A^* \mid \exists i \in I\, \exists t \in T \quad i \xrightarrow[\mathcal{A}]{f} t\} = L(\mathcal{A}) \ . \end{aligned}$$

■

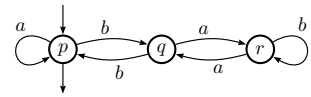
The theory of Büchi grammars continues with the definition of pure grammars, whose study is the subject of an exercise.

Exercises

• 6.15 **Pure Büchi grammars.** We say that a Büchi grammar $\mathcal{G} = \langle A, N, K, U \rangle$ is *pure* if $N = \emptyset$.

(a) Show that every rational language is generated by a pure Büchi grammar.

(b) Construct the pure Büchi grammar corresponding to the automaton $\mathcal{P}_2$ and give the series of derivations that produces the word $ababa$.



• 6.16 **'Lead or Delay' action revisited.** We return to the definition of the map χ_B: let B^* be an alphabet and $H = (B^* \times 1_{B^*}) \cup (1_{B^*} \times B^*) \cup \{\mathbf{0}\}$. We define a map ψ from $B^* \times B^*$ to H by *cf. Exer. 2.7, p. 237*

$$\forall f, g \in B^* \qquad (f,g)\psi = \begin{cases} (g^{-1}f, 1_{B^*}) & \text{if} \quad g \leqslant f \\ (1_{B^*}, f^{-1}g) & \text{if} \quad f \leqslant g \\ \mathbf{0} & \text{otherwise} \end{cases},$$

and a map χ_B from $H \times (B^* \times B^*)$ to H by:

$$\forall f, g, u, v \in B^* \qquad \begin{cases} ((f, 1_{B^*}), (u,v))\,\chi_B & = (f\,u, v)\psi \\ ((1_{B^*}, g), (u,v))\,\chi_B & = (u, g\,v)\psi \\ (\mathbf{0}, (u,v))\,\chi_B & = \mathbf{0} \end{cases}.$$

Express χ_B with the help of the Dyck reduction and verify that it is an action of $B^* \times B^*$ on H.

7 Rationals in commutative monoids

A monoid M is called *commutative* if its law is commutative: that is, if $x\,y = y\,x$ for all x and y in M. This assumption, or property, is most natural: in general, when one 'counts' – without specifying what or how – one does so in a commutative monoid. It is highly 'structuring': the notions seen up to here, such as rational subsets and congruences, will take particularly simple forms.

We will show that, as in the free monoid (and in the free group) but for reasons and with proofs that are fundamentally different, the rational subsets of a (finitely generated) commutative monoid form an effective Boolean algebra and are all unambiguous (Theorems 7.5 and 7.6). These two results are essentially reproduced from the magisterial presentation given by their authors, Eilenberg and Schützenberger.[64]

7.1 The natural order on $A^{\oplus}$

Following common usage, the operation of a commutative monoid is written additively[65] (and called addition) and its neutral element is written 0, or 0_M. We write $p+p = 2p$ and, more generally, $n\,p$, where n is an integer, for the sum $p+p+\cdots+p$ of n terms. Since the symbol for the monoid's operation is different, that for the generated monoid – that is, for star – is also, and $B^{\oplus}$ denotes the submonoid generated by the subset B of M.

[64] In [83].

[65] In the next chapter, when we deal with *series on a commutative monoid*, we shall be obliged to revert to multiplicative notation.

In a commutative monoid, the notions of left and right ideal and of two-sided ideal coincide, and we simply talk of an *ideal*. Every ideal I of M is of the form $I = K + M$ since in particular $I = I + M$.

7.1.1 The free commutative monoid

Let A be a finite set of k elements. The *free commutative monoid* generated by A is the quotient of A^* by the congruence generated by the relations $a\,b = b\,a$ for all a and b in A. It is written[66] $A^{\oplus}$. The monoid $A^{\oplus}$ is isomorphic to $\mathbb{N}^k$, whose elements are the k-tuples of integers $x = (x_1, x_2, \ldots, x_k)$. (It is also isomorphic to $a_1^* \times a_2^* \times \cdots \times a_k^*$, where the a_i are the letters of A, but this isomorphism is uninteresting when we use additive notation). The *length* of x in $A^{\oplus}$, hence in $\mathbb{N}^k$, written $|x|$, is the sum of the components:

$$|x| = \sum_1^k x_i \quad .$$

The *natural order* on $A^{\oplus}$ is the partial order defined by

$$\forall x, y \in A^{\oplus} \qquad x \leqslant y \;\Leftrightarrow\; \exists z \in A^{\oplus} \quad y = x + z \quad . \tag{7.1}$$

Viewed on $\mathbb{N}^k$, this order is the cartesian product of the orders on $\mathbb{N}$ and is written

$$\forall x, y \in \mathbb{N}^k \qquad x \leqslant y \;\Leftrightarrow\; \forall i \in [k] \quad x_i \leqslant y_i \quad . \tag{7.2}$$

We write $x \not\leqslant y$ if x is not less than or equal to y.

7.1.2 Dickson's Lemma

cf. Sec. 5.1, p. 293 Recall that if E is a partially ordered set, an *ideal* of E (for the order) is a non-empty subset I of E which is closed 'above', and that the ideal generated by a subset B of E is written $\mathsf{I}(B)$:

$$\mathsf{I}(B) = \{x \in E \mid \exists y \in B \quad y \leqslant x\} \quad .$$

It then follows from the definition (Equation (7.1)) that in $A^{\oplus}$ *the ideals for the natural order coincide with the ideals for the monoid structure*: in both cases, these are sets of the form $B + A^{\oplus}$.

Recall also that a *chain* is a subset (of a partially ordered set) all of whose elements are pairwise comparable and an *antichain* is a non-empty subset all of whose elements are pairwise *incomparable*.

By (7.2), we deduce that a strictly decreasing chain in $A^{\oplus}$, hence in $\mathbb{N}^k$, whose first element is x has at most $|x|$ elements, so *every strictly decreasing chain in* $A^{\oplus}$ *is finite*. It follows that, for every non-empty subset S of $A^{\oplus}$, the set of minimal elements of S, written $\min S$, is non-empty and is thus an antichain.

[66] By doing this, we make the same conflation as in the case of free monoids where the star denotes both the generated monoid and the generated *free* monoid. However, this never gets us into trouble.

Theorem 7.1 *The natural order on $A^{\oplus}$ is a partial well ordering.*

This result is simple but fundamental. It is equivalent, since there is no infinite strictly decreasing chain in $A^{\oplus}$, to the following result. *cf. Th. 5.1, p. 294*

Lemma 7.1 [**Dickson**] *Every antichain in $A^{\oplus}$ (for the natural order) is finite.*

Proof. We take advantage of the equivalence between conditions (iii) and (vi) of Theorem 5.1 and show that $A^{\oplus}$ has the finite basis property: that is, that for every subset S of $A^{\oplus}$ there exists a finite subset F of S such that S is included in $F + A^{\oplus}$.

We prove the lemma by induction on the cardinal k of A and reasoning about $\mathbb{N}^k$. If $k = 1$, the natural order is total and we take F to be *the* minimum element of S.

Let $y = (y_1, y_2, \ldots, y_k)$ be an arbitrary fixed element of S. An element z of $\mathbb{N}^k$ which is not in $y + \mathbb{N}^k$ has at least one of its coordinates z_i strictly less than the corresponding coordinate y_i of y.

For each i in $[k]$, and for each v, $0 \leqslant v < y_i$, the set $S_{i,v}$ of elements of S whose ith coordinate is equal to v is isomorphic (for the natural order) to a subset of $\mathbb{N}^{k-1}$. By the induction hypothesis, there exists a finite subset $F_{i,v}$ of $S_{i,v}$ such that $S_{i,v}$ is included in $F_{i,v} + \mathbb{N}^{k-1}$. There is only a finite number of pairs (i, v); the set $F = \{y\} \cup \left\{\bigcup_{i,v} F_{i,v}\right\}$ is finite and contained in S, and S is contained in $F + \mathbb{N}^k$. ■

We shall formalise this reasoning in the next subsection with the *lexicographic decomposition* of the subsets of $\mathbb{N}^k$, to apply it easily to the proof of other propositions. The ideals of $A^{\oplus}$ will often be dealt with using the following characterisation.

Corollary 7.2 *Let I be a subset of $A^{\oplus}$. The following conditions are equivalent:*

(i) *there exists a finite subset F such that $I = F + A^{\oplus}$;*

(ii) *there exists a finite subset F such that $\complement I = \{z \mid \forall y \in F \quad y \not\leqslant z\}$;*

(iii) *I is an ideal.*

Proof. (i) $\Rightarrow$ (ii) If $I = F + A^{\oplus}$, then:

$$\begin{aligned}\complement I &= \{z \mid \neg\exists y \in F\,,\ \neg\exists x \in A^{\oplus} \quad z = y + x\}\\ &= \{z \mid \forall y \in F\,,\ \forall x \in A^{\oplus} \quad z \neq y + x\} = \{z \mid \forall y \in F \quad y \not\leqslant z\}\ .\end{aligned}$$

(ii) $\Rightarrow$ (iii) We have a series of implications:

$$\begin{aligned} z \in I \quad \Longrightarrow \quad & \big(\ \forall y \in F\,,\ y \leqslant z \ \Rightarrow\ \forall m \in A^{\oplus} \quad y \leqslant z + m \ \big)\\ & \Longrightarrow \quad I + A^{\oplus} \subseteq I\ .\end{aligned}$$

Propositions (i) and (iii) are equivalent by Theorem 5.1. ■

Corollary 7.3 *The ideals of a finitely generated commutative monoid satisfy the ascending chain condition.*

Proof. Let $\alpha\colon A^{\oplus} \to M$ be a surjective morphism, with A finite. If I is an ideal of M, $I\alpha^{-1}$ is an ideal of $A^{\oplus}$, equal to $F + A^{\oplus}$ for some finite F, and $I = G + M$ with $G = (F)\alpha$, a finite subset of M.

Let $I_1 \subseteq I_2 \subseteq \cdots \subseteq I_n \subseteq \cdots$ be a chain of ideals of M. The union of all the I_n is an ideal I of M, equal to $G + M$ with G finite; G is included in an I_n, for some n, hence so is I, and $I = I_n$. ■

Exercises

• 7.1 We say that a finite antichain X of an ordered set E is *wide* if the cardinal of every antichain of E that contains X is bounded.

(a) Show that every antichain of $\mathbb{N}^2$ is wide.

(b) Give a necessary and sufficient condition for an antichain of $\mathbb{N}^k$, $k \geqslant 3$, to be wide.

• 7.2 We say that a language L over A^* is a *bounded rational language*[67] if it is a finite union of products of *rays*: that is, $L = \bigcup S_i$ and each of the S_i can be written

cf. also Exer. I.8.4, p. 177

$$S_i = u_0\, v_1^*\, u_1\, v_2^* \cdots u_n\, v_n^*\, u_{n+1} \quad ,$$

with the u_j and the v_j in A^*; by convention, all the v_j are different from 1_{A^*} and if we set $n = 0$ in the preceding expression, S_i is reduced to one word: every finite language is bounded rational. Prove:

Proposition 7.4 *Every rational language contains a bounded rational language with the same commutative image.*

7.2 The lexicographic order on $\mathbb{N}^k$

The natural order on $\mathbb{N}$ is not only a total order but a *well ordering*. The finite set $[k]$ is also naturally well ordered. In the usual way we deduce a well ordering on $\mathbb{N}^k$ (that is, the set of maps from $[k]$ to $\mathbb{N}$) called the *lexicographic order*: $x = (x_1, x_2, \ldots, x_k)$ is less than $y = (y_1, y_2, \ldots, y_k)$ if x_i is less than y_i, where i is the smallest index for which the coordinates of x and y differ. In other words

$$\forall x, y \in \mathbb{N}^k \qquad x \preccurlyeq y \quad \Longleftrightarrow \quad \big[\, i = \mathsf{min}\{j \in [k] \mid x_j \neq y_j\} \;\Rightarrow\; x_i < y_i \,\big] \quad .$$

Contrary to the natural order, the lexicographic order is not defined intrinsically on $A^{\oplus}$, but modulo the choice of an order on A; that is, of an isomorphism between $A^{\oplus}$ and $\mathbb{N}^k$. The lexicographic order satisfies the relation

$$\forall x, y, z \in \mathbb{N}^k \qquad x \preccurlyeq y \quad \Longleftrightarrow \quad x + z \preccurlyeq y + z \quad . \tag{7.3}$$

Note also that even though the lexicographic order is a well ordering – that is, there is no infinite strictly decreasing chain – the set of minorants of a given element is, in general, infinite.

[67] An unhappy, even detestable term, but one imposed by usage (principally that of the theory of *context-free* languages [101]).

7.2.1 Congruences of $\mathbb{N}^k$

The natural order and the lexicographic order will enable us to give a first description of the congruences of $\mathbb{N}^k$:

Theorem 7.2 *Let η be a congruence of $\mathbb{N}^k$. There exists a finite subset of $\mathbb{N}^k$, $F(\eta)$, such that:*

(i) *the set $S(\eta) = \{x \mid \forall y \in F(\eta) \quad y \not\leqslant x\}$ is a cross-section of η;*

(ii) *the congruence η is generated by the set of pairs*

$$R(\eta) = \{(x, y) \mid x \in S(\eta)\,,\ y \in F(\eta) \quad x \equiv y \bmod \eta\} \quad .$$

Proof. Let η be a congruence of $\mathbb{N}^k$ and S the set of minimum elements (for the lexicographic order) in each class modulo η:

$$S = \{x \mid \forall y \in \mathbb{N}^k \quad x \equiv y \bmod \eta \Rightarrow x \preccurlyeq y\} \quad .$$

By definition, S is a cross-section of η. Let F be the set of minimal elements of the complement of S:

$$F = \min\left(\complement S\right).$$

The set $\complement S$ is contained in $F + \mathbb{N}^k$, hence S contains $\complement(F + \mathbb{N}^k)$. Let x be in S and suppose that there exists y in F such that $y \leqslant x$, hence such that there exists z such that $x = y + z$.

Since y is in $\complement S$, there exists y', congruent to y modulo η and strictly less than y in the lexicographic order. The element $y' + z$ is congruent to x modulo η and, by (7.3), is strictly less than it, contradicting the definition of S.

Thus, $S = \complement(F + \mathbb{N}^k)$ and, taking $S(\eta) = S$ and $F(\eta) = F$, and by Corollary 7.2, we have shown (i).

The congruence ζ generated by $R(\eta)$ – a finite set of pairs since it has the same number of elements as $F(\eta)$ – is, by definition, the finest congruence such that, for all pairs (x, y) in $R(\eta)$, x and y are congruent. It is thus finer than η and each of its classes is contained in a class modulo η.

Suppose that ζ is strictly finer than η: that is, there exists at least one class modulo η that contains at least two classes modulo ζ. The minimum element z of one of these two classes necessarily belongs to $\complement S(\eta) = F(\eta) + \mathbb{N}^k$ and can be written $z = y + u$ for some y in $F(\eta)$. It is congruent, modulo ζ, to $z' = x + u$ where (x, y) belongs to $R(\eta)$. Hence $z' \preccurlyeq z$, which contradicts the assumption that z is minimal in its class modulo ζ: thus, η and ζ coincide. ■

Note again that, by Corollary 7.2, the set $S(\eta)$ is the *complement of an ideal* of $\mathbb{N}^k$.

Corollary 7.5 *The congruences of a finitely generated commutative monoid are finitely generated.*

Proof. The congruences of $\mathbb{N}^k$ are finitely generated by Theorem 7.2. If M is a finitely generated commutative monoid, let $\alpha\colon \mathbb{N}^k \to M$ be a surjective morphism, for some k.

For every congruence ξ of M, we write $(\xi)\alpha^{-1}$ for the congruence $\alpha \circ \xi \circ \alpha^{-1}$ of $\mathbb{N}^k$:

$$x \equiv y \bmod (\xi)\alpha^{-1} \quad \Longleftrightarrow \quad x\alpha \equiv y\alpha \bmod \xi \ .$$

We immediately verify[68] that α^{-1} is a lattice isomorphism of congruences of M on the sub-lattice of the congruences of $\mathbb{N}^k$ which are coarser than the map equivalence of α. Hence we deduce that ξ is finitely generated. ■

Remark 7.1 The congruences (of a monoid) are naturally ordered by the 'coarser than' relation: that is, by inclusion of their graphs. We can thus speak of a *chain of congruences* and of an ascending chain condition for congruences. We immediately see that the congruences of a monoid satisfy the ascending chain condition if and only if they are finitely generated. Corollary 7.5 can thus be reformulated in the following form, often more useful for proofs. □

Corollary 7.6 *The congruences of a finitely generated commutative monoid satisfy the ascending chain condition.* ■

Notation The definition of a morphism such as $\alpha\colon \mathbb{N}^k \to M$ sometimes requires us to write out the generators of $\mathbb{N}^k$: we write e_i for the element all of whose coordinates are 0 except for the ith, which is 1.

7.2.2 Lexicographic decomposition

Many results on $\mathbb{N}^k$ can be proved by induction on the dimension k. We have already seen an example in the proof of Dickson's Lemma. The reasoning is always based on the same model which we can formalise once and for all. By simple reference to this 'induction engine', later proofs will be both clarified and shortened.

We will use the natural ordering on the different coordinates of an element of $\mathbb{N}^k$ – that is, the basis of the lexicographic order – to decompose every subset of $\mathbb{N}^k$ into the *disjoint union* of a *finite* collection of sets that are naturally isomorphic to subsets of $\mathbb{N}^{k-1}$ – and we will be able to use the induction hypothesis – and of a subset of $\mathbb{N}^k$, for which will the property will be (more or less) obvious.

More precisely, for every subset X of $\mathbb{N}^k$ and every $y = (y_1, y_2, \ldots, y_k)$ in $\mathbb{N}^k$, we write X^y for the set of elements of X greater than or equal to y and X_y, its complement in X:

$$X^y = \{x \in X \mid y \leqslant x\} = y + \{X - y\} \qquad \text{and} \qquad X_y = \{x \in X \mid y \not\leqslant x\} \ .$$

We define the set $I(y)$ of pairs:

$$I(y) = \{(i, s) \mid 1 \leqslant i \leqslant k \quad \text{and} \quad 0 \leqslant s < y_i\}$$

and the set $P(y)$ of elements of $\mathbb{N}^k$, indexed by $I(y)$:

$$y_{i,s} = (y_1, \ldots, y_{i-1}, s, 0, \ldots, 0) \ .$$

[68] This is an instance of what is sometimes called the *Second Isomorphism Theorem*.

The virtue of these definitions is that for all x in X_y there exists a *unique* $i \in [k]$ such that

$$\forall t < i \qquad y_t \leqslant x_t \qquad \text{and} \qquad x_i < y_i \ , \tag{7.4}$$

and hence exactly one pair (i, s) in $I(y)$ – we take $s = x_i$ – such that

$$x = y_{i,s} + x' \qquad \text{and} \qquad x' \in \mathbb{N}^k \ , \quad x'_i = 0 \ . \tag{7.5}$$

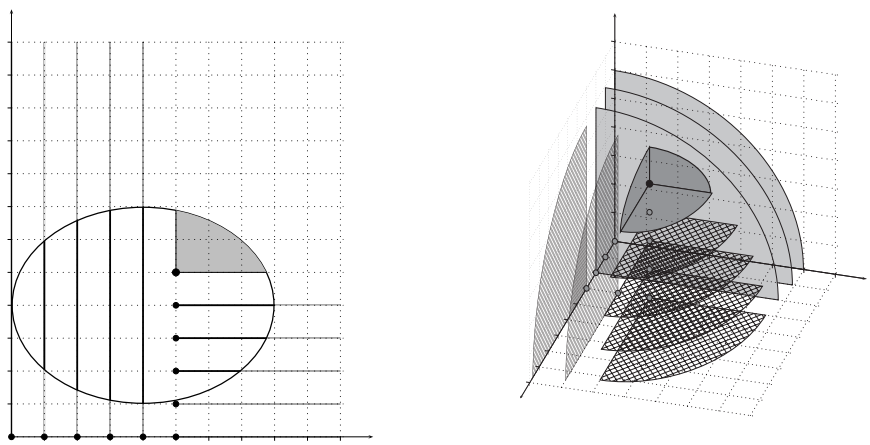
If we write $\mathbb{N}^k_i$ for the submonoid of $\mathbb{N}^k$ consisting of the elements whose ith coordinate is zero:

$$\mathbb{N}^k_i = \{x \in \mathbb{N}^k \mid x_i = 0\} \ ,$$

we can rewrite (7.4) and (7.5) in the form

$$X_y = \bigcup_{(i,s)\in I(y)} \Big[y_{i,s} + \{(X - y_{i,s}) \cap \mathbb{N}^k_i\} \Big] \tag{7.6}$$

where the union is disjoint. We call this process the *lexicographic decomposition* of X. It is illustrated in Figure 7.1 for dimensions 2 and 3.



(a) In two dimensions (b) In three dimensions

Figure 7.1: Lexicographic decomposition

Exercises

- 7.3 Explain how we can sensibly give the same name to the lexicographic order defined on A^* as to that defined on $\mathbb{N}^k$.

- 7.4 (a) Describe a well ordering on $A^{\oplus}$ that satisfies (7.3) and is not a lexicographic order, whatever order is chosen on A.

  (b) Describe a well ordering on $A^{\oplus}$ that satisfies (7.3) and such that the set of minorants of every element is finite.

- 7.5 **Simplifiable commutative monoid.** A commutative monoid M is *simplifiable* if

$$\forall x, y, z \in M \qquad x + z = y + z \quad \Longrightarrow \quad x = y \ .$$

We shall study the simplifiability of commutative monoids in terms of congruences.

(a) Let η be a congruence on M. For each m in M, we define the congruence η_m by

$$\forall x, y \in M \qquad x \equiv y \bmod \eta_m \quad \Longleftrightarrow \quad x + m \equiv y + m \bmod \eta \ .$$

Verify

$$\forall m \in M \qquad \eta \subseteq \eta_m \tag{7.7}$$

$$\forall m \in M \qquad \eta \subseteq \xi \quad \Longrightarrow \quad \eta_m \subseteq \xi_m \tag{7.8}$$

$$\forall m, m' \in M \qquad \eta_{m+m'} = (\eta_m)_{m'} \tag{7.9}$$

Verify that the quotient M by η is simplifiable if and only if for all m in M, $\eta_m = \eta$. (We say in this case that η is *simplifiable*).

(b) Prove:

Proposition 7.7 *Let η be a congruence on a finitely generated monoid M. The family of congruences $\{\eta_m \mid m \in M\}$ contains exactly one simplifiable congruence, and this is its maximum element.*

(c) Prove:

Proposition 7.8 *Let M be a commutative monoid. There exist a commutative group $K(M)$ and a morphism (of monoids) $\gamma\colon M \to K(M)$ such that every morphism from M to a commutative group factorises through γ. The monoid M is simplifiable if and only if γ is injective.*

7.6 (a) Let (x, y) be a pair of elements of $\mathbb{N}^2$. Give a necessary and sufficient condition for the congruence generated by (x, y) to be simplifiable.

(b) Describe all the simplifiable quotients of $\mathbb{N}^2$.

7.3 Subtractive submonoids and affine sets

Just as in an arbitrary monoid K, for every subset P of K and every p in K, we write $p^{-1}P = \{m \in K \mid pm \in P\}$; if M is a commutative monoid, for every subset S of M and every x in M, we write

$$S - x = \{z \in M \mid x + z \in S\} \qquad \text{and} \qquad S - T = \bigcup_{x \in T} S - x \ .$$

A submonoid S of M is *subtractive* if $S - s$ is contained in S for all s in S; that is, if

$$\forall x, y \in M \qquad x \in S \quad \text{and} \quad x + y \in S \quad \Longrightarrow \quad y \in S \ ,$$

or again if $S - S = S$. The image under an inverse morphism of a submonoid is a submonoid, and we also have:

Property 7.1 *Let $\alpha\colon M \to N$ be a surjective morphism. If S is a subtractive submonoid of N, then $S\alpha^{-1}$ is a subtractive submonoid of M.* ■

Lemma 7.9 *A subtractive submonoid S of $A^{\oplus}$ is finitely generated by the minimal elements of $S \setminus \{0\}$.*

Proof. Let $P = \min\{S \setminus \{0\}\}$, then $P^{\oplus}$ is contained in S. If $P^{\oplus} \neq S$, let x be a minimal element of $S \setminus P^{\oplus}$. There exist p in P and y in $A^{\oplus}$ such that $x = p + y$.

Since S is subtractive, y is in S; since $y < x$, y is in $P^{\oplus}$, hence also x, which is a contradiction. ■

Hence we deduce, by Property 7.1:

Proposition 7.10 *A subtractive submonoid of a finitely generated commutative monoid is finitely generated.* ■

Definition 7.1 We say that a subset S of a commutative monoid is *affine*[69] if, for all s in S, $S-s$ is a submonoid; that is, if s, $s+x$ and $s+y$ are in S, then $s+x+y$ is in S.

An element s of an affine set S is *stable* if $S-s$ is a subtractive submonoid. □

Examples 7.1 (i) A first example which justifies the chosen terminology, the intersection of every affine subspace of $\mathbb{Q}^k$ with $\mathbb{N}^k$, is an affine subset of $\mathbb{N}^k$.
(ii) A subtractive submonoid is an affine set, but that is not the case for an arbitrary submonoid.
(iii) Figure 7.2 shows a part (of course) of the affine subset of $\mathbb{N}^2$ generated by the elements $(0,1)$, $(1,3)$ and $(5,6)$, an example which belongs to neither of the two previous families. □

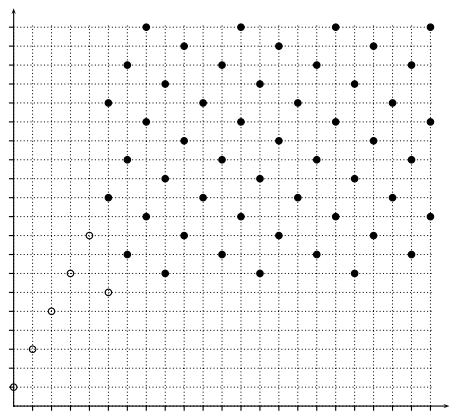


Figure 7.2: An affine set in $\mathbb{N}^2$; the stable elements are in black.

The following result in some sense justifies this new definition.

Proposition 7.11 *The graph of a congruence α on a commutative monoid M is an affine subset of $M \times M$.*

Proof. It is a direct consequence of the regularity and transitivity of α: if (s_1, s_2), $(s_1 + x_1, s_2 + x_2)$ and $(s_1 + y_1, s_2 + y_2)$ are in α, then

$$s_1 + x_1 + y_1 \equiv s_2 + x_2 + y_1 \equiv s_1 + x_2 + y_1 \equiv s_2 + x_2 + y_2 \bmod \alpha \ ,$$

and $(s_1 + x_1 + y_1, s_2 + x_2 + y_2)$ is in α. ■

[69] *Note added in translation:* Eilenberg and Schützenberger, *art. cit.*, call these sets *slices*.

The following properties of affine sets are easy to verify.

Properties 7.2 *Let S be an affine subset of M.*

(i) *For all x in M, $S - x$ is affine.*

(ii) *If N is a commutative monoid and $\alpha\colon N \to M$ is a morphism, then $S\alpha^{-1}$ is an affine subset of N.*

(iii) *The intersection of two affine subsets is affine.* ■

The affine subsets are not subtractive submonoids, but 'almost'. This is what the next result expresses.

Proposition 7.12 *Every affine subset of $\mathbb{N}^k$ has a stable element.*

Proof. Let S be an affine set in $\mathbb{N}^k$ and s in S. Figure 7.2 illustrates the intuition that 'the larger s is, the bigger $S - s$ is', which is formally stated as

$$\forall s, t \in S \qquad s \leqslant t \quad \Longrightarrow \quad S - s \subseteq S - t \ . \tag{7.10}$$

In fact, $t = s + x \in S$ and $y \in S - s$; that is, if $s + y \in S$, then $s + x + y \in S$, or $y \in S - t$. Conversely, s is stable if and only if $S - s$ 'could not get any bigger'; that is,

$$s \in S \quad \text{is stable} \quad \Longleftrightarrow \quad \forall t \in S \qquad s \leqslant t \ \Rightarrow \ S - s = S - t \ . \tag{7.11}$$

In fact, $S - s$ is subtractive if and only if for all x in $S - s$, we have that every y in $S - (s + x)$ is in $S - s$.

For every s in S, let I_s be the ideal generated by the non-zero elements of $S - s$:

$$I_s = [\{S - s\} \setminus \{0\}] + \mathbb{N}^k \ .$$

From (7.10) it follows that $s \leqslant t$ implies $I_s \subseteq I_t$ and since the ideals of $\mathbb{N}^k$ satisfy the ascending chain condition, there exists an s in S such that $I_s = I_t$ for all the t in S greater than s. We shall show that s is a stable element and, to this end, if $I_s = I_t$ for all t greater than s then $S - s = S - t$, which is equivalent by (7.11).

Let y be a minimal element of the union of $(S - u) \setminus (S - s)$ for all $u > s$ and $t = s + x$ the corresponding u: $y \in S - (s + x)$. If $y \in I_s$, then $y = v + w$ with $v \in (S - s) \setminus \{0\}$. Thus s, $s + x$ and $s + v$, hence also $s + x + v$, are in S and $w \in (S - (s + x + v)) \setminus (S - s)$, which contradicts the minimality of y. ■

We therefore deduce:

Proposition 7.13 *Every affine subset of a finitely generated commutative monoid is rational.*

Proof. Let M be a finitely generated commutative monoid, $\alpha\colon \mathbb{N}^k \to M$ a surjective morphism, and S an affine subset of M. Since $S\alpha^{-1}$ is affine (Property 7.2 (ii)) and the image of a rational under a morphism is rational, it suffices to prove the proposition for the affine subsets of $\mathbb{N}^k$. We proceed by induction on k and use the lexicographic decomposition.

Let S be an affine set in $\mathbb{N}^k$ and y in S a stable element. We have $S = S^y \cup S_y$. On one hand,

$$S^y = y + (S - y)$$

is rational since $S - y$ is a subtractive monoid, hence finitely generated (Proposition 7.10). On the other,[70]

$$S_y = \bigcup_{(i,s)\in I(y)} \left\{ y_{i,s} + (S - y_{i,s}) \cap \mathbb{N}_i^k \right\} \quad ,$$

and each of the $(S - y_{i,s}) \cap \mathbb{N}_i^k$ is an affine subset (Properties 7.2 (i) and (iii)) isomorphic to an affine subset of $\mathbb{N}^{k-1}$, hence rational by the induction hypothesis: the finite union of their translations is rational. ■

From this we obtain what is, in a way, the aim of all these developments:

Proposition 7.14 *The graph of a congruence of a finitely generated morphism M is a rational set of $M \times M$.* ■

Exercises

7.7 Let $\alpha\colon A^{\oplus} \to M$ be a morphism in a simplifiable (commutative) monoid M. Show that, for all m in M, there exist two finite sets B and V in $A^{\oplus}$ such that $(m)\alpha^{-1} = B + V^{\oplus}$.

• 7.8 Give some examples of submonoids of $\mathbb{N}^2$ which are not finitely generated.

• 7.9 Verify Properties 7.2.

• 7.10 **Stable submonoid.** We say that a finitely generated submonoid M of $\mathbb{N}^k$ is *stable* if it is the intersection $M = C \cap H$ of a *closed cone*[71] C in $\mathbb{Q}_+^k$ and a subgroup H of $\mathbb{Z}^k$.

Show that M is stable if and only if

$$\forall u \in M\,,\ \forall v \in \mathbb{N}^k \qquad u + v^{\oplus} \subseteq M \quad \Longrightarrow \quad v \in M \ . \tag{7.12}$$

7.4 Semi-linear and semi-simple sets

As we have already noted, commutativity allows us to denote rational sets by particularly simple expressions, expressions whose star height is at most 1. We now give even more precise description.

cf. Exer. I.6.5, p. 161 and p. 204

Let $B = \{b_1, b_2, \ldots, b_k\}$ be a finite subset of M. We have

$$B^{\oplus} = \{n_1 b_1 + n_2 b_2 + \cdots + n_k b_k \mid n_i \in \mathbb{N}\} = {b_1}^{\oplus} + {b_2}^{\oplus} + \cdots + {b_k}^{\oplus} \ . \tag{7.13}$$

We can also write

$$B^{\oplus} = (\mathbb{N}^k)\varphi$$

where φ is the morphism from $\mathbb{N}^k$ to M defined by $(e_i)\varphi = b_i$ for all i in $[k]$. The

[70] $I(y)$ was defined on p. 329.

[71] A subset C of $\mathbb{Q}_+^k$ is a *cone* if it contains all the lines that go through it and the origin; that is, for all $x \in C$, and all $r \in \mathbb{Q}_+$, $r\,x \in C$. A set X in $\mathbb{Q}_+^k$ is *closed* if it contains the limit of every sequence that converges in $\mathbb{Q}_+^k$.

submonoid $B^{\oplus}$ is a *free* commutative monoid if the sum $b_1{}^{\oplus} + b_2{}^{\oplus} + \cdots + b_k{}^{\oplus}$ is unambiguous or, which is the same thing, if φ is *injective*. (Note that a free commutative monoid with rank greater than or equal to 2 is *not a free monoid*.)

Definition 7.2 A subset X of a commutative monoid M is:

(i) *linear* if $X = x + B^{\oplus}$ where x is an arbitrary element of M and B is a *finite* subset of M;

(ii) *simple* if X is linear ($X = x + B^{\oplus}$) and if $B^{\oplus}$ is a free commutative monoid with basis B and the sum $x + B^{\oplus}$ is unambiguous;

(iii) *semi-linear* if it is the finite union of linear sets;

(iv) *semi-simple* if it is the finite disjoint union of simple sets. □

Remark 7.2 The second condition of the definition of a simple set: 'the sum $x+B^{\oplus}$ is unambiguous' always holds if M is *simplifiable*, which it is in the two cases $M = \mathbb{N}^k$ and $M = \mathbb{Z}^k$ with which we are principally concerned. □

By definition, a semi-linear set in M is rational, and a semi-simple set is unambiguously rational. The remarkable fact is that the converse of these two assertions is true.

Proposition 7.15 *A rational set in a commutative monoid is semi-linear; an unambiguous rational set is semi-simple.*

Proof. Let B and C be two subsets of M; by (7.13), we have that

$$B^{\oplus} + C^{\oplus} = (B \cup C)^{\oplus} \tag{7.14}$$

and if the sum on the left hand side is unambiguous then $(B \cup C)^{\oplus}$ is free commutative with basis $B \cup C$. Let

$$X = \bigcup_{i \in I} (x_i + B_i{}^{\oplus}) \qquad \text{and} \qquad Y = \bigcup_{j \in J} (y_j + C_j{}^{\oplus}) \tag{7.15}$$

be two semi-linear sets: I, J, every B_i and every C_j are finite. By distributivity over (7.15), we have

$$X + Y = \bigcup_{(i,j) \in I \times J} \left[(x_i + y_j) + (B_i \cup C_j)^{\oplus} \right] \quad , \tag{7.16}$$

hence $X + Y$ is semi-linear. If the decompositions of X and Y in (7.15) are semi-simple, and if the sum $X + Y$ is unambiguous, then the union in (7.16) is disjoint and every $(x_i + y_j) + (B_i \cup C_j)^{\oplus}$ is simple: $X + Y$ is semi-simple.

Furthermore, we have $(x + B^{\oplus})^{\oplus} = \{0\} \cup \left[x + (\{x\} \cup B)^{\oplus} \right]$ and, more generally, if X is given by (7.15),

$$X^{\oplus} = \{0\} \cup \bigcup_{J \subseteq I} \left[\left(\sum_{j \in J} x_j \right) + \left[\bigcup_{j \in J} (\{x_i\} \cup B_i) \right]^{\oplus} \right]$$

and hence $X^{\oplus}$ is semi-linear. For $X^{\oplus}$ to be a free submonoid in M with basis X,

then X must reduce to a single element and $X^{\oplus}$ has to be isomorphic to $\mathbb{N}$, in which case $X^{\oplus}$ is simple.

Thus the family of semi-linear sets in M, closed under union by definition, is closed under sum and star; as it contains the finite subsets, it contains all the rational sets of M. Likewise, the family of semi-simple sets, closed under disjoint union, is closed under unambiguous sum and free star; as it contains all the finite subsets, it contains all the unambiguous rational subsets. ■

7.5 Rationals of $\mathbb{N}^k$

We shall now prove the two principal results of this section:

Theorem 7.3 *The rational subsets of $\mathbb{N}^k$ form an effective Boolean algebra.*

Theorem 7.4 *The rational subsets of $\mathbb{N}^k$ are unambiguous.*

In the next subsection we deduce from this, thanks to Proposition 7.14, the same properties for the rational subsets of finitely generated commutative monoids. We start by proving a lemma which essentially states that every finitely generated submonoid is a finite union of free commutative submonoids. The following step is the proof that the semi-simple subsets of $\mathbb{Z}^k$ form a Boolean algebra and, to this end, that the positive integer solutions of a Diophantine linear system form a semi-simple subset of $\mathbb{N}^k$, an interesting result in itself (Proposition 7.17). Theorems 7.3 and 7.4 can then be deduced together, by a sort of somersault.

The effectiveness of the whole process is ensured by the fact that each successive step is effective (which we prove just for the solution of a sufficient class of Diophantine systems).

7.5.1 The Freedom Lemma

Lemma 7.16 *Let $X = x + B^{\oplus}$ be a linear set in a* simplifiable *commutative monoid M. Then X can also be written as a* finite union *of simple sets:*

$$X = \bigcup_{i \in I} \{x_i + D_i{}^{\oplus}\} \ ,$$

with $D_i \subseteq B$ for all i in I.

Remark 7.3 The union of $x_i + D_i{}^{\oplus}$ is not in general disjoint and the lemma does not state that X is semi-simple, which is the point of Theorem 7.4. □ *cf. also Exer. 7.11*

Proof. We can clearly suppose that $x = 0$: if $B^{\oplus} = \bigcup_{i \in I} \{y_i + D_i{}^{\oplus}\}$, then $x + B^{\oplus} = \bigcup_{i \in I} \{(x + y_i) + D_i{}^{\oplus}\}$. We prove the lemma by induction on the number of elements p of B. If $X = B^{\oplus}$ is a free commutative monoid, and in particular if $p = 1$ (since M is simplifiable), the proposition is true.

If $B^{\oplus}$ is not free, there exists at least one relation

$$\sum_{i=1}^{p} n_i\, b_i = \sum_{j=1}^{p} m_j\, b_j \quad , \tag{7.17}$$

where the n_i and the m_j are in $\mathbb{N}$. Possibly by renumbering the elements of B and since M is simplifiable, (7.17) can be rewritten as

$$\sum_{i=1}^{q} t_i\, b_i = \sum_{j=q+1}^{p} t_j\, b_j \quad , \tag{7.18}$$

where $1 \leqslant q < p$ and the t_l are integers which are not all zero. For each i in $[q]$, we define the sets

$$Y_i = \{s\, b_i \mid 0 \leqslant s < t_i\} \quad , \qquad B_i = B \setminus \{b_i\} \quad , \qquad Z_i = Y_i + B_i^{\oplus} \quad ,$$

and we set

$$Z = \bigcup_{i \in [q]} Z_i \quad .$$

Showing that $X = Z$ completes the proof since we can apply the induction hypothesis to each set $s\, b_i + B_i^{\oplus}$.

Since $Y_i \subset X$ and $B_i \subset X$ for all i and since X is a submonoid, we have $Z \subseteq X$. Suppose that $X \setminus Z$ is non-empty, and let $r = (r_1, r_2, \ldots, r_p)$ such that

$$z = \sum_{l=1}^{p} r_l\, b_l \quad , \tag{7.19}$$

belongs to $X \setminus Z$, and choose r to be minimal in the *lexicographic order* of $\mathbb{N}^p$. If, for all i in $[q]$, we have $t_i \leqslant r_i$, then substituting (7.18) into (7.19) we get

$$z = \sum_{i=1}^{q} (r_i - t_i)\, b_i + \sum_{j=q+1}^{p} (r_j + t_j)\, b_j \quad ,$$

which contradicts the minimality of r. If there exists an i in $[q]$ such that $r_i < t_i$, then z belongs to Z_i, which is a contradiction. ■

7.5.2 Positive solutions of Diophantine linear systems

A system of m linear equations with integer coefficients and k unknowns (a Diophantine system) can be seen as a morphism from $\mathbb{N}^k$ to $\mathbb{Z}^m$. Another statement that paves the way to the proof of Theorems 7.3 and 7.4 is the characterisation by the following proposition of the set of positive integer solutions of a Diophantine system.

Proposition 7.17 *Let $\alpha\colon \mathbb{N}^k \to \mathbb{Z}^m$ be a morphism. For all c in $\mathbb{Z}^m$, $(c)\alpha^{-1}$ is a semi-simple subset of $\mathbb{N}^k$.*

Proof. Set $X = (c)\alpha^{-1} = \{x \in \mathbb{N}^k \mid x\alpha = c\}$. If X is finite, it is semi-simple. In particular, if $k = 1$, X is either empty or reduces to a single element (if the system is not completely degenerate; if it is, $X = \mathbb{N}$ is simple).

If X is infinite, it contains at least two comparable elements t and $t' = t + y$, with $y \neq 0$. But $t\alpha = t'\alpha = c$ implies $y\alpha = 0$. Every element x of X can be written uniquely in the form

$$x = n\,y + z \qquad \text{with} \quad n \in \mathbb{N}\ , \quad z \in X \quad \text{and} \quad y \not\leqslant z\ ,$$

so that X is the *unambiguous sum*

$$X = y^{\oplus} + X_y \tag{7.20}$$

and $X_y = \{z \in X \mid y \not\leqslant z\}$ can be written as a disjoint union (Equation (7.6)):

$$X_y = \bigcup_{(i,s)\in I(y)} \Big[y_{i,s} + \{(X - y_{i,s}) \cap \mathbb{N}_i^k\}\Big]\ . \tag{7.21}$$

For each (i, s) in $I(y)$,

$$(X - y_{i,s}) \cap \mathbb{N}_i^k = \{x \in \mathbb{N}_i^k \mid x\alpha = c - (y_{i,s})\alpha\} \tag{7.22}$$

is a semi-simple subset by induction: X_y, and hence X, is semi-simple. ■

Remark 7.4 This proof is *not effective*: to make it so, we must also give an algorithm which allows us to compute a y such that $y\alpha = 0$ or to decide that none exists and in that case to compute all the solutions of $x\alpha = c$. This is possible, and even standard, but we will content ourselves here with a simplified proof which holds for the morphisms we use subsequently. □

cf. bibliographic notes p. 373

Definition 7.3 We say that a morphism $\alpha\colon \mathbb{N}^k \to \mathbb{Z}^m$ is *split* if there exist two morphisms $\beta\colon \mathbb{N}^p \to \mathbb{N}^m$ and $\gamma\colon \mathbb{N}^q \to \mathbb{N}^m$ such that $p + q = k$ (and hence $\mathbb{N}^k = \mathbb{N}^p \times \mathbb{N}^q$) and, for all (x, y) in $\mathbb{N}^p \times \mathbb{N}^q$, $(x, y)\alpha = x\beta - y\gamma$. □

Lemma 7.18 *Let $\alpha\colon \mathbb{N}^p \times \mathbb{N}^q \to \mathbb{Z}^m$ be a split morphism and let c be in $\mathbb{Z}^m$. We can effectively compute an expression for $(c)\alpha^{-1}$ as a semi-simple subset of $\mathbb{N}^{p+q}$.*

Proof. (i) We first show that we can decide whether $X = (c)\alpha^{-1}$ is empty or not, and, if it is not, compute at least one of its elements.

Note that the morphism $\beta\colon \mathbb{N}^p \to \mathbb{N}^m$ is determined by the choice of a set E of p vectors in $\mathbb{N}^m$ and similarly $\gamma\colon \mathbb{N}^q \to \mathbb{N}^m$ is determined by the choice of a set F of q vectors in $\mathbb{N}^m$ whose opposites form a set F' of q vectors of $\mathbb{Z}^m$. We set $B = E \cup F'$. Furthermore, $X = (c)\alpha^{-1}$ is non-empty if and only if c belongs to $B^{\oplus}$.

We can decide whether B is free, and, if it is, decide whether there exists an element x of $\mathbb{Q}^k$ in $(c)\alpha^{-1}$, compute x if it exists and hence decide whether x belongs to $\mathbb{N}^k$.

If B is not free, Lemma 7.16 allows us to write $B^{\oplus}$ as a finite union of sets $x_i + D_i^{\oplus}$, (where the D_i are free and included in B) and hence to answer the question by solving the previous problem a finite number of times, replacing c by $c - x_i$ and B by D_i.

(ii) By induction on k, we effectively compute a description of $X = (c)\alpha^{-1}$ as a semi-simple set.

If $k = 1$, X is either empty or reduces to a single element that we can compute (or is equal to $\mathbb{N}$ if the system is degenerate).

If $k > 1$, we first look for a solution of $y\alpha = 0$. If one exists, we use Equations (7.20) and (7.21) to solve a finite set of systems of size $k - 1$.

If $(0)\alpha^{-1}$ is empty, we look for a solution of $x\alpha = c$. If none exists, the property is verified. Otherwise, we compute one, say t, and $X = (c)\alpha^{-1}$ is the *disjoint union* of

$$X^t = \{t\}$$

and of X_t, which can be written as the disjoint union

$$X_t = \bigcup_{(i,s)\in I(t)} \left[t_{i,s} + \{(X - t_{i,s}) \cap \mathbb{N}_i^k\} \right] \quad ,$$

every element of which is, by induction, an effectively computable semi-simple set. ■

7.5.3 Semi-simple subsets of $\mathbb{Z}^k$

We now show, as the penultimate step, that the semi-simple subsets of $\mathbb{Z}^k$ form a Boolean algebra.

Proposition 7.19 *The semi-simple subsets of $\mathbb{Z}^k$ are closed under intersection.*

Proof. By distributivity of intersection over union, it suffices to show that the intersection of two simple sets is semi-simple. Let

$$X = x + B^{\oplus} \qquad \text{and} \qquad Y = y + C^{\oplus}$$

be two simple sets; that is,

$$X = x + (\mathbb{N}^p)\varphi \qquad \text{and} \qquad Y = y + (\mathbb{N}^q)\psi \quad ,$$

where $\varphi\colon \mathbb{N}^p \to \mathbb{Z}^k$ and $\psi\colon \mathbb{N}^q \to \mathbb{Z}^k$ are two *injective* morphisms. We define two morphisms from $\mathbb{N}^{p+q} = \mathbb{N}^p \times \mathbb{N}^q$ to $\mathbb{Z}^k$ in the following way:

$$\begin{aligned} \sigma\colon \mathbb{N}^p \times \mathbb{N}^q \to \mathbb{Z}^k \qquad & \text{by} \quad (z,t)\sigma = z\varphi - t\psi \quad , \\ \tau\colon \mathbb{N}^p \times \mathbb{N}^q \to \mathbb{Z}^k \qquad & \text{by} \quad (z,t)\tau = z\varphi \quad . \end{aligned}$$

Then let

$$W = \{(z,t) \in \mathbb{N}^p \times \mathbb{N}^q \mid (z,t)\sigma = y - x\} = \{(z,t) \in \mathbb{N}^p \times \mathbb{N}^q \mid x + z\varphi = y + t\psi\} \quad .$$

By Proposition 7.17, W is semi-simple. Since φ and ψ are injective, τ is injective on W and $X \cap Y = x + W\tau$ is semi-simple. ■

We note that the morphisms σ and τ are *split*. Proposition 7.19 is hence *effective*. To show closure under complement, we prove a preliminary property.

Lemma 7.20 *The following subsets of* $\mathbb{Z}^k$*:*

$$\mathbb{Z}^k \quad , \qquad \mathbb{Z}^k \setminus \mathbb{N}^k \quad , \qquad \mathbb{N}^p \times \mathbb{N}^q \setminus \mathbb{N}^p \times 0 \quad , \quad \textit{for } p+q=k \, ,$$

are semi-simple.

Proof. We first note that if

$$X_1 = (x_1 + {B_1}^{\oplus}) \subseteq M_1 \qquad \text{and} \qquad X_2 = (x_2 + {B_2}^{\oplus}) \subseteq M_2$$

are two simple sets in two commutative monoids M_1 and M_2 such that their complements $Y_1 = M_1 \setminus X_1$ and $Y_2 = M_2 \setminus X_2$ are semi-simple, then

$$X_1 \times X_2 = (x_1 + x_2) + (B_1 \cup B_2)^{\oplus}$$

and

$$M_1 \times M_2 \setminus X_1 \times X_2 = (Y_1 \times Y_2) \cup (X_1 \times Y_2) \cup (Y_1 \times X_2)$$

are two semi-simple sets in $M_1 \times M_2$. This property extends by induction to the product of an arbitrary number of commutative monoids M_i.

The lemma then follows from $\mathbb{N} = 1^{\oplus}$, $\mathbb{Z} \setminus \mathbb{N} = \{-1\} + \{-1\}^{\oplus}$ and $\mathbb{N} \setminus \{0\} = \{1\} + \{1\}^{\oplus}$ being simple. ∎

Proposition 7.21 *The semi-simple subsets of* $\mathbb{Z}^k$ *are closed under complement.*

Proof. Since the complement of a union is the intersection of the complements, it is enough, thanks to Proposition 7.19, to show that the complement of a simple subset of $\mathbb{Z}^k$, $X = x + B^{\oplus}$, is semi-simple. Since

$$\mathbb{Z}^k \setminus X = x + (\mathbb{Z}^k \setminus B^{\oplus}) \quad ,$$

we may assume that $x = 0$ and $X = B^{\oplus} = \{b_1, b_2, \ldots, b_p\}^{\oplus}$.

We write $B^{\circledcirc}$ for the subgroup of $\mathbb{Z}^k$ generated by B and we have

$$\mathbb{Z}^k \setminus B^{\oplus} = \left(\mathbb{Z}^k \setminus B^{\circledcirc}\right) \cup \left(B^{\circledcirc} \setminus B^{\oplus}\right) \quad ,$$

where the union is disjoint: it suffices to show that each of the terms is semi-simple.

We first suppose that $p = k$. Since the $\{b_1, b_2, \ldots, b_k\}$ are linearly independent, $B^{\circledcirc}$ is isomorphic to $\mathbb{Z}^k$; the isomorphism sends $B^{\oplus}$ to $\mathbb{N}^k$. By Lemma 7.20, $\mathbb{Z}^k \setminus \mathbb{N}^k$ is semi-simple, hence $B^{\circledcirc} \setminus B^{\oplus}$ is too. Since $p = k$, the *quotient group* $\mathbb{Z}^k / B^{\circledcirc}$ is finite, and $\mathbb{Z}^k \setminus B^{\circledcirc}$ is a disjoint finite union of cosets $z + B^{\circledcirc}$. Each of them is semi-simple since $B^{\circledcirc}$, isomorphic to $\mathbb{Z}^k$, is semi-simple by Lemma 7.20.

If $k = p + q$, we can find q elements $\{b_{p+1}, b_{p+2}, \ldots, b_k\}$ such that the set $C = \{b_1, b_2, \ldots, b_k\}$ is made up of linearly independent vectors; that is, such that $C^{\oplus}$ is simple. We then write

see proof of Th. III.7.1, p. 491, if necessary.

$$\mathbb{Z}^k \setminus B^{\oplus} = \left(\mathbb{Z}^k \setminus C^{\oplus}\right) \cup \left(C^{\oplus} \setminus B^{\oplus}\right) \quad .$$

By the above, $\mathbb{Z}^k \setminus C^{\oplus}$ is semi-simple. The monoid $C^{\oplus}$ is isomorphic to $\mathbb{N}^p \times \mathbb{N}^q$ under an isomorphism that sends $B^{\oplus}$ to $\mathbb{N}^p \times 0$: hence, $C^{\oplus} \setminus B^{\oplus}$ is semi-simple by Lemma 7.20. ∎

7.5.4 Proof of Theorems 7.3 and 7.4

The two results will be proved simultaneously, and indeed we cannot separate them, by the proof that the rationals of $\mathbb{N}^k$ are semi-simple subsets of $\mathbb{Z}^k$. We start by showing an elementary *Fatou property*:

cf. § 6.1.3, p. 303

Property 7.3 *A semi-linear (resp. semi-simple) subset of $\mathbb{Z}^k$ included in $\mathbb{N}^k$ is a semi-linear (resp. semi-simple) subset of $\mathbb{N}^k$.*

Proof. Let $X = \bigcup \left(z_i + C_i^{\oplus}\right)$ be a semi-linear (or semi-simple) subset of $\mathbb{Z}^k$; that is, the z_i and C_i are in $\mathbb{Z}^k$. Suppose $X \subseteq \mathbb{N}^k$; then, each z_i belongs to $\mathbb{N}^k$ and, for all i, and for each c in C_i, $z_i + n\,c$ belongs to $\mathbb{N}^k$ for every integer n. This implies that all the components of c are non-negative: $C_i \subseteq \mathbb{N}^k$ and X is a semi-linear (or semi-simple) subset of $\mathbb{N}^k$. ■

We show that the rationals of $\mathbb{Z}^k$ – that is, the semi-linear subsets – are semi-simple. Let X be a semi-linear subset of $\mathbb{Z}^k$. By Lemma 7.16, X is a finite union, *not necessarily disjoint*, of *simple* subsets:

$$X = X_1 \cup X_2 \cup \cdots \cup X_p \ .$$

By Proposition 7.21, each subset $\mathbb{Z}^k \setminus X_i$ is semi-simple. Hence, by Proposition 7.19,

$$X' = \mathbb{Z}^k \setminus X = \bigcap_{i \in [p]} \left(\mathbb{Z}^k \setminus X_i\right)$$

is semi-simple and $X = \mathbb{Z}^k \setminus X'$ is too.

Now let X be a rational of $\mathbb{N}^k$. We can view X as a rational of $\mathbb{Z}^k$ and hence X is a semi-simple subset of $\mathbb{Z}^k$, hence also of $\mathbb{N}^k$ by Property 7.3. ■

In particular, an ideal of $\mathbb{N}^k$, finitely generated, is rational and hence so is its complement. From the characterisation of a cross-section $S(\eta)$ for every congruence η of $\mathbb{N}^k$ (Theorem 7.2), we obtain:

Corollary 7.22 *Every congruence of $\mathbb{N}^k$ admits a rational cross-section.* ■

cf. p. 254

Remark 7.5 The recognisable subsets of $\mathbb{N}^k$ are characterised by Corollary 2.20: a subset of $\mathbb{N}^k$ is recognisable if and only if it is a finite union of products of recognisable subsets of $\mathbb{N}$; that is, if and only if it is – apart from a finite set – a finite union of 'infinite grids' of dimension r ($r \leqslant k$) whose sides are 'parallel to the axes'. It is then known that the following holds.

Proposition 7.23 [**Ginsburg & Spanier**] *It is decidable whether a rational subset of $\mathbb{N}^k$ is recognisable.*

This result about the decidability of $\operatorname{Rec}\mathbb{N}^k$ in $\operatorname{Rat}\mathbb{N}^k$ contrasts with the corresponding undecidability results[72] when dealing with direct products of free monoids

[72] See Corollary IV.4.2 p. 570, Subsection IV.5.4 p. 588 and Problem IV.6.1 p. 615 (*Note added in translation:* the latter not being a problem anymore; *cf.* Note IV.28, p. 615).

with several generators. Ginsburg and Spanier's proof[73] rests on the description of rational subsets of $\mathbb{N}^k$ by formulas in 'Presburger arithmetic', an aspect which we have not dealt with. It must surely be possible to make a proof entirely in the framework set up here.[74] □

Exercise

7.11 Show that we can prove a stronger version of Lemma 7.16:

Lemma 7.24 *Let $X = x + B^{\oplus}$ be a linear set of a simplifiable commutative monoid M. Then X can also be written as a* disjoint *finite union of simple sets*

$$X = \bigcup_{i \in I} \{x_i + D_i{}^{\oplus}\}$$

where $D_i \subseteq B$ for all i in I.

7.6 Rationals of commutative monoids

The rational subsets of commutative monoids inherit properties of those of $\mathbb{N}^k$, thanks to the rationality of the congruences of $\mathbb{N}^k$, which allow us to prove:

Proposition 7.25 *Let $\alpha\colon \mathbb{N}^k \to M$ be a surjective morphism. The inverse image $X\alpha^{-1}$ of a rational X of M is a rational of $\mathbb{N}^k$.*

Proof. The graph $\widehat{\alpha\alpha^{-1}}$ of the map equivalence of α is a rational of $\mathbb{N}^k \times \mathbb{N}^k$. Let X, a rational of M, and R, a rational of $\mathbb{N}^k$, be such that $R\alpha = X$. We then note that

$$X\alpha^{-1} = \left[\widehat{\alpha\alpha^{-1}} \cap (\mathbb{N}^k \times R)\right] \pi_1 \ ,$$

where π_1 is the first projection of $\mathbb{N}^k \times \mathbb{N}^k$ on to $\mathbb{N}^k$. ■

Theorem 7.5 *The rational subsets of a finitely generated commutative monoid form an effective Boolean algebra.*

Proof. Let $\alpha\colon \mathbb{N}^k \to M$ be a surjective morphism, for some k. As α is surjective, $M \setminus X = \left[\mathbb{N}^k \setminus (X\alpha^{-1})\right]\alpha$ and $\operatorname{Rat} M$, closed under complement, is a Boolean algebra. ■

Corollary 7.26 *The rational subsets of a commutative monoid are closed under intersection and difference.*

Proof. Let X and Y be two rationals of a commutative monoid M; there exists a submonoid M' of M, finitely generated, such that X and Y are two rationals of M'. Since $\operatorname{Rat} M'$ is a Boolean algebra, $X \cap Y$ and $X \setminus Y$ are two rationals of M', hence of M. ■

---

[73] *Cf.* [102]

[74] *Note added in translation:* When writing the French edition, I was not aware that Ph. Gohon already answered that question almost twenty years earlier in [106].

Theorem 7.6 *The rational subsets of a commutative monoid are unambiguous.*

Proof. Let X be a rational of a commutative monoid M, which we assume is finitely generated, and $\alpha\colon \mathbb{N}^k \to M$ a surjective morphism.

Let S be a rational cross-section of $\alpha\alpha^{-1}$. Then $W = S \cap X\alpha^{-1}$ is a rational of $\mathbb{N}^k$, hence unambiguous, and $W\alpha = X$ is an unambiguous rational since α is injective over S hence over W. ∎

Exercises

- 7.12 Prove:

Corollary 7.27 *Let $\beta\colon M \to N$ be a morphism on commutative monoids. If M is finitely generated and X is a rational of N, then $X\beta^{-1}$ is a rational of M.*

- 7.13 Prove the Fatou property for commutative monoids:

Corollary 7.28 *Let M' be a finitely generated submonoid of a commutative monoid M. A rational of M contained in M' is a rational of M'.*

- 7.14 Prove:

Corollary 7.29 *If X and Y are two rationals of a finitely generated commutative monoid M, then $Y - X$ is a rational of M.*

8 Star height of group languages

cf. Sec. I.6, p. 157 We have already said that the determination of the star height of a rational language is a hard problem in the general case. It turns out that in the case of group languages: that is, languages whose syntactic monoid is a group, the problem can be solved with ideas we have developed in this chapter.

Theorem 8.1 *The universal automaton of a rational group language K contains a sub-automaton of minimal loop complexity that recognises K.*

Since the universal automaton of a rational language is finite, we can enumerate its sub-automata, keeping those that recognise the language, and from among them find those of minimal complexity. We therefore have:

Corollary 8.1 [**McNaughton**] *The star height of a rational group language is computable.* ∎

cf. p. 167 Furthermore, the same theorem allows us to prove Theorem I.6.2 directly, which we proved in Chapter I by a subtle combinatorial method.[75]

Corollary 8.2 *The language W_q over $\{a, b\}^*$ that consists of the words in which the letters a and b appear the same number of times modulo 2^q is of star height q.*

[75] *The ultimate goal of mathematics is to eliminate any need for intelligent thought. Alfred N. Whitehead.*

Proof. The syntactic monoid of W_q is the group $\mathbb{Z}/2^q\mathbb{Z}$ and the image of W_q in this group is the identity. We have already seen, by a simple calculation, that in this case the trim universal automaton is isomorphic to (the Cayley graph of) the group itself (with the images of a and b as generators), which is of loop complexity q, another simple calculation. ■

cf. Exer. 6.4, p. 304
cf. Exer. I.6.3, p. 161

Remark 8.1 A group language K being neither a left nor a right ideal, $\mathcal{U}_K$ always contains a non-accessible state and a non-co-accessible state. In the rest of this section we shall consider only the *trim part* of $\mathcal{U}_K$, which we shall still call the *universal automaton* of K, and write $\mathcal{U}_K$. □

cf. Exer. 4.4, p. 278

Example 8.1 We return, with a slight modification,[76] to Example I.6.3. Let H_2 and H_3 be the languages over $A^* = \{a, b\}^*$ consisting of words whose number of a's is *congruent* to the number of b's *plus* 1 modulo 2 and 3 respectively and H_6 their union:

cf. p. 166

$$H_2 = \{f \mid |f|_a - |f|_b \equiv 1 \bmod 2\} \ , \qquad H_3 = \{f \mid |f|_a - |f|_b \equiv 1 \bmod 3\}$$
$$\text{and} \qquad H_6 = \{f \mid |f|_a - |f|_b \equiv 1, 3, 4 \text{ or } 5 \bmod 6\} \ .$$

The minimal automaton of H_6 is the 'double ring' of length 6 whose loop complexity is 3. The minimal automata of H_2 and H_3 have complexity 1 and 2, hence the star height of H_6 is at most 2 (*cf.* Figure 8.1).

cf. Exam. I.6.2 (f), p. 160

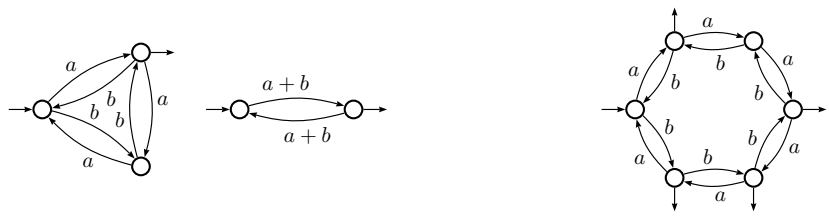


Figure 8.1: An automaton of minimal loop complexity (left) which is not the minimal automaton (right) for H_6

Figure 8.2 shows the universal automaton of H_6 in simplified form. We see, all the better for its grey background, a sub-automaton of this universal automaton which recognises H_6, with a minimal complexity. This sub-automaton is equal to the union of the minimal automata of H_2 and H_3 seen above, and this is not a coincidence. □

To prove Theorem 8.1, we shall begin by relating the loop complexity of an automaton to that of its image under a conformal morphism.

Remember Def. I.6.2, p. 160 and Def. 3.1, p. 258

Theorem 8.2 [**McNaughton**] *If $\varphi\colon \mathcal{B} \to \mathcal{A}$ is a conformal morphism, the complexity of $\mathcal{B}$ is greater than or equal to that of $\mathcal{A}$: that is,* $\mathsf{lc}(\mathcal{B}) \geqslant \mathsf{lc}(\mathcal{A})$.

[76] The definition of the languages is a little less simple, but the corresponding universal automaton is a bit simpler.

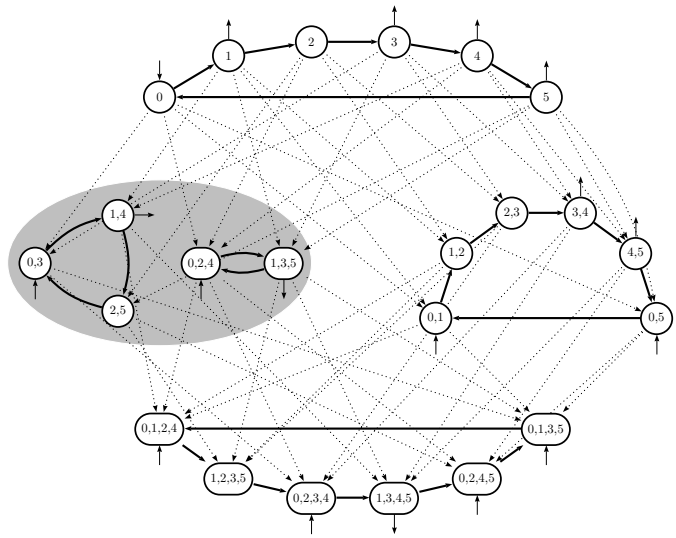


Figure 8.2: The universal automaton of H_6. The bold arrows represent a double transition, one labelled a in the direction of the arrow and one labelled b in the opposite direction; the dotted arrows represent the spontaneous transitions, and the universal automaton is obtained by eliminating them by backward closure.

cf. p. 258 The example of Figure 3.2 shows that if φ is not conformal, it can very well happen that $\mathsf{lc}(\mathcal{B})$ is *strictly less* than $\mathsf{lc}(\mathcal{A})$. We first show a lemma:

Lemma 8.3 *Let $\varphi\colon \mathcal{B} \to \mathcal{A}$ be a conformal morphism. For every ball $\mathcal{P}$ in $\mathcal{A}$, there exists a ball $\mathcal{Q}$ in $\mathcal{B}$ such that the restriction of φ to $\mathcal{Q}$ is a conformal morphism from $\mathcal{Q}$ to $\mathcal{P}$.*

Proof. This lemma (like the theorem) is in fact a proposition about graphs, but we shall use automata-theoretic notions to simplify the proof. We assume, possibly by changing them all, that each transition of $\mathcal{A}$ bears a distinct label, and that each state of $\mathcal{A}$ is both initial and final; this may change the language accepted by $\mathcal{A}$ but has no effect on its loop complexity. The words of the language recognised by $\mathcal{A}$ (resp. by a sub-automaton $\mathcal{P}$ of $\mathcal{A}$) describe paths in the *graph* $\mathcal{A}$ (resp. in the graph $\mathcal{P}$). The transitions of $\mathcal{B}$ are labelled in such a way that φ is an automata morphism and each state of $\mathcal{B}$ is both initial and final.

Let $\mathcal{P}$ be a ball in $\mathcal{A}$ and $\mathcal{R} = \mathcal{P}\varphi^{-1}$. Set $n = \|\mathcal{R}\|$ and $m = \|\mathcal{P}\|$ to be the number of states of $\mathcal{R}$ and $\mathcal{P}$ respectively and consider a circuit (hence a word) w which visits all the paths in $\mathcal{P}$ of length less than 2^{n+m}. The circuit w^n is a path in $\mathcal{P}$ which can be lifted to a path in $\mathcal{R}$ (since φ is conformal). By the proof of the block star lemma, a factor w^k is the label of a *circuit* in $\mathcal{R}$; let $\mathcal{Q}$ be the ball in $\mathcal{R}$, and hence in $\mathcal{B}$, that contains this circuit. By construction, $\mathcal{Q}$ recognises all words of length less than 2^{n+m} of the language recognised by $\mathcal{P}$, hence $\mathcal{Q}$ is equivalent to $\mathcal{P}$, hence all the paths in $\mathcal{P}$ become paths in $\mathcal{Q}$: thus, φ is conformal from $\mathcal{Q}$ to $\mathcal{P}$. ■

cf. Lem. I.1.14, p. 71

cf. Exer. I.3.10, p. 111 and p. 195

Proof of Theorem 8.2. Suppose that the property is false, and proceed by *reductio ad absurdum*. Among the automata which are sent by a conformal morphism to an automaton of strictly greater complexity, let $\mathcal{B}$ be an automaton of minimal complexity d, and let $\mathcal{A}$, of complexity c, be the image of $\mathcal{B}$ under a conformal morphism: thus, $c > d$.

If $d = 0$, the length of the paths in $\mathcal{B}$ is bounded and it is impossible for φ to be conformal, hence $d > 0$.

By definition, there exists a ball $\mathcal{P}$ in $\mathcal{A}$ of complexity c and, by Lemma 8.3, a ball $\mathcal{Q}$ in $\mathcal{B}$ whose image under φ is $\mathcal{P}$. This ball is of complexity at most d but also, by the minimality of d, at least d. There exists a state q in $\mathcal{Q}$ such that

$$\mathsf{lc}(\mathcal{Q} \setminus \{q\}) = d - 1 \ .$$

Let $p = q\varphi$, $\mathcal{P}' = \mathcal{P} \setminus \{p\}$ and $\mathcal{Q}' = \mathcal{Q} \setminus \{p\varphi^{-1}\}$; we have

$$\mathsf{lc}(\mathcal{Q}') \leqslant \mathsf{lc}(\mathcal{Q} \setminus \{q\}) = d - 1 \qquad \text{and} \qquad \mathsf{lc}(\mathcal{P}') \geqslant c - 1 > d - 1 \ .$$

Every path in $\mathcal{P}'$ is a path in $\mathcal{P}$ which does not visit p, hence the image of a path in $\mathcal{Q}$ which does not go through any of the vertices of $p\varphi^{-1}$; that is, the image of a path in $\mathcal{Q}'$: thus, φ is a conformal morphism from $\mathcal{Q}'$ to $\mathcal{P}'$, which contradicts the assumed minimality of d. ■

Proof of Theorem 8.1. Let $\mathcal{B}$ be an automaton of minimal complexity which recognises K and $\varphi\colon \mathcal{B} \to \mathcal{U}_K$ a morphism from $\mathcal{B}$ to the universal automaton of K. If φ is a *conformal* morphism from $\mathcal{B}$ to its image $\mathcal{B}\varphi$ in $\mathcal{U}_K$, this sub-automaton of $\mathcal{U}_K$ is of lesser or equal complexity to that of $\mathcal{B}$ by Theorem 8.2 and the property is proved. However, in the general case φ is not conformal. The proof comes down to showing that nonetheless φ is conformal on some sub-automata of $\mathcal{B}$ (on some balls) which are crucial for the complexity. We start by proving some properties of balls of the universal automaton of a group language.

In what follows, $K \subseteq A^*$ is a group language, $\alpha\colon A^* \to G$ is the syntactic morphism, $P = K\alpha$, and $\mathcal{A}_K = \langle G, A, \delta, 1_G, P \rangle$ is a complete accessible deterministic automaton that recognises K. For w in A^* and g in G we therefore write $g \cdot w$ for $g(w\alpha)$, multiplication in G.

Let us begin by recalling some properties that are independent of the assumption that G is a group: that $\mathcal{U}_K$, the universal automaton of K, is obtained by considering the factorisations (L, R) of P in G and that if *cf. Prop. 4.1, p. 277*

$$(L_1, R_1) \xrightarrow[\mathcal{U}_K]{a} (L_2, R_2)$$

is a transition of $\mathcal{U}_K$, then $L_1(a\alpha)R_2 \subseteq P$ and hence

$$L_1 \cdot a \subseteq L_2 \qquad \text{and} \qquad (a\alpha)^{-1}R_2 \subseteq R_1 \ .$$

Lemma 8.4 *The balls of $\mathcal{U}_K$ are deterministic and complete.*

Proof. Let (L_1, R_1) and (L_2, R_2) be two states of $\mathcal{U}_K$ belonging to the same ball. There exists u and v in A^* such that $L_1 \cdot u \subseteq L_2$ and $L_2 \cdot v \subseteq L_1$. As G is a

group, the action of every element is injective and $\|L_1\| \leqslant \|L_2\| \leqslant \|L_1\|$, hence $\|L_1\| = \|L_2\|$ and $L_1 \cdot u = L_2$. That is, L_2 is *uniquely determined* by L_1 and u: the ball is deterministic.

Furthermore, if (L, R) is a factorisation of P, then $(L(u\alpha), (u\alpha)^{-1}R)$ is also a factorisation of P, for all u in A^*, and there exists a transition labelled u from the first to the second. For all u, there exists v such that $(u\,v)\alpha = 1_G$, and hence a transition labelled v from $(L(u\alpha), (u\alpha)^{-1}R)$ to (L, R). Thus, $(L(u\alpha), (u\alpha)^{-1}R)$ belongs to the same ball as (L, R), and the ball is complete. ∎

Lemma 8.5 *For every integer k, there exists a word w_k in A^* whose image in G is 1_G and such that every computation of length k of every ball C in $\mathcal{U}_K$ is contained in every computation of C labelled w_k.*

Proof. Every word whose image in G is 1_G labels a circuit in every ball of $\mathcal{U}_K$ and for each source vertex. For each ball, and each vertex of this ball, we construct a circuit which visits every computation of length k of this ball. The product of the labels of all these circuits is a word w_k that answers the question. ∎

We now turn to the proof of the theorem itself.

The automaton $\mathcal{B}$, an automaton of minimal loop complexity which recognises K, has n states. Let g be in P, a final state of $\mathcal{A}_K$, and u_g be a word in A^* that is sent to g by α. For every integer k, the word $(w_k)^n u_g$ is in K and is hence accepted by $\mathcal{B}$. The Block Star Lemma, applied to the factors w_k, ensures that there exists a state p_k of $\mathcal{B}$ such that there exists a circuit with source p_k labelled by a certain power (w_k^l). Let $\mathcal{D}_k$ be the ball in $\mathcal{B}$ which contains p_k, and hence this circuit. We thus obtain an infinite sequence of balls $\mathcal{D}_k$ in which at least one ball $\mathcal{D}$ in $\mathcal{B}$ appears infinitely often.

Let $\mathcal{C}$ be the ball in $\mathcal{U}_K$ which contains the image of $\mathcal{D}$ under the morphism $\varphi\colon \mathcal{B} \to \mathcal{U}_K$. For every path c in $\mathcal{C}$, there exists a k greater than the length of c, an integer l and a state p of $\mathcal{D}$ such that there exists a loop in $\mathcal{D}$ with source p labelled $(w_k)^l$. This same word $(w_k)^l$ labels a loop in $\mathcal{C}$ which contains all the computations of length less than or equal to k; it thus contains c in particular. That is, c is the image of a computation of $\mathcal{D}$, hence on one hand, $\mathcal{C}$ is the image of $\mathcal{D}$ under φ and on the other, the restriction of φ to $\mathcal{D}$ is *conformal*. By Theorem 8.2, $\mathsf{lc}(\mathcal{D}) \geqslant \mathsf{lc}(\mathcal{C})$.

Let (L, R) be the factorisation, which is the image of p under φ (the state p that was defined just above). Since $(w_k)^{l'}$ is in $\mathsf{Past}_{\mathcal{B}}(p)$, 1_G is in $\mathsf{Past}_{\mathcal{U}_K}((L, R))$ and hence 1_G is in L; that is, (L, R) is an initial state of $\mathcal{U}_K$. Likewise, $(w_k)^{l''} u_g$ is in $\mathsf{Fut}_{\mathcal{B}}(p)$ and g is in R. Every word u of A^* such that $u\alpha = g$ labels a computation of $\mathcal{C}$ with source (L, R) and destination $(L\,g, g^{-1}R)$, a final state of $\mathcal{U}_K$, since $1_G \in g^{-1}R$. Hence u is accepted by $\mathcal{C}$.

cf. Exer. 3.2

We can repeat this construction for each g in P and finally obtain a set of balls of $\mathcal{U}_K$ that recognise all of K and each of which has complexity less than or equal to at least one ball in $\mathcal{B}$. The complexity of the set is at most equal to that of $\mathcal{B}$, which was assumed to be minimal. ∎

The construction performed in this proof can be both deepened and generalised. The structure of the universal automaton of a group language can be described precisely using techniques introduced in Subsection 4.3. This analysis can then be extended to the universal automaton of a *reversible language*[77] and leads to the following generalisation of Theorem 8.1 (see the bibliographic notes for complete references.) *cf. p. 286*

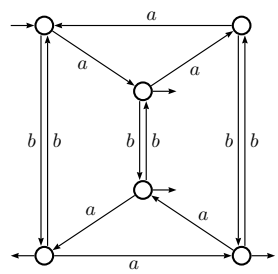
Theorem 8.3 [**Lombardy & Sakarovitch**] *The universal automaton of a reversible language K contains a (quasi-reversible) sub-automaton of minimal loop complexity which recognises K.* ■

Exercises

• 8.1 Give an example of a non-conformal (but surjective) morphism from one ball to another which has the same loop complexity.

• 8.2 Show that the minimal automaton of L has the least loop complexity of all the deterministic automata that recognise a language L over A^*.

8.3 Compute the star height of the language recognised by the adjacent automaton.



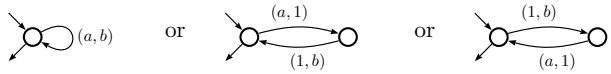
[77] A language is reversible if it is recognised by an automaton in which the letters induce partial bijections between the states. *Note added in translation:* A language is *quasi-reversible* if it is recognised by an automaton in which the transitions that contradict the property that "the letters induce partial bijections between the states" do not belong to a strongly connected component.

SOLUTIONS TO THE EXERCISES

Nicht Kunst und Wissenschaft allein,
Geduld will bei dem Werke sein.
Goethe

1. Automata and rational sets

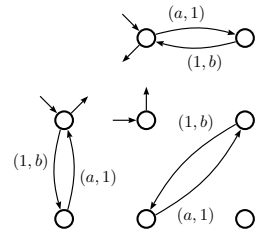
1.1 *Let* $\mathbb{N}^2 = \{a\}^* \times \{b\}^*$ *. Show that the set* $(a,b)^* = \{(a^n, b^n) \mid n \in \mathbb{N}\}$ *is accepted by a finite automaton over* $\mathbb{N}^2$*.*

Solution: Three automata: 

1.2 *Demonstrate with an example that the construction of the cartesian product used in Chapter I to accept the intersection of behaviours of two automata does not work for automata over arbitrary (that is, non-free) monoids. Observe that the same remark holds for automata with spontaneous transitions.*

Solution: The construction applied opposite to two automata which both recognise the subset $(a,b)^*$ of $\mathbb{N}^2$ gives an automaton whose behaviour is empty (and we are in the special case where the minimal generating set of the monoid is unique and where the transition labels are taken from this generating set).

Likewise, the product of two automata with spontaneous transitions below gives an automaton whose behaviour is empty.



1.5 *Let* N *be a submonoid of* M*. Show that* $P \in \operatorname{Rat} M$ *and* $P \subseteq N$ *do not imply necessarily that* $P \in \operatorname{Rat} N$ *.*

Solution: Take $M = \{a,b\}^*$, and $P = N = (b^*a)^*$. Then N is a submonoid of M but is not finitely generated; hence $N \notin \operatorname{Rat} N$.[78]

1.6 *Show that if* R *is a rational set of a monoid* M *without divisors of the identity, then so is* $R \setminus 1_M$ *.*

Solution: By induction on the depth of the rational expression which denotes R:

$$\begin{aligned}
\left|(\mathsf{E}+\mathsf{F})\right| \setminus 1_M &= (\left|\mathsf{E}\right| \setminus 1_M) \cup (\left|\mathsf{F}\right| \setminus 1_M) \quad , \\
\left|(\mathsf{E}\cdot\mathsf{F})\right| \setminus 1_M &= (\left|\mathsf{E}\right| \setminus 1_M) \cdot (\left|\mathsf{F}\right| \setminus 1_M) \cup (\left|\mathsf{E}\right| \setminus 1_M) \cdot \delta_{1,\left|\mathsf{F}\right|} \cup \delta_{1,\left|\mathsf{E}\right|} \cdot (\left|\mathsf{F}\right| \setminus 1_M) \quad , \\
\left|(\mathsf{E}^*)\right| \setminus 1_M &= (\{\left|\mathsf{E}\right| \setminus 1_M\}^*) \cdot (\left|\mathsf{E}\right| \setminus 1_M) \quad ,
\end{aligned}$$

the two last equations holding only when M has no divisors of the identity.

1.7 *Give an example of a (finitely generated) monoid whose rational subsets are closed under intersection but not under complement.*

Solution: If A is infinite, $\operatorname{Rat} A^*$ is closed under intersection but not under complement; but the solution is less obvious if we want to satisfy the property with a finitely generated monoid.

Let $A = \{a,b,c\}$ and M be the quotient of A^* by all the relations $w\,w = a\,a$ for all w in A^*. We know that there exist an infinite number of words without a square factor over

See Th. 0.3.1, p. 23

[78] *Note added in translation:* This example thus shows that the rational submonoids of a free monoid do not have the Fatou property (see p. 303 and p. 520).

the alphabet of three letters, hence M is *infinite*. The complement of $a\,a$ in M is infinite and contains no element that itself contains a square factor: this subset cannot be rational. Furthermore, it is immediate that $\operatorname{Rat} M$ is closed under intersection.

1.8 ***Local languages.*** *A language L over A^* is local if it can be written*

$$L = (DA^* \cap A^*F) \setminus (A^*VA^*) \text{ , with } D, F \subseteq A \text{ and } V \subseteq A^2 \text{ .}$$

(a) *Let $A = \{a, b\}$. Show that the following languages are local:*

$$a^+ \text{ , } \{ab\}^+ \text{ , } a^+ \cup b^+ \text{ , } \{a\} \text{ , } \{ab\} \text{ .}$$

Show that the following languages are not local: $a^+ \cup \{ab\}^+$, $a^+\{ab\}^+$, $\{aa\}$.

(b) *Show that the family of local languages is closed under intersection and star. Show that it is not closed under union or product.*

(c) *Show that the set of factors of a local language is a local language.*

(d) *Show that a language L over A^* is local if and only if for all f, g, h and k in A^* and all a in A, we have that if $f\,a\,g$ and $h\,a\,k$ are in L then $f\,a\,k\,s$ in L.*

(e) *Prove Proposition 1.9:* A language is rational if and only if it is the image of a local language under a (strictly alphabetic) morphism. *cf. p. 228*

Solution [partial]: (a)

$$a^+ = (\{a\}A^* \cup A^*\{a\}) \setminus A^*\{a\,b\}A^* \text{ ; } \quad \{a\,b\}^+ = (\{a\}A^* \cup A^*\{b\}) \setminus A^*\{a\,a, b\,b\}A^* \text{ ;}$$
$$a^+ \cup b^+ = (\{a, b\}A^* \cup A^*\{a, b\}) \setminus A^*\{a\,b, b\,a\}A^* \text{ ; } \quad a = (\{a\}A^* \cup A^*\{a\}) \setminus A^*\{a\,b, a\,a\}A^* \text{ ;}$$
$$\{a\,b\} = (\{a\}A^* \cup A^*\{b\}) \setminus A^*\{a\,a, b\,a, b\,b\}A^* \text{ .}$$

(e) Every local language is rational since it is obtained as the difference of two rational languages.

Conversely, let $R = L(\mathcal{A})$ be a rational language, with $\mathcal{A} = \langle Q, A, E, I, T \rangle$. As in the proof of Theorem 1.1, set $Z = E$. Set *See p. 225*

$$D = \{(i, a, q) \in Z \mid i \in I\} \text{ , } \quad F = \{(p, a, t) \in Z \mid t \in T\} \text{ ,}$$
$$V = \{(p, a, q)(p', a, q') \in Z^2 \mid q \neq p'\} \text{ ,}$$

then $L = (DZ^* \cap Z^*F) \setminus (Z^*VZ^*)$. The map $\eta\colon Z \to A$, which consists of keeping the label of a transition $(p, a, q)\eta = a$, extends to a strictly alphabetic morphism $\eta\colon Z^* \to A^*$. We verify that $R = L\eta$.

1.12 *Prove that in $M = \{a, b\}^* \times \{c\}^*$, the subset $V_1 \cup W_1 = (a, c)^*(b, 1)^* \cup (a, 1)^*(b, c)^*$ is inherently ambiguous.*

Solution: We set $A = \{a, b, c\}$ and $\alpha\colon A^* \to M$ to be the (surjective) canonical morphism. Assume, proceeding by *reductio ad absurdum*, that $V_1 \cup W_1$ is unambiguous and let K be a rational of A^* such that $K\alpha = V_1 \cup W_1$ and α is injective over K (Proposition 1.12 (ii)). *See p. 231*

Let N be the number of states of the automaton (over A) that accepts K and let u be in A^* such that $u\alpha = (a^N b^{N+N!}, c^N)$. As α respects the order of a's and b's, there exists a factorisation $u = v\,w$ such that $|v|_b = 0$ and $|v_a| = N$, and also $|w|_c \leqslant N$. From the Star Lemma, we obtain a factorisation $v = x\,y\,z$ where y is an iterating factor of u. We have $|y|_a = |y|_c = p \leqslant N$ and p is a divisor of $N!$ (that is, $p\,k = N$). Thus, $t = x\,y^{k+1}\,z\,w$ is in K and $t\alpha = (a^{N+N!}b^{N+N!}, c^{N+N!})$.

We apply the symmetric reasoning to the word u' of K such that $u'\alpha = (a^{N+N!}b^N, c^N)$ and we obtain $t' = v'\,x'\,y'^{k'+1}\,z'$ in K such that $t'\alpha = (a^{N+N!}b^{N+N!}, c^{N+N!})$. The two words t and t', which we shall write $t = r\,w$ and $t' = v'\,r'$, are distinct since this double factorisation is entirely determined by the equalities $|r|_b = |v'|_b = 0$, $|w|_b = |r'|_b = N + N!$ and by our having $|w|_c \leqslant N$ and $|r'|_c \geqslant N!$. This contradicts the hypothesis that α is injective over K.

cf. Rem. I.1.11, p. 72

Remark SE.1.1 This proof is representative of a general method in automata theory: a judicious choice of particular words and the application of the Star Lemma to reach a contradiction. We shall return to the same result in Chapter V where we shall be able to give it a more structural interpretation. □

cf. Sec. V.2.4, p. 677

1.13 ***Rationally enumerable monoids.*** *Let M be a monoid. Let $\alpha\colon A^* \to M$ be a surjective morphism; we will say that a set R of A^* is a* cross-section *(of A^*) for α if α is injective over R and $R\alpha = M$: that is if R is a set of representatives for the equivalence $\alpha\,\alpha^{-1}$.*

(a) *Show that if there exists a rational cross-section for α, then there exists a rational cross-section for every surjective morphism $\beta\colon B^* \to M$. We will say in this case that M is* rationally enumerable.

(b) *Show that M is rationally enumerable if and only if $M \in \mathrm{URat}\, M$.*

Solution: (a) Let $\alpha\colon A^* \to M$ and $\beta\colon B^* \to M$ be two surjective morphisms: there exists[79] a morphism $\theta\colon A^* \to B^*$ such that $\alpha = \theta\,\beta$.

$$\begin{tikzcd} & A^* \arrow[dl, dashed, "\theta"'] \arrow[d, "\alpha"] \\ B^* \arrow[r, "\beta"'] & M \end{tikzcd}$$

If R is a cross-section for α, then $S = R\theta$ is a cross-section for β. In fact, $S\beta = (R\theta)\beta = R\alpha = M$ and if α is injective over R, θ is also injective over R and β is injective over $R\theta = S$.

(b) If M is rationally enumerable, there exists a surjective morphism $\alpha\colon A^* \to M$ and $R \in \mathrm{Rat}\, A^* = \mathrm{URat}\, A^*$ such that $R\alpha = M$ and α is injective over R: $M \in \mathrm{URat}\, M$ (Proposition 1.12 (i)).

See p. 231

For every monoid M, there exists an alphabet A and a surjective morphism $\alpha\colon A^* \to M$. If $M \in \mathrm{URat}\, M$, there exists an R in $\mathrm{URat}\, A^*$ such that $R\alpha = M$ (Proposition 1.12 (ii)): M is rationally enumerable.

2. Actions and recognisable sets

2.1 *Let $\delta\colon S \times M \to S$ and $\tau\colon M \to \mathfrak{T}_S$ be two maps such that*

$$\forall s \in S\,,\ \forall m \in M \qquad (s)[m\tau] = (s,m)\delta \ .$$

Verify that δ is an action if and only if τ is a morphism.

Solution: For all s in S and all m and m' in M, the series of equalities

$$((s)[m\tau])[m'\tau] = ((s,m)\delta, m')\delta = (s, mm')\delta = (s)[(mm')\tau]$$

shows that τ is a morphism when δ is an action and, conversely, the series

$$((s,m)\delta, m')\delta = ((s)[m\tau])[m'\tau] = (s)[(mm')\tau] = (s, mm')\delta$$

shows that δ is an action when τ is a morphism.

2.2 *Verify that the map δ_2 given in Example 2.4 really does correspond to an action.*

Solution: $M = A^* \times B^*$ is isomorphic to the quotient of $\big[\big(\bigcup_{a\in A}(a,1)\big) \cup \big(\bigcup_{b\in B}(1,b)\big)\big]^*$ by the congruence generated by the set of pairs $\{(a,1)(1,b) = (1,b)(a,1) \mid a \in A\,,\ b \in B\}$. We then verify

$$\begin{aligned} p\cdot(a,1)(1,b) &= q\cdot(1,b) = s \quad\text{and}\quad p\cdot(1,b)(a,1) = r\cdot(a,1) = s\ ,\\ q\cdot(a,1)(1,b) &= q\cdot(1,b) = s \quad\text{and}\quad q\cdot(1,b)(a,1) = s\cdot(a,1) = s\ ,\\ r\cdot(a,1)(1,b) &= s\cdot(1,b) = s \quad\text{and}\quad r\cdot(1,b)(a,1) = r\cdot(a,1) = s\ ,\\ s\cdot(a,1)(1,b) &= s\cdot(1,b) = s \quad\text{and}\quad s\cdot(1,b)(a,1) = s\cdot(a,1) = s\ . \end{aligned}$$

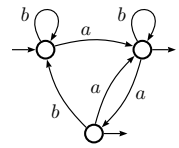
The conclusion is then a consequence of the solution to Exercise 2.3.

[79] *Cf.* Lem. 0.4.1, p. 25.

2.3 *Let A be an alphabet (not necessarily finite). Let R be a set (not necessarily finite) of pairs (u,v) of words from A^*, and M the quotient of A^* by the congruence generated by R. Let S be an arbitrary set and δ a map from $S\times A$ to S . Show that δ defines an action of M on S if for all s in S and every pair (u,v) in R we have $(s,u)\delta=(s,v)\delta$.*

Solution: Let $\alpha\colon A^*\to A^*/[R]=M$ be the canonical morphism. By Proposition 0.2.1, and for all f and g in A^*, $f\alpha=g\alpha$ if and only if $f \overset{\star}{\underset{R}{\longmapsto}} g$. This last relation taken with the hypothesis implies that, for all s in S, $(s,f)\delta=(s,g)\delta$. *See p. 17*

2.5 *Compute the transition monoid defined by the deterministic automaton in the adjacent figure.*



Solution: The corresponding morphism τ_2 from A^* to $\mathbb{B}^{3\times3}$ is defined by the two matrices

$$a\tau_2=\begin{pmatrix}0&1&0\\0&0&1\\0&1&0\end{pmatrix}=x \qquad\text{and}\qquad b\tau_2=\begin{pmatrix}1&0&0\\0&1&0\\1&0&0\end{pmatrix}=y\ .$$

We verify that $x^3=x$, $y^2=y$, $yx=x$ and $x^2y=y$, from which we obtain the multiplication table of the monoid.

1	x	y	x^2	xy
x	x^2	xy	x	y
y	x	y	x^2	xy
x^2	x	y	x^2	xy
xy	x^2	xy	x	y

2.6 *Let P be a subset recognised by a (right) action of M on S. Describe a* left *action of M on a set T which also recognises P and such that T is finite when S is.*

Solution: Let $\tau\colon M\to\mathfrak{T}_S$ be the morphism defined by δ (Equation (2.2)). This morphism defines in turn a *left action* λ of M on $\mathfrak{T}_S$: $\forall t\in\mathfrak{T}_S$, $\forall m\in M$ $(m,t)\lambda=(m\tau)t$. If P is recognised by δ, $P=P\tau\tau^{-1}$ whence $P=\{m\in M \mid (m,1_{\mathfrak{T}_S})\lambda\in P\tau\}$ is also recognised by λ. *cf. Propty 2.2, p. 235*

2.7 ***'Lead or Delay' action.*** *Let B be an alphabet and H_B the set defined by $H_B=(B^*\times1_{B^*})\cup(1_{B^*}\times B^*)\cup\{\mathbf{0}\}$. We define a map ψ from $B^*\times B^*$ to H_B by*

$$\forall f,g\in B^*\qquad (f,g)\psi=\begin{cases}(g^{-1}f,1_{B^*}) & \text{if}\quad g\leqslant f\\ (1_{B^*},f^{-1}g) & \text{if}\quad f\leqslant g\\ \mathbf{0} & \text{otherwise}\end{cases}$$

then a map ω_B from $H_B\times(B^*\times B^*)$ to H_B by

$$\forall f,g,u,v\in B^*\qquad\begin{cases}((f,1_{B^*}),(u,v))\,\omega_B &=& (f\,u,v)\psi\\ ((1_{B^*},g),(u,v))\,\omega_B &=& (u,g\,v)\psi\\ (\mathbf{0},(u,v))\,\omega_B &=& \mathbf{0}\end{cases}$$

(a) *Verify that ω_B is an action of $B^*\times B^*$ on H_B. Interpret – that is, justify – the name given to this action.*

(b) *What is the transition monoid of this action if $\|B\|\geqslant2$? if $\|B\|=1$?*

Solution: (a) A case by case verification, which is long-winded but elementary, gives

$$\forall u,v,x,y\in B^*\qquad(((f,1_{B^*}),(u,v))\,\omega_B,(x,y))\,\omega_B=((f,1_{B^*}),(u\,x,v\,y))\,\omega_B\ ,$$

remembering that $(f^{-1}g)^{-1}h=g^{-1}(f\,h)$ if $|f|\leqslant|g|$. The notions introduced in Section 6 of this chapter will enable us to give a more elegant proof (*cf.* Exer. 6.16, p. 323 and p. 367). *cf. Exer. I.1.14, p. 65*

In the set $B^*\times B^*$, we view an element $(f,1_{B^*})$ as a lead, and an element $(1_{B^*},g)$ as a delay, of the first component with respect to the second. When this lead or delay is multiplied

by a pair (u,v), we consider the resulting pair $(f\,u,v)$, or $(u,g\,v)$. If one of the components is a prefix of the other, the (left) quotient of the one by the other is the lead or delay resulting from this multiplication by (u,v). Otherwise, the two components are 'incompatible' and no multiplication by other elements of $B^*\times B^*$ could make them 'compatible', a situation which it is reasonable to represent by a zero.

(b) For all $(u,v)\in B^*\times B^*$ and every a and b in B, we have

$$(v,1_{B^*})\cdot(u,v)=(u,1_{B^*})\ ,\quad (1_{B^*},u)\cdot(u,v)=(1_{B^*},v)\ ,$$
$$(v\,a,1_{B^*})\cdot(u,v)=(a\,u,1_{B^*})\quad\text{and}\quad (v\,b,1_{B^*})\cdot(u,v)=(b\,u,1_{B^*})\ .$$

If $(x,y)\neq(u,v)$ is such that $(v,1_{B^*})\cdot(x,y)=(u,1_{B^*})$, this implies that $x=w\,u$ and $y=v\,w$ with $|w|\geqslant 1$. Hence $w=a\,w'$ and $(v\,b,1_{B^*})\cdot(x,y)=\mathbf{0}$.

Thus, ω_B induces an injective map from $B^*\times B^*$ to $\mathcal{T}_{H_B}$ and its transition monoid is isomorphic to $B^*\times B^*$.

This proof assumes that there are two distinct letters in B. If $\|B\|=1$, B^* is isomorphic to $\mathbb{N}$ and the transition monoid of ω_B is isomorphic to $\mathbb{Z}$.

2.8 *Compute the transition monoid of the automaton $\mathcal{A}_1$.*

Solution: It is the quotient of $\{a,b\}^*$ by $a^2=a$, $b^2=b$, and $(a\,b)^2=a\,b$, a monoid with 7 elements.

2.9 (a) *Let $\mathcal{A}$ be a deterministic automaton with n states. What is the maximum size of its transition monoid? Can this bound be reached?*

(b) *Do the same for an automaton $\mathcal{A}$ that is not necessarily deterministic.*

Solution [partial]: (a) If $\mathcal{A}=\langle Q,A,E,i,T\rangle$, its transition monoid is a submonoid of the monoid of functional relations from Q to itself, which has n^{n+1} elements. If $\mathcal{A}$ is *complete*, its transition monoid is a submonoid of $\mathfrak{T}_Q$, the monoid of maps from Q to itself (defined everywhere), which has n^n elements.

Let $A=\{a,b,c,d\}$ and define the actions of each letter in the following manner: a is a circular permutation on Q, b is a transposition – that is, two elements of Q are swapped and the others are left unpermuted – c is equal to a except that one element is left unpermuted, and d is equal to c except that its action on the element left unpermuted by c is undefined. These four actions generate the monoid of functional relations from Q to itself, and $\{a,b,c\}$ generates $\mathfrak{T}_Q$.

cf. Exer. I.8.1, p. 175 and p. 206.

2.11 *Show that every rational language is recognised by a left regular deterministic automaton – that is, by an automaton $\mathcal{A}=\langle Q,A,\delta,i,T\rangle$ such that:*

$$\forall f,g\in A^*\qquad i\cdot f=i\cdot g\quad\Longrightarrow\quad[\,\forall u\in A^*\quad i\cdot u\,f=i\cdot u\,g\,]\ .$$

cf. Rem. 2.4, p. 241.

Solution: Let N be a monoid that recognises L and $\mathcal{N}=\langle N,A,E,\{1_N\},S\rangle$ the automaton defined for the proof of Proposition 2.4. This automaton is deterministic, and left regular.

2.13 *Show that:*

(a) *The complement of a subset of M recognised by $\alpha\colon M\to N$ is also recognised by α.*

(b) *Let P and P' be two subsets of M, recognised respectively by $\alpha\colon M\to N$ and $\alpha'\colon M\to N'$. Then the morphism $\alpha\times\alpha'\colon M\to N\times N'$ defined by $(m)\alpha\times\alpha'=(m\alpha,m\alpha')$ recognises $P\cap P'$ (and $P\cup P'$).*

(c) *Let Q be a subset N recognised by $\beta\colon N\to S$; then $P=Q\alpha^{-1}$ is recognised by $\alpha\,\beta$.*

Solution: (a) If P is recognised by α, P is saturated by the congruence $\alpha\alpha^{-1}$ and likewise for the *complement* of P.

(b) We first verify that the given definition indeed makes $\alpha\times\alpha'$ a morphism:

$$(m\,m')\alpha\times\alpha'=(m\,m'\alpha,m\,m'\alpha')=(m\alpha,m\alpha')\,(m'\alpha,m'\alpha')=(m\alpha\times\alpha')\,(m'\alpha\times\alpha')\ .$$

Let $Q = P\alpha$ and $Q' = P'\alpha'$ (and hence $P = Q\alpha^{-1}$ and $P' = Q'\alpha'^{-1}$). We then have

$$(Q\times Q')(\alpha\times\alpha')^{-1} = \{m \in M \mid m\alpha \in Q \quad \text{and} \quad m\alpha' \in Q'\}$$
$$= \{m \in M \mid m \in P \quad \text{and} \quad m \in P'\} = P \cap P' \ .$$

Similarly, $P \cup P' = [(Q\times N') \cup (N\times Q')](\alpha\times\alpha')^{-1}$.

(c) By hypothesis, $Q = Q\beta\beta^{-1}$ hence $P = Q\alpha^{-1} = Q\beta\beta^{-1}\alpha^{-1} = (Q\beta)(\alpha\circ\beta)^{-1}$.

2.14 *Let P be a subset of M recognised by an action δ of M on S. Show that for every subset K of M, PK^{-1} is recognised by δ.*

Solution: For s in S and m in M set $s\cdot m^{-1} = \{t \in S \mid t\cdot m = s\}$. Hence, by additivity, for R contained in S and K contained in M

$$R\cdot m^{-1} = \bigcup_{r\in R} r\cdot m^{-1} \qquad \text{and} \qquad R\cdot K^{-1} = \bigcup_{m\in K} R\cdot m^{-1} = \{t \in S \mid t\cdot m \in R\} \ .$$

By definition, there exists Q in S such that $P = \{n \in M \mid s_0\cdot n \in Q\}$, hence

$$P\,m^{-1} = \{n\,m \in M \mid n\,m \in P\} = \{n \in M \mid s_0\cdot n\,m \in Q\} = \{n \in M \mid s_0\cdot n \in Q\cdot m^{-1}\} \ ,$$

and $PK^{-1} = \{n \in M \mid s_0\cdot n \in Q\cdot K^{-1}\}$ is recognised by δ.

2.15 *Show that the definition of the left quotient (resp. right quotient) of in a monoid coincides in a group with multiplication on the left (resp. on the right) by the 'normal' inverse.*

Solution: Let G be a group. Write, for this exercise, $\bar{g}$ for the inverse of the element g of G. Let g and h be in G; by definition $g^{-1}\{h\} = \{k \mid g\,k = h\}$. Since we are in a group, $g\,k = h$ implies $k = \bar{g}\,h$ and thus $g^{-1}\{h\} = \bar{g}\,h$. Hence, for all subsets P and Q of G, $g^{-1}P = \bar{g}\,P$ and $Q^{-1}P = \bar{Q}\,P$ with $\bar{Q} = \bigcup_{g\in Q}\bar{g}$.

2.16 *Let U be the monoid $\{u, x, z\}$ whose neutral element is u, and z the zero, with x such that $x\,x = z$ and $M = \{(u,0),(x,0)\} \cup z\times\mathbb{Z}$ be a submonoid of $U\times\mathbb{Z}$.*

Show that the singleton $(x,0)$ is recognisable in M but that neither $(x,0)^2$, nor $(x,0)^*$, nor $(x,0)^+$ are.[80]

Solution: The projection (omitting the second component) is a morphism from M to U which recognises $(u,0)$, $(x,0)$ and $z\times\mathbb{Z}$.

Furthermore, $z\times\mathbb{Z}$ is a submonoid of M, isomorphic to $\mathbb{Z}$. For every subset X of M,

$$X = [X \cap \{(u,0),(x,0)\}] \bigcup [X \cap z\times\mathbb{Z}]$$

and hence X is recognisable if and only if $X \cap z\times\mathbb{Z}$ is recognisable. Thus, $(x,0)^2 = (z,0)$ is not recognisable since 0 is not recognisable in $\mathbb{Z}$. *If necessary, cf. Sec. 6.1, p. 301*

Likewise, $(x,0)^*$ and $(x,0)^+$ are not recognisable since $(x,0)^* \cap z\times\mathbb{Z} = (x,0)^+ \cap z\times\mathbb{Z} = (z,0)$.

2.17 *Let M be an equidivisible monoid, m be in M and P and Q be two (arbitrary) subsets of M. Prove:*[81]

$$m^{-1}(P\,Q) = (m^{-1}P)\,Q \cup (P^{-1}m)^{-1}Q \ , \qquad \text{and} \qquad m^{-1}P^+ = \left[(P^{*-1}m)^{-1}P\right] P^* \ .$$

Solution: We have:

$$m^{-1}(P\,Q) = \{s \mid m\,s \in P\,Q\} = \{s \mid \exists p \in P, \exists q \in Q \quad m\,s = p\,q\} \ ,$$

[80] Essentially Exer. III.12.9 of [80], which Eilenberg attributes to S. Winograd.

[81] *Cf.* Exer. III.12.10 of [80].

hence, since M is equidivisible

$$\begin{aligned} m^{-1}(PQ) &= \{s \mid \exists r, \exists p \in P, \exists q \in Q \quad p = m\,r \quad \text{and} \quad s = r\,q\} \\ &\qquad\qquad \cup \{s \mid \exists r, \exists p \in P \quad p\,r = m \quad \text{and} \quad r\,s \in Q\} \\ &= (m^{-1}P)\,Q \cup (P^{-1}m)^{-1}Q \ , \end{aligned}$$

and likewise

$$\begin{aligned} m^{-1}P^+ &= \{s \mid \exists n \in \mathbb{N}_+ \quad m\,s \in P^n\} \\ &= \{s \mid \exists k, l \in \mathbb{N}, \exists r \in (P^k)^{-1}m \quad s \in (r^{-1}P)\,P^l\} \\ &= \left[(P^{*-1}m)^{-1}P\right] P^* \ . \end{aligned}$$

see p. 247 **2.19** *Verify Property 2.12:*

(i) *For every action δ of M on S, there exists a (unique) map σ from M to S which is an action morphism from the regular representation of M on itself to δ (and which we call the* canonical morphism *of δ).*

(ii) *Let δ and η be two actions from M to S and T respectively and let σ and τ be their two canonical morphisms. If φ is an action morphism from δ to η, then $\tau = \sigma\varphi$.*

Solution: (i) The map $\sigma\colon M \to S$ is necessarily defined by $m\sigma = s_0 \underset{S}{\cdot} m$, since $s_0 = 1_M\sigma$ and $m\sigma = (1_M \underset{M}{\cdot} m)\sigma$. This is indeed an action morphism since

$$\forall m, n \in M \quad (m \underset{M}{\cdot} n) = s_0 \underset{S}{\cdot} m\,n = (s_0 \underset{S}{\cdot} m) \underset{S}{\cdot} n \,.$$

(ii) Let δ and η be two actions of M on S and T respectively, σ and τ their two canonical morphisms, and φ an action morphism of δ on η. It follows that

$$\forall m \in M \quad (m\sigma)\varphi = (s_0 \underset{S}{\cdot} m)\varphi = s_0\varphi \underset{T}{\cdot} m = t_0 \underset{T}{\cdot} m = m\tau \ .$$

Hence $\tau = \sigma\varphi$.

$$\begin{tikzcd} S \arrow[rr, "\check{m\delta}"] \arrow[dd, "\varphi"'] & & S \arrow[dd, "\varphi"] \\ & M \arrow[ul, "\sigma"'] \arrow[ur, "\sigma"] \arrow[r, "\check{m\iota}"] \arrow[dl, "\tau"] \arrow[dr, "\tau"'] & M \\ T \arrow[rr, "\check{m\eta}"'] & & T \end{tikzcd}$$

see p. 249 **2.20** *Prove that: 'the relation ρ_P is the coarsest right regular equivalence of M that saturates P' (Property 2.14).*

Solution: (i) The relation ρ_P is right regular since it is obviously equal to the map equivalence of the initial morphism on *the action* ρ_P (Property 2.13). More directly[82]

$$\begin{aligned} m\ \rho_P\ m' \quad \Rightarrow \quad m^{-1}P = m'^{-1}P \quad &\Rightarrow \quad \forall t \in M \quad t^{-1}(m^{-1}P) = t^{-1}(m'^{-1}P) \\ &\Rightarrow \forall t \in M \quad (m\,t)^{-1}P = (m'\,t)^{-1}P \quad \Rightarrow \quad \forall t \in M \quad m\,t\ \rho_P\ m'\,t \ . \end{aligned}$$

(ii) P is saturated for ρ_P; suppose $p\ \rho_P\ p'$:

$$p \in P \quad \Rightarrow \quad 1_M \in p^{-1}P = p'^{-1}P \quad \Rightarrow \quad p' \in P \ .$$

(iii) Let α be a right regular equivalent of M which saturates P and m and m' such that $m\ \alpha\ m'$. Suppose that there exists t in $m^{-1}P \setminus m'^{-1}P$; then $m\,t \in P$ and $m'\,t \notin P$.

As α is right regular, $m\ \alpha\ m' \Rightarrow m\,t\ \alpha\ m'\,t$, which contradicts the hypothesis that P is saturated by α. Thus $m\ \alpha\ m' \Rightarrow m\ \rho_P\ m'$, so ρ_P is coarser than α.

[82] And if we get in a muddle with this ρ_P , which denotes two different objects.

2.22 *Prove that 'the relation σ_P is the coarsest congruence on M that saturates P' (Property 2.15).* see p. 249

Solution: (i) σ_P is an equivalence relation by its definition. σ_P saturates P because m is in P if and only if $(1_M, 1_M)$ belongs to $\mathsf{C}_P(m)$.

(ii) Suppose $m\,\sigma_P\,n$ and let t be in M. We have

$$\mathsf{C}_P(m\,t) = \{(u,v) \mid u\,m\,t\,v \in P\} = \{(u,v) \mid (u,t\,v) \in \mathsf{C}_P(m)\} = \mathsf{C}_P(n\,t) \ .$$

Likewise, $\mathsf{C}_P(t\,m) = \mathsf{C}_P(t\,n)$, hence σ_P is a congruence. Exer. 0.2.4, p. 17

(iii) Let α be a congruence of M which saturates P and m and m' such that $m\ \alpha\ m'$. Suppose that there exists (u,v) in $\mathsf{C}_P(m) \setminus \mathsf{C}_P(m')$; then $u\,m\,v \in P$ and $u\,m'\,v \notin P$.

As α is a congruence, $m\ \alpha\ m' \Rightarrow u\,m\,v\ \alpha\ u\,m'\,v$, which contradicts the hypothesis that P is saturated by α. Thus $m\ \alpha\ m' \Rightarrow m\,\sigma_P\,m'$: σ_P is coarser than α.

2.23 *Let P be a subset of M. Show that M/σ_P is the transition monoid of the action ρ_P, that is, of the canonical action of M on M/ρ_P (Proposition 2.13).* see p. 250

Solution: The monoid M acts on the set $M/\rho_P = \{[m]_{\rho_P} \mid m \in M\}$ and we have

$$n \in [m]_{\rho_P} \quad \Longleftrightarrow \quad n^{-1}P = m^{-1}P \ .$$

Two elements p and q of M have the same image in the transition monoid if they have the same action on every element of M/ρ_P; we therefore have the series of equivalences:

$$\begin{aligned} p\alpha = q\alpha \quad &\Longleftrightarrow \quad \forall m \in M \quad [m]_{\rho_P} \cdot p = [m]_{\rho_P} \cdot q \quad \Longleftrightarrow \quad \forall m \in M \quad [m\,p]_{\rho_P} = [m\,q]_{\rho_P} \\ &\Longleftrightarrow \quad \forall m \in M \quad (m\,p)^{-1}P = (m\,q)^{-1}P \\ &\Longleftrightarrow \quad \forall m \in M \quad \{n \mid m\,p\,n \in P\} = \{n \mid m\,q\,n \in P\} \\ &\Longleftrightarrow \quad \forall m,n \in M \quad (m\,p\,n \in P \Leftrightarrow m\,q\,n \in P) \quad \Longleftrightarrow \quad \mathsf{C}_P(p) = \mathsf{C}_P(q) \end{aligned}$$

2.24 ***Disjunctive subset.*** (a) *Verify that for all u and v in M and all subsets P of M, $(u^{-1}P)\,v^{-1} = u^{-1}(P\,v^{-1})$ (which we will therefore write $u^{-1}P\,v^{-1}$ and call the* bi-quotient*).*

(b) *Show that P is disjunctive if and only if the Boolean algebra generated by the bi-quotients of P is the whole of $\mathfrak{P}(M)$.*

(c) *Give an example of a monoid which contains no disjunctive subset.*

Solution: (a) We have $(u^{-1}P)\,v^{-1} = \{m \in M \mid m\,v \in u^{-1}P\} = \{m \in M \mid u\,m\,v \in P\}$, which is a left–right symmetric expression.

(b) This property is none other than a reformulation of the characterisation of the syntactic congruence by contexts: let m and n be in M; it follows that

$$n \in \bigcap_{u\,m\,v \in P} u^{-1}P\,v^{-1} \quad \Longleftrightarrow$$
$$\forall u,v \in M \quad u\,m\,v \in P \Rightarrow u\,n\,v \in P \qquad \text{that is,} \qquad \mathsf{C}_P(m) \subseteq \mathsf{C}_P(n) \ ,$$

$$n \in \bigcap_{u\,m\,v \notin P} \overline{u^{-1}P\,v^{-1}} \quad \Longleftrightarrow$$
$$\forall u,v \in M \quad u\,m\,v \notin P \Rightarrow u\,n\,v \notin P \qquad \text{that is,} \qquad \mathsf{C}_P(n) \subseteq \mathsf{C}_P(m) \ .$$

If P is disjunctive, σ_P is the identity and hence $\mathsf{C}_P(m) = \mathsf{C}_P(n)$ implies $m = n$; that is

$$\left(\bigcap_{u\,m\,v \in P} u^{-1}P\,v^{-1}\right) \bigcap \left(\bigcap_{u\,m\,v \notin P} \overline{u^{-1}P\,v^{-1}}\right) = \{m\} \ , \qquad \text{(SE.2.1)}$$

and the atoms of the Boolean algebra generated by the bi-quotients are the elements of M.

Conversely, if the bi-quotients of P generate $\mathfrak{P}(M)$, each m is completely defined by the bi-quotients that contain it: that is, Equation (SE.2.1) is satisfied and σ_P is the identity.

(c) The monoid $U_3 = \{1, x, y, z\}$, where elements different from 1 are right zeros – that is, $x\,x = y\,x = z\,x = x$, *etc.*– answers the question.

2.26 *Write $\mathcal{I}(L)$ for the set of iterating factors of L. Show that $\mathcal{I}(L)$ is rational when L is (Proposition I.8.1).*

see p. 177

cf. the solution in Chap. I p. 208

Solution: Let $\alpha\colon A^* \to M$ be a morphism on a finite M that recognises L: $L = P\alpha^{-1}$ for $P \subseteq M$. The subset $\mathcal{I}(P)$ of M defined by $\mathcal{I}(P) = \{m \in M \mid \exists p, q \in M \quad p\,m^*q \subseteq P\}$ is such that $\mathcal{I}(L) = [\mathcal{I}(P)]\,\alpha^{-1}$.

2.29 [83]*Let M be an equidivisible monoid.*

(a) *Show that* $\operatorname{Rec} M$ *is closed under product.*

(b) *Show that* $\operatorname{Rec} M$ *is closed under* $+$ *; that is,* $\quad P \in \operatorname{Rec} M \;\Rightarrow\; P^+ \in \operatorname{Rec} M$.

Solution: (i) First method (for (a) only).

Let P and Q be two recognisable subsets of M, recognised respectively by $\alpha\colon M \to N$ and $\beta\colon M \to V$ (N and V are finite).

Since M is equidivisible, the map $\sigma\colon M \to (\mathfrak{P}(M)\times\mathfrak{P}(M))^{2\times 2}$ defined by

$$\forall m \in M \qquad (m)\sigma = \begin{pmatrix} (m,1) & \{(u,v) \mid u\,v = m\} \\ 0 & (1,m) \end{pmatrix}$$

cf. Exer. 0.6.11, p. 32

is a morphism. Furthermore, the morphism $\alpha\times\beta\colon M\times M \to N\times V$ extends to a morphism (of semirings) $\alpha\times\beta\colon (\mathfrak{P}(M)\times\mathfrak{P}(M))^{2\times 2} \to (\mathfrak{P}(N)\times\mathfrak{P}(V))^{2\times 2}$. Thus, $\tau = \sigma\circ\alpha\times\beta$ is a morphism from M to a finite monoid (of matrices).[84]

We can immediately verify that $P\,Q = (Z)\tau^{-1}$ where Z is the set of matrices in $(\mathfrak{P}(N)\times\mathfrak{P}(V))^{2\times 2}$ whose upper right entry is included in $P\alpha\times Q\beta$.

(ii) Second method (for (a) and (b) together).

cf. Exer. 2.17, p. 246

Since M is equidivisible, the quotients of $P\,Q$ and of P^+ by an element m are given by

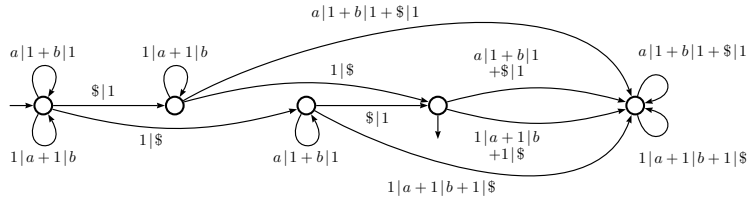
$$m^{-1}(P\,Q) = (m^{-1}P)\,Q \cup (P^{-1}m)^{-1}Q$$
$$m^{-1}P^+ = \left[(P^{*-1}m)^{-1}P\right]\,P^*$$

cf. Th. 2.3, p. 247

If P and Q are recognisable, they have only a finite number of distinct quotients and the above show that $P\,Q$ on one hand and P^+ on the other have only a finite number of quotients and are thus recognisable.

2.30 *Let \$ be a symbol which, by convention, does not belong to any alphabet. For each alphabet A, write $A_\$$ for the set $A \cup \{\$\}$. Let δ_3 be the minimal action of $A_\$^* \times B_\$^*$ which recognises the subset $A^*\$\times B^*\$$. Compute $\mathcal{D}_{\delta_3}^{(C)}$, the automaton of δ_3.*

Solution:



[83] *Cf.* Exer. III.12.10 of [80]; solution by the second method.

[84] *Note added in translation:* The monoid $M\tau$ is the *Schützenberger product* of N and V (*cf.* [81, 185, 187]).

3. Morphisms and coverings

3.1 *Verify that:*

(a) *the composition of two automata morphisms is an automata morphism.*

(b) $\pi_{\mathcal{A}}\colon \mathcal{A}\times\mathcal{B} \to \mathcal{A}$ *is an automata morphism (automata over a free monoid).*

(c) $\varphi\colon \mathcal{B} \to \mathcal{A}$ *is an automata morphism if and only if* $\varphi_{\mathsf{n}}\colon \mathcal{B}_{\mathsf{n}} \to \mathcal{A}_{\mathsf{n}}$ *is a morphism of labelled graphs.*

Solution [partial]: (b) First of all, $\pi_{\mathcal{A}}$ is indeed a morphism of labelled graphs:

$$((p,r),a,(q,s))\pi_{\mathcal{A}}\iota = (p,a,q)\iota = p \quad \text{and} \quad ((p,r),a,(q,s))\iota\pi_{\mathcal{A}} = (p,r)\pi_{\mathcal{A}} = p$$

and on the other hand (p,r) is in $I\times J$ (resp. in $T\times U$) if and only if $p = (p,r)\pi_{\mathcal{A}}$ is in I (resp. in T).

3.2 *Verify that if* $\varphi\colon \mathcal{B} \to \mathcal{A}$ *is a morphism, then for every state* r *of* $\mathcal{B}$ *we have* $\mathsf{Past}_{\mathcal{B}}(r) \subseteq \mathsf{Past}_{\mathcal{A}}(r\varphi)$ *and* $\mathsf{Fut}_{\mathcal{B}}(r) \subseteq \mathsf{Fut}_{\mathcal{A}}(r\varphi)$.

Solution: Let $\mathcal{A} = \langle Q,M,E,I,T\rangle$ and $\mathcal{B} = \langle R,M,F,J,U\rangle$ be two automata over M. We have

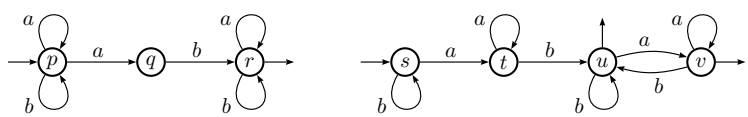
$$\mathsf{Past}_{\mathcal{A}}(q) = \{m \in M \mid \exists i \in I \quad i \xrightarrow[\mathcal{A}]{m} q\} = |\langle Q,M,E,I,\{q\}\rangle| \qquad \text{and}$$
$$\mathsf{Fut}_{\mathcal{A}}(q) = \{m \in M \mid \exists t \in T \quad q \xrightarrow[\mathcal{A}]{m} t\} = |\langle Q,M,E,\{q\},T\rangle|$$

If $\varphi\colon \mathcal{B} \to \mathcal{A}$ is an automata morphism, then φ is also a morphism from $\langle R,M,F,J,\{r\}\rangle$ to $\langle Q,M,E,I,\{r\varphi\}\rangle$ and from $\langle R,M,F,\{r\},U\rangle$ to $\langle Q,M,E,\{r\varphi\},T\rangle$, for all r in R. The two inclusions we are after then follow from Property 3.1. *see p. 256*

3.5 *Verify that in general there is no automata morphism between a (non-deterministic) automaton and an equivalent deterministic automaton.*

Solution: We consider the automaton $\mathcal{A}_1$ and its determinisation (Figure SE.3.1). If there is a morphism φ from $\mathcal{A}_{1\mathsf{det}}$ to $\mathcal{A}_1$, the constraints on the initial and final states imply that $s\varphi = p$, $u\varphi = v\varphi = r$. If $t\varphi = q$, the transition $t \xrightarrow[\mathcal{A}_{1\mathsf{det}}]{a} t$ has no image; if $t\varphi = p$, $t \xrightarrow[\mathcal{A}_{1\mathsf{det}}]{b} u$ has no image; if $t\varphi = r$, then $s \xrightarrow[\mathcal{A}_{1\mathsf{det}}]{a} t$ has no image: contradiction.

We verify that there is no morphism from $\mathcal{A}_1$ to $\mathcal{A}_{1\mathsf{det}}$ in the same way.



(a) The automaton $\mathcal{A}_1$ (b) The automaton $\mathcal{A}_{1\mathsf{det}}$

Figure SE.3.1: Determinisation does not in general produce a morphism

3.12 *Show that a covering (resp. a co-covering) of a deterministic (resp. co-deterministic) automaton is a deterministic (resp. co-deterministic) automaton.*

Solution: It follows from the definition of a deterministic automaton that $\mathcal{A}$ is deterministic if and only if for all p in Q there exists, for each a in A, at most one edge (p,a,q) in $\mathsf{Out}_{\mathcal{A}}(p)$, and $\mathcal{A}$ has a unique initial state i. If $\mathcal{B}$ is a covering of $\mathcal{A}$, there is a bijection between $\mathsf{Out}_{\mathcal{B}}(r)$ and $\mathsf{Out}_{\mathcal{A}}(r\varphi)$ and hence for all r in R there exists, for each a in A, at most one edge (r,a,s) in $\mathsf{Out}_{\mathcal{B}}(r)$, and there is a unique initial state in the inverse image of i under φ: $\mathcal{B}$ is deterministic. The proof of the dual property is analogous.

3.14 *Show that a covering (resp. a co-covering) of an unambiguous automaton is an unambiguous automaton (Corollary 3.8).*

Solution: By definition, $\mathcal{A}$ is unambiguous if and only if for each f in A^* there is at most one successful path in $\mathcal{A}$ with label f. If $\mathcal{B}$ is a covering (or a co-covering) of $\mathcal{A}$, there is a bijection between the successful computations of $\mathcal{B}$ and those of $\mathcal{A}$, and $\mathcal{B}$ is also unambiguous.

3.15 *Show that a covering (resp. a co-covering) of a co-deterministic (resp. deterministic) automaton is an unambiguous automaton.*

Solution: A deterministic (or co-deterministic) automaton is unambiguous and we can apply Corollary 3.8, which we have just proved.

cf. Propty 3.9, p. 268

3.16 *Prove that if $\mathcal{A}$ is an automaton over M and δ is an action of M on a set S, δ is a valuation of $\mathcal{A}\times\delta$ and that if $\big((p,s),t\big)$ is a state of $(\mathcal{A}\times\delta)\times\delta$, then $s=t$.*

Solution: By definition, every state $\big((p,s),t\big)$ of $(\mathcal{A}\times\delta)\times\delta$ is accessible from an initial state $\big((i,s_0),s_0\big)$, and a computation $\big((i,s_0),s_0\big) \xrightarrow[(\mathcal{A}\times\delta)\times\delta]{m} \big((p,s),t\big)$ implies that $s_0\cdot m = t$ and that $(i,s_0) \xrightarrow[\mathcal{A}\times\delta]{m} (p,s)$ is a computation in $\mathcal{A}\times\delta$, and hence that $s_0\cdot m = s$. Thus $s=t$ and δ is a valuation of $\mathcal{A}\times\delta$.

3.17 *Construct the S-co-covering and the S-co-immersions of the automaton $\mathcal{A}_1$.*

Solution: See Figure SE.3.2 and Figure SE.3.3 below.

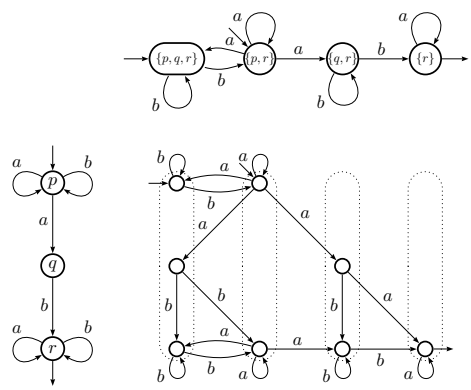


Figure SE.3.2: The S-co-covering of $\mathcal{A}_1$

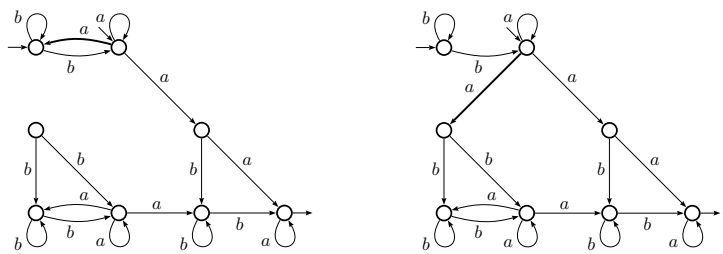


Figure SE.3.3: The two S-co-immersions of $\mathcal{A}_1$

3.18 *Show that the S-covering of a deterministic automaton is isomorphic to that automaton.*

Solution: If $\mathcal{A} = \langle Q, A, E, I, T \rangle$ is deterministic, $\mathcal{A}_{\mathsf{det}} = \mathcal{A}$ and the accessible subset of $\mathcal{A}\times\mathcal{A}$ is formed, again because $\mathcal{A}$ is deterministic, from states (p,p) for p in Q and from edges $\big((p,p), a, (q,q)\big)$ for (p,a,q) in E.

3.19 *Show that property (ii) of Theorem 3.2 is not satisfied in general if the product of an automaton $\mathcal{A}$ is taken with an equivalent deterministic automaton other than $\mathcal{A}_{\mathsf{det}}$.*

Solution: We construct the product of $\mathcal{A}_1$ by the minimal deterministic automaton $\mathcal{B}_1$ of $L(\mathcal{A}_1)$: *cf.* Figure SE.3.4 below. The projection $\pi_{\mathcal{B}_1}$ is not In-surjective in state (u,q) (shown black in the figure).

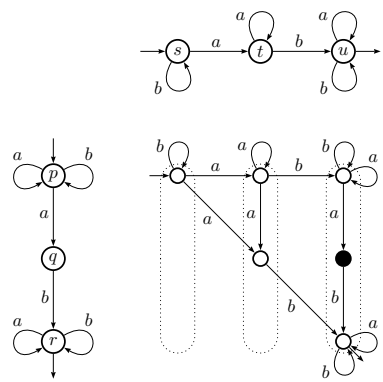


Figure SE.3.4: Another covering of $\mathcal{A}_1$

4. Universal automaton

4.1 *Compute the 2-factorisations of the language $a\,A^*$.*

Solution: $(1_{A^*} \cup a\,A^*, a\,A^*)$, $(a\,A^*, A^*)$ and $(A^*, \emptyset)$.

4.2 *Compute the 2-factorisation of the subset $K_2 = \{x, x^2, x\,y\}$ in the monoid M_2, the quotient of $\{x,y\}^*$ by the relations $x^3 = x$, $y^2 = y$, $y\,x = x$ and $x^2 y = y$.*

Solution: The reader should have recognised in M_2 the transition monoid computed in Exercise 2.5, p. 351.

By judicious arrangement of its table (adjacent), the five 2-factorisations included in K_2 are clearly visible: $(\{x^2\}, \{1, x^2, x, x\,y\})$, $(\{x, x\,y\}, \{1, y, x^2, x\})$, $(\{1, y, x^2\}, \{x^2, x, x\,y\})$, $(\{x^2, x, x\,y\}, \{1, x^2, x\})$ and $(\{1, y, x^2, x, x\,y\}, \{x^2, x\})$.

y	1_M	x^2	x	$x\,y$
y	y	x^2	x	$x\,y$
y	x^2	x^2	x	$x\,y$
$x\,y$	x	x	x^2	y
$x\,y$	$x\,y$	x	x^2	y

4.3 *How many factorisations are there of the subset $K = \{1,3,4,5\}$ in $\mathbb{Z}/6\mathbb{Z}$?* *cf. Exam. 4.2, p. 275*

Solution: 17.

4.4 *Give a necessary and sufficient condition on the subset K for $(\emptyset, M)$ (resp. $(M, \emptyset)$) to be a factorisation of K.*

Solution: The pair $(\emptyset, M)$ is *not* a factorisation of K if and only if K is a *right ideal* of M: that is, $K\,M \subseteq K$; and symmetrically, $(M, \emptyset)$ is *not* a factorisation of K if and only if K is a *left ideal* of M.

cf. p. 253

4.5 *Use the decomposition of a recognisable subset into a finite number of factorisations to prove Prop. 2.19:* if a recognisable subset P of M is contained in the product XY of two subsets X and Y of M, there exists a finite set I of terms $S_i T_i$ with S_i contained in X and T_i included in Y such that

$$P = \bigcup_{i \in I} S_i T_i \ .$$

Furthermore, if X and Y are recognisable, the S_i and T_i are too.

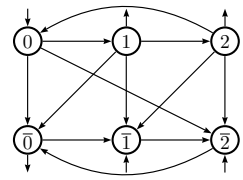
Solution: From $P \subseteq XY$ and $P = \bigcup_{q \in Q_P} L_q R_q$, we deduce $P = \bigcup_{q \in Q_P} (X \cap L_q)\ (Y \cap R_q)$. The inclusion from right to left is obvious. Conversely, each m of P can be written, by hypothesis, $m = xy$, with x in X and y in Y. The sub-2-factorisation (x, y) is dominated by (at least) one factorisation (L_q, R_q) and $m \in (X \cap L_q)\ (Y \cap R_q)$.

cf. Exer. I.2.12, p. 92

4.6 *Let* $\mathsf{Conj}\colon A^* \to A^*$ *be the relation which associates a word* f *with the set of its conjugates:* $\mathsf{Conj}(f) = \{vu \mid u, v \in A^* \quad uv = f\}$. *Show that* $\mathsf{Conj}(K)$ *is rational when* K *is.*

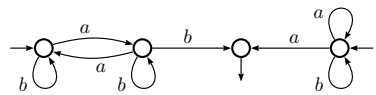
Solution: Every factorisation $f = uv$ of a word f in K belongs to a 2-factorisation (L_q, R_q) of K. Thus $\mathsf{Conj}(K) = \bigcup_{q \in Q_K} R_q L_q$ and the conclusion follows as the R_q and L_q are rational, and Q_K is finite, when K is rational.[85]

4.9 *Compute the* $\{1\}$*-universal automaton of the subset* $\{1, 2\}$ *of* $\mathbb{Z}/3\mathbb{Z}$ *(additive notation).*

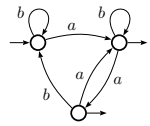


Solution: The factorisations of $\{1, 2\}$ are: $(0, \{1, 2\})$, $(1, \{0, 1\})$, $(2, \{0, 2\})$, $(\{1, 2\}, 0)$, $(\{0, 2\}, 2)$ and $(\{0, 1\}, 1)$. The corresponding states are written respectively 0, 1, 2, $\overline{0}$, $\overline{1}$ and $\overline{2}$.

4.10 *Compute the universal automaton of the language recognised by the adjacent automaton* $\mathcal{L}_1$.

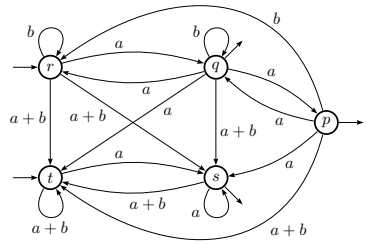


Solution: The determinisation of $\mathcal{L}_1$ is the adjacent automaton (Exercise I.3.5 (g)), whose transition monoid was shown in Exercise 2.5 to be the monoid M_2 of Exercise 4.2 above.



The universal automaton thus has five states that correspond to the five factorisations computed above:

$$p = (\{x^2\}, \{1, x^2, x, xy\}) \ ,$$
$$q = (\{x, xy\}, \{1, y, x^2, x\}) \ ,$$
$$r = (\{1, y, x^2\}, \{x^2, x, xy\}) \ ,$$
$$s = (\{x^2, x, xy\}, \{1, x^2, x\})$$
and
$$t = (\{1, y, x^2, x, xy\}, \{x^2, x\}) \ .$$



The transitions of the $\{a, b\}$-universal automaton, shown in the adjacent figure, are easily determined.

4.14 *Let* K *be a recognisable subset of* M *and* A *a fixed alphabet. Show that for all substitutions* $\sigma\colon A^* \to M$, *there is only a finite number of possible languages* $K\check{\sigma}$.[86]

Solution: The language $K\check{\sigma}$ is determined by the set of the $B_{p,q} = \{a \in A \mid a\sigma \subseteq F_{p,q}\}$ and for each $B_{p,q}$ there is only a finite number (at most $2^{\|A\|}$) of possible choices.

---

[85] *Note added in translation:* moreover, the solution shows that if K is recognised by N, $\mathsf{Conj}(K)$ is recognised by the Schützenberger product of N by itself (*cf.* Exer. 2.29, p. 356, and also [130]).

[86] *Cf.* [66, Cor. 6.3].

4.15 *Find the best approximation to the language* $C_1 = (a^*b + b\,b^*a)^*$ *by the languages* $H_1 = a^2$, $H_2 = b$ *and* $H_3 = b\,a$.

Solution: Set $X = \{x, y, z\}$ and $\sigma_1 \colon X^* \to M$ defined by $x\sigma_1 = H_1$, $y\sigma_1 = H_2$ and $z\sigma_1 = H_3$. We find, having made the calculations, that $C_1\breve{\sigma}_1 = (y + z + x\,(x + z)^*y)^*$ and hence $(C_1\breve{\sigma}_1)\sigma_1 = (b + b\,a + a^2(a^2 + b\,a)^*b)^*$.[87]

4.16 *Give a version of the universal automaton and Theorem 4.5 in terms of rational expressions.*

Solution: Let K be a recognisable subset of M and $\mathcal{U}_K$ its universal automaton. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of variables in bijection with the set of *distinct* factors of K. Replace each factor $F_{p,q}$ in $\mathcal{U}_K$ by the corresponding variable and compute, by any method, a *rational expression* E_K *over* X which denotes the language accepted by this automaton.

This expression E_K can play for K a role analogous to that of $\mathcal{U}_K$. In particular, if $\mathcal{K} = \{K_1, K_2, \ldots, K_l\}$ is a family of subsets of M, the best approximation to K by $\mathcal{K}$ is given by the interpretation $\tau_{\mathcal{K}}$ of E_K where, for each x in X, $(x)\tau_{\mathcal{K}}$ is the union of the K_i which are contained in the factor of K corresponding to x.[88]

4.17 *Find all the languages whose square is included in the language* $(a\,(a + b) + b)^*a$.

Solution: There is a single maximal sub-solution: $a^*b\,[(a\,(a + b) + b)^*a]$.[89]

4.18 *Given a recognisable subset* K *of* M*, find all the subsets* S *such that* $S^3 \subseteq K$ *and that are maximal for this property.*

Solution: As in Example 4.4, every sub-factorisation (S, S, S) is dominated by a factorisation of the form $(L_p, F_{p,q}, R_q)$ (note that such a triple *is not* always a factorisation of a pair (p, q)). All the maximal sub-solutions are hence of the form $L_p \cap F_{p,q} \cap R_q$.

4.19 *What is the relationship between the maximal solutions of* $X^2 \subseteq H$ *and the* square root $\mathcal{R}_2(H)$ *computed in Exercise I.8.1 and in § 2.5.1?* *cf. p. 175 and p. 251*

Solution: $\mathcal{R}_2(H) = \bigcup\{S \mid S \quad \text{maximal solution of} \quad X^2 \subseteq H\}$.

5. The importance of being well ordered

5.1 *Show that* if A is a finite alphabet, every ideal of A^* for division is a rational language *(Corollary 5.2) without using Theorem 5.3.*

Solution: If L is an ideal, $L = \mathsf{I}(K)$ for some finite K. Furthermore, we have $\mathsf{I}(K) = K \between A^*$, and the shuffle of two rational languages is rational. *cf. Prop. I.1.7, p. 65*

6. Rationals in the free group

6.1 *What is the syntactic monoid of a subset* P *of* G*? Show that a singleton of an infinite group is disjunctive. State Proposition 2.1 for groups.*

Solution [partial]: A subset of a group G is a union of cosets of a subgroup of finite index if and only if it is a union of cosets of a *normal* subgroup of finite index.

---

[87] *Cf.* [66, p. 53].

[88] This is the 'factorial function' of K [66, p. 54].

[89] *Cf.* [66, p. 55].

6.3 *Show that* $\mathrm{Rec}\, G$ *is closed under product.*

Solution: A group is *equidivisible* and the closure under product of the recognisable subsets of an equidivisible monoid was proved in Exercise 2.29.

6.5 *Show that the loop complexity of the Cayley graph of the additive group* $\mathbb{Z}/2^q\mathbb{Z}$ *with* $\{1, -1\}$ *as the generating set is* q *.*

Solution: The Cayley graph in question is the 'double ring' of length 2^q , and the result was proved in Exercise I.6.3. *See p. 203*

6.6 ***Enumeration of Dyck words.***[90] *For every word* w *in* D_A^**, we define by induction two integers:* $h(w)$*, the* height *of* w*, and* $p(w)$*, the* width *of* w *(Definition 6.1) as*

$$h(1_{\widetilde{A}^*}) = 0$$

$$h(w) = \begin{cases} 1 + h(w') & \text{if} \quad w \in D_A \quad \text{and} \quad w = z\, w'\, \overline{z} \ , \quad (z \in \widetilde{A}, w' \in D_A^*) \\ \mathsf{max}_{1 \leqslant i \leqslant n}\{h(w_i)\} & \text{if} \quad w \notin D_A \quad \text{and} \quad w = w_1 \cdots w_n \ , \quad (w_i \in D_A) \end{cases}$$

$$p(1_{\widetilde{A}^*}) = 0$$

$$p(w) = \begin{cases} 1 & \text{if} \quad w = z\overline{z} \ , \quad (z \in \widetilde{A}) \\ p(w') & \text{if} \quad w \in D_A \quad \text{and} \quad w = z\, w'\, \overline{z} \ , \quad (z \in \widetilde{A}, w' \in D_A^+) \\ \mathsf{max}_{1 \leqslant i \leqslant n}(\{p(w_i)\}, n) & \text{if} \quad w \notin D_A \quad \text{and} \quad w = w_1 \cdots w_n \ , \quad (w_i \in D_A) \end{cases}$$

(a) *Compute* $h(f)$ *and* $p(f)$ *for* $f = \overline{a}\,\overline{a}\,a\,\overline{a}\,a\,a$ *.*

(b) *Show that there is only a finite number of Dyck words of a given height and width. More precisely, prove the following lemma.*

Lemma 6.16 *Let* h *and* p *be two positive integers and* $N(h, p)$ *the integer defined by*

$$N(h, p) = 2p\frac{p^h - 1}{p - 1} \quad .$$

If a word w *in* D_A^* *has a length greater than or equal to* $N(h, p)$*, then either* $p(w)$ *is greater than or equal to* p*, or* $h(w)$ *is greater than or equal to* h*.*

(c) *Show that the bound of Lemma 6.16 is exact, even for an alphabet with one letter.*

Solution: (a) $f = \overline{a} \cdot \overline{a}\,a \cdot \overline{a}\,a \cdot a$ hence $h(f) = p(f) = 2$.

Note that the height and the width of a word are defined from the factorisation into prime Dyck words and not from a chain of reductions, as we can do when we make equivalent definitions for parenthetic words. For example, the chain

$$\overline{a}\,\overline{a}\,\underline{a\,\overline{a}}\,a\,a \to \overline{a}\,\underline{\overline{a}\,a}\,a \to \underline{\overline{a}\,a} \to 1_{\widetilde{A}^*}$$

corresponds to a width of 1 and a height of 3.

(b) By induction on h, we show that the statement is true of all p.

Let $w = w_1\, w_2 \cdots w_n \in D_A^*$, with $w_i \in D_A$. If $h(w) = 1$, then $h(w_i) = 1$ for all i and hence $w_i = z_i\, \overline{z}_i$ with $z_i \in \widetilde{A}$. Thus, $|w| = 2n = 2p$ and the base case is proved.

Suppose $h(w) = h + 1$. Then $w = w_1\, w_2 \cdots w_n$, with $w_i = z_i\, w_i' \overline{z}_i \in D_A$ and $z_i \in \widetilde{A}$. For each i, $h(w_i') = h_i \leqslant h$. Let $p_i = p(w_i) = p(w_i')$. By the induction hypothesis, we have

$$|w_i'| \leqslant 2p_i\frac{p_i^{h_i} - 1}{p_i - 1} \leqslant 2p_i\frac{p_i^{h} - 1}{p_i - 1} \quad \text{and hence} \quad |w| \leqslant 2n + \sum_{i=1}^{n} 2p_i\frac{p_i^{h_i} - 1}{p_i - 1} \quad .$$

[90] Adapted from [12].

Since $p = \mathsf{max}(n, \mathsf{max}\{p_i \mid 1 \leqslant i \leqslant n\})$, it follows that

$$|w| \leqslant 2p + p\,(2p\,\frac{p^h - 1}{p-1}) = 2p\,(1 + p\,(\frac{p^h - 1}{p-1})) = N(h+1, p) \quad .$$

(c) For all h and p, let $v_{h,p}$ be defined by

$$v_{1,1} = a\,\overline{a} \quad v_{1,p} = (v_{1,1})^p \quad , \quad \forall p \in \mathbb{N} \quad \text{and} \quad v_{h+1,p} = (a\,v_{h,p}\,\overline{a})^p \quad , \quad \forall h \in \mathbb{N} \quad .$$

It is clear that $|v_{h,p}| = N(h,p)$, $h(v_{h,p}) = h$ and $p(v_{h,p}) = p$.

$\boxed{\mathbf{6.7}}$ ***A combinatorial lemma.***[91] *Prove:*

Lemma 6.17 *Let f_1 , f_2 , …, f_n , be n words in $\widetilde{A}^*$ whose product is in D_A^*: thus, it holds $(f_1\,f_2 \cdots f_n)\rho_A = 1_{\widetilde{A}^*}$. There exists an index l, $1 \leqslant l < n$, such that*

$$|(f_l\,f_{l+1})\rho_A| \leqslant \mathsf{max}_{i \in [n]} |(f_i)\rho_A| \quad .$$

Solution: Without loss of generality, we can assume that all the f_i are reduced and that none is equal to $1_{\widetilde{A}^*}$ (if one is, the lemma is trivially verified).
As $|(f_1)\rho_A| > 0$ and $|(f_1\,f_2 \cdots f_n)\rho_A| = 0$, there exists an l, $1 \leqslant l < n$, such that

$$|(f_1 \cdots f_{l-1})\rho_A| \leqslant |(f_1 \cdots f_{l-1}\,f_l)\rho_A| \quad \text{and} \quad |(f_1 \cdots f_{l-1}\,f_l)\rho_A| > |(f_1 \cdots f_{l-1}\,f_l\,f_{l+1})\rho_A|$$

and we shall show that this l answers the question. There are two cases.

(i) $(f_1 \cdots f_{l-1})\rho_A = g_1\,g_2$

$$(f_1 \cdots f_{l-1}\,f_l)\rho_A = g_1\,g_3 \quad \text{with} \quad |g_2| \leqslant |g_3| \quad , \quad \text{hence} \quad f_l = \overline{g}_2\,g_3$$

$$(f_1 \cdots f_{l-1}\,f_l\,f_{l+1})\rho_A = g_1\,g_4\,g_6 \quad \text{with} \quad g_3 = g_4\,g_5 \quad , \quad |g_6| < |g_5| \quad , \quad \text{hence} \quad f_{l+1} = \overline{g}_5\,g_6 \,.$$

It follows that $(f_l\,f_{l+1})\rho_A = (\overline{g}_2\,g_4\,g_5\,\overline{g}_5\,g_6)\rho_A = (\overline{g}_2\,g_4\,g_6)\rho_A$ and hence

$$|(f_l\,f_{l+1})\rho_A| \leqslant |g_2| + |g_4| + |g_6| < |g_2| + |g_4| + |g_5| = |f_l| \quad .$$

(ii) $(f_1 \cdots f_{l-1})\rho_A = g_1\,g_2\,g_3$

$$(f_1 \cdots f_{l-1}\,f_l)\rho_A = g_1\,g_2\,g_4 \quad \text{with} \quad |g_3| \leqslant |g_4| \quad , \quad \text{hence} \quad f_l = \overline{g}_3\,g_4$$

$$(f_1 \cdots f_{l-1}\,f_l\,f_{l+1})\rho_A = g_1\,g_5 \quad \text{with} \quad |g_5| < |g_2| + |g_4| \quad , \quad \text{hence} \quad f_{l+1} = \overline{g}_4\overline{g}_2\,g_5$$

It follows that $(f_l\,f_{l+1})\rho_A = (\overline{g}_3\,\overline{g}_2\,g_5)\rho_A$ and hence

$$|(f_l\,f_{l+1})\rho_A| \leqslant |g_2| + |g_3| + |g_5| < |g_2| + |g_4| + |g_5| = |f_{l+1}| \quad .$$

$\boxed{\mathbf{6.9}}$ *Show that every element f of $F(A)$ distinct from $1_{F(A)}$ is contained in a recognisable subset of $F(A)$ which does not contain $1_{F(A)}$ (in other words,* the free group is residually finite*).*

Solution:[92] We identify the element f with the reduced word of $\widetilde{A}^*$ which corresponds to it, $f = x_1\,x_2 \cdots x_n$, and we write S for the set $\{0, 1, \ldots, n\}$. Each letter a of A then defines a partial action from S to itself by

$$\forall i \in S \qquad \begin{cases} (i-1) \cdot a \quad = i & \text{if} \quad x_i = a \\ (i+1) \cdot a \quad = i & \text{if} \quad x_i = \overline{a} \\ i \cdot a & \text{undefined in all other cases} \end{cases}$$

[91] Due to M. Nivat [179]; his proof is reproduced verbatim. The result was further developed with a kind of converse in [14].

[92] Proof by F. Levi [148].

It is indeed a partial map, and it is injective, since the word f is reduced; that is, contains no factor $\overline{a}\,a$ or $a\,\overline{a}$. It can be extended arbitrarily to a permutation φ_a on S.

The map from A to the finite group G of permutations of S defined by $a \mapsto \varphi_a$ extends to a morphism from $F(A)$ to G. The inverse image of the neutral element of G under this morphism is a recognisable subset, which obviously contains $1_{F(A)}$, but which does not contain f since the image of f is a permutation which sends 0 to n.

6.10 *Deduce Proposition 6.18 from Theorem 6.1.*

Solution: For all $R \subseteq A^*$, $R\eta_\Pi = (R\alpha_\Pi)\iota_\Pi$. If R is rational, $R\alpha_\Pi$ is too (Prop. 1.7), and hence $(R\alpha_\Pi)\iota_\Pi$ by Theorem 6.1.

6.11 *Let R be a rational of A^* and $\varphi\colon A^* \to F(B)$ a morphism. Show that we can decide if the intersection $R \cap (1_{F(B)})\varphi^{-1}$ is empty or not.*

Solution: The intersection $R \cap (1_{F(B)})\varphi^{-1}$ is non-empty if and only if $1_{F(B)} \in R\varphi$, which we can decide since $R\varphi$ is a rational of $F(B)$, which we can effectively compute given R.

6.12 *Convert the proof of Lemma 6.20 into an algorithm that computes the matrix $H_{\mathcal{A}}$ and whose complexity is $O(\|Q\|^3)$.*

Solution:[93] The automaton $\mathcal{A}$ is given by the representation (λ, μ, ν) of size Q. The matrix H is initialised to I, and for each a in $\widetilde{A}$, the matrix $a\pi$ is initialised to $a\mu$. Finally, a *queue*, called $EDGE$, is initialised with the set of transitions – that is, of triples in $(Q \times \widetilde{A} \times Q)$ – of $\mathcal{A}$. The program 'in Pascal' given in Figure SE.6.1

(a) computes the matrix $H_{\mathcal{A}}$, that is, at the end of the `while` loop $H = H_{\mathcal{A}}$, (and also the representation π);

(b) and its complexity is $O(\|Q\|^3)$.

6.14 *Give a simple proof (one which does not use Theorem 6.3) of the fact that it is decidable whether a finitely generated subgroup of the free group is of finite index.*

Solution:[94] Following established convention, we write $[G : H]$ for the index H in G and we shall use the elementary lemma:

Lemma 6.1 *Let $H \subseteq K \subseteq G$ be three groups. The index of H in G is finite if and only if that of H in K and that of K in G are, in which case we have:*

$$[G : H] = [G : K]\,[K : H] \ .$$

Let $H = \langle f_1, f_2, \ldots, f_l \rangle$ be a finitely generated subgroup of $\mathrm{F}(A)$. For $[\mathrm{F}(A) : H]$ to be finite it is necessary and sufficient that (setting $A = \{a_1, a_2, \ldots, a_k\}$) the following k indices are finite:

$$[\langle H \cup \{a_1\} \rangle : H] \ , \quad [\langle H \cup \{a_1, a_2\} \rangle : \langle H \cup \{a_1\} \rangle] \ , \quad \cdots \ ,$$
$$[\langle H \cup \{a_1, a_2, \ldots, a_k\} \rangle : \langle H \cup \{a_1, a_2, \ldots, a_{k-1}\} \rangle] \ .$$

The problem then becomes that of the computation of an index of the form $[\langle K \cup \{f\} \rangle : K]$, which is given by the formula

$$[\langle K \cup \{f\} \rangle : K] = \begin{cases} +\infty & \text{if} \quad \langle f \rangle \cap K = \emptyset \\ \min\{n \mid f^n \in K\} & \text{otherwise} \end{cases}$$

and which is effectively computable in the Boolean algebra of the rational subsets of $\mathrm{F}(A)$.

[93] This algorithm is given in [25]. The reader will also note the resemblance of this procedure to the one which computes the transitive closure of a graph in complexity $O(d\,m')$, *cf.* Exer. I.1.20, p. 73 and p. 185.

[94] After an idea of P. Marchand and R. Schott (unpublished).

```
while NonEmpty(EDGE)
 do (i,a,j) := Front(EDGE); Pop(EDGE);
    for all q ∈ Q
     do if ā π_{j,q} ≠ 0 then
           do if H_{i,q} = 0 then
                 do H_{i,q} := 1;
                    for all z ∈ Ã
                     do for all p ∈ Q
                           do if z π_{q,p} ≠ 0 then
                                 do if z π_{i,p} = 0 then
                                       do z π_{i,p} := 1;
                                          Enter((i,z,p), EDGE)
                                       od
                                 od
                           od
                     od
                 od
           od
        if ā π_{q,i} ≠ 0 then
           do if H_{q,j} = 0 then
                 do H_{q,j} := 1;
                    for all z ∈ Ã
                     do for all p ∈ Q
                           do if z π_{j,p} ≠ 0 then
                                 do if z π_{q,p} = 0 then
                                       do z π_{q,p} := 1;
                                          Enter((q,z,p), EDGE)
                                       od
                                 od
                           od
                     od
                 od
           od
     od
 od
```

Figure SE.6.1: Computation of the matrix $H_{\mathcal{A}}$

6.15 ***Pure Büchi grammars.***[95] *We shall say that a Büchi grammar* $\mathcal{G} = \langle A, N, K, U \rangle$ *is* pure *if* $N = \emptyset$. *cf. p. 321*

(a) *Show that every rational language is generated by a pure Büchi grammar.*

(b) *Construct the pure Büchi grammar corresponding to the automaton* $\mathcal{P}_2$ *(Figure SE.6.2) and give the series of derivations that produces the word* $a\,b\,a\,b\,a$.

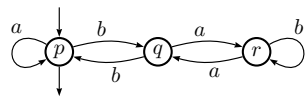


Figure SE.6.2: The 'divider by 3'

Solution: (a) Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton which we shall specify later. All the following computations are performed in $\mathcal{A}$; we first make some definitions. A computation c,

$$c \;:=\; p = p_0 \xrightarrow{a_1} p_1 \xrightarrow{a_2} p_2 \longrightarrow \cdots \xrightarrow{a_n} p_n = q \;\;,$$

is *very simple* if all the p_i, $0 \leqslant i \leqslant n$, are distinct; it is *almost very simple* if its source is repeated in the rest of the computation and it is the only repeated state; that is, there exists k, $1 \leqslant i \leqslant n$, such that $p = p_0 = p_k$, and if the computation of source p_1 (the maximum proper suffix of c) is very simple. (N.B. a computation of length 0 is very simple, labelled 1_{A^*}, and a computation $p \xrightarrow{a} p$, where a is in A, is almost very simple.)[96]

[95] Statement and solution adapted from [45].

[96] *Note added in translation:* These definitions are slightly more restricted than the one of (almost) simple computations used in the proof of Proposition 6.2, p. 302, and which are consistent with the

An almost very simple computation c can be written uniquely in the form $c := p \xrightarrow{w} p \xrightarrow{v} q$. Every computation $r \xrightarrow{f} q$ can be written uniquely in the form $r \xrightarrow{h} p \xrightarrow{g} q$, where $p \xrightarrow{g} q$ is almost very simple.

Suppose now that $\mathcal{A} = \langle Q, A, E, I, t \rangle$ is *co-deterministic*, which is always possible without restricting $L(\mathcal{A})$. A computation which ends at t is called *terminal*; and since $\mathcal{A}$ is co-deterministic, each word in A^* is the label of *at most one* terminal computation.

Let U be the set of labels of very simple successful computations:

$$U = \{u \mid \exists i \in I \quad i \xrightarrow{u} t \quad \text{is a very simple computation}\} \ ,$$

and let K be the set of pairs $(v, w\,v)$ where $w\,v$ is the label of an almost very simple terminal computation c and where w is the label of the loop of c:

$$K = \{(v, wv) \mid \exists p \in Q \quad p \xrightarrow{w} p \xrightarrow{v} t \quad \text{is an almost very simple computation}\} \ .$$

Then the pure Büchi grammar $\mathcal{G} = \langle \emptyset, A, K, U \rangle$ is what we seek; that is, is such that $L(\mathcal{G}) = L(\mathcal{A})$.

We first show that $L(\mathcal{A}) \subseteq L(\mathcal{G})$, by *reductio ad absurdum*. Let f be a word of minimal length such that $f \in L(\mathcal{A}) \setminus L(\mathcal{G})$ and let $c := i \xrightarrow{f} t$ be the successful computation whose label it is. If c is very simple, $f \in U \subset L(\mathcal{G})$: contradiction. If c is not very simple, it can be factorised into

$$c := i \xrightarrow{h} p \xrightarrow{w} p \xrightarrow{v} t \qquad \text{where} \qquad p \xrightarrow{w} p \xrightarrow{v} t \qquad \text{is almost very simple.}$$

It follows that

$$h\,v \vdash_{\mathcal{B}\,K} h\,w\,v = f \ ,$$

and as $h\,v$ is the label of the successful computation $c' := i \xrightarrow{h} p \xrightarrow{v} t$, then $h\,v$ belongs to $L(\mathcal{G})$ by hypothesis (of the minimality of f), and hence so does f, which is a contradiction.

We show conversely, by induction on the length of derivations, that $L(\mathcal{G}) \subseteq L(\mathcal{A})$. By definition, $U \subseteq L(\mathcal{A})$ and the property is true for length 0. Let g be such that there exists u such that

$$u \overset{n}{\vdash_{\mathcal{B}\,K}} f \vdash_{\mathcal{B}\,K} g \ .$$

By definition, there exists $(v, w\,v)$ in K such that $f = h\,v$ and $g = h\,w\,v$, and an almost very simple computation $p \xrightarrow{w} p \xrightarrow{v} t$. By the induction hypothesis, $f \in L(\mathcal{A})$ and there exists a (unique) computation $i \xrightarrow{h} q \xrightarrow{v} t$.

Because $\mathcal{A}$ is co-deterministic, there exists only one terminal computation labelled v and hence $p = q$. It follows that $i \xrightarrow{h} p \xrightarrow{w} p \xrightarrow{v} t$ is a (successful) computation of $\mathcal{A}$ and $g \in L(\mathcal{A})$.

(b) The very simple terminal computations of the automaton $\mathcal{P}_2$ are (*cf.* Figure SE.6.2)

$$\longrightarrow p \longrightarrow \ , \qquad q \xrightarrow{b} p \longrightarrow \ , \qquad r \xrightarrow{a} q \xrightarrow{b} p \longrightarrow \ .$$

and the almost very simple terminal computations are

$$\longrightarrow p \xrightarrow{a} p \longrightarrow \ , \qquad \longrightarrow p \xrightarrow{b} q \xrightarrow{b} p \longrightarrow \ , \qquad r \xrightarrow{b} r \xrightarrow{a} q \xrightarrow{b} p \longrightarrow \ ,$$
$$\text{and} \qquad q \xrightarrow{a} r \xrightarrow{a} q \xrightarrow{b} p \longrightarrow \ .$$

---

widely accepted definition of *simple path*. They will be used only in the solution of this exercise, which they make simpler, for they insure the unicity of decomposition.

We deduce the associated pure Büchi grammar, $\mathcal{G}_1 = \langle \emptyset, A, K_1, U_1 \rangle$, with

$$K_1 = \{(1_{A^*}, a), (1_{A^*}, bb), (b, aab), (ab, bab)\} \qquad \text{and} \qquad U_1 = \{1_{A^*}\} \ .$$

Finally, the series of derivations that produces the word $abababa$:

$$\begin{array}{llll} 1_{A^*} \underset{\mathcal{B}_{K_1}}{\vdash} a = (1_{A^*} \cdot a)\tau_A & \underset{\mathcal{B}_{K_1}}{\vdash} abb = (a \cdot bb)\tau_A & & \\ \underset{\mathcal{B}_{K_1}}{\vdash} abaab = (abb \cdot \overline{b}aab)\tau_A & \underset{\mathcal{B}_{K_1}}{\vdash} ababab = (abaab \cdot \overline{b}\overline{a}bab)\tau_A & & \\ & & \underset{\mathcal{B}_{K_1}}{\vdash} abababa = (ababab \cdot a)\tau_A & \end{array}$$

$\boxed{\textbf{6.16}}$ ***'Lead or Delay' action revisited.*** *We return to the definition of the map χ_B: let B^* be an alphabet and $H = (B^* \times 1_{B^*}) \cup (1_{B^*} \times B^*) \cup \{\mathbf{0}\}$. We define a map ψ from $B^* \times B^*$ to H by* *cf. Exer. 2.7, p. 237*

$$\forall f, g \in B^* \qquad (f, g)\psi = \begin{cases} (g^{-1}f, 1_{B^*}) & \text{if} \quad g \leqslant f \\ (1_{B^*}, f^{-1}g) & \text{if} \quad f \leqslant g \\ \mathbf{0} & \text{otherwise} \ , \end{cases}$$

and a map χ_B from $H \times (B^* \times B^*)$ to H by:

$$\forall f, g, u, v \in B^* \qquad \begin{cases} ((f, 1_{B^*}), (u, v))\, \chi_B & = (f\, u, v)\psi \\ ((1_{B^*}, g), (u, v))\, \chi_B & = (u, g\, v)\psi \\ (\mathbf{0}, (u, v))\, \chi_B & = \mathbf{0} \ . \end{cases}$$

Express χ_B with the help of the Dyck reduction and verify that it is an action of $B^* \times B^*$ on H.

Solution: We write, logically, ρ_B for the Dyck reduction on $\widetilde{B}^*$. Let $H' = B^* \cup \overline{B}^*$; there is an obvious bijection between H' and the non-zero elements of H, namely $(f, 1_{B^*})$ is associated with f and $(1_{B^*}, g)$ with $\overline{g}$. The maps ψ and φ are thus translated by

$$(f, g)\psi = (\overline{g}\, f)\rho_B \circ \iota_{H'} \qquad \text{and} \qquad (z, (u, v))\varphi = (\overline{v}\, z\, u)\rho_B \circ \iota_{H'} \ .$$

To show that φ is an action thus means showing that for all z in H' and all u, v, x and y in B^*, we have

$$(\overline{y}\, \overline{v}\, z\, u\, x)\rho_B \circ \iota_{H'} = (\overline{y}\, [(\overline{v}\, z\, u)\rho_B \circ \iota_{H'}]\, x)\rho_B \circ \iota_{H'} \ .$$

Corollary 6.9 implies

$$(\overline{y}\, \overline{v}\, z\, u\, x)\rho_B = (\overline{y}\, [(\overline{v}\, z\, u)\rho_B]\, x)\rho_B$$

and hence the equality in the case where $(\overline{v}\, z\, u)\rho_B$ belongs to H'.

If $(\overline{v}\, z\, u)\rho_B$ does not belong to H', then $(\overline{v}\, z\, u)\rho_B$ belongs to $\overline{B}^*\, \overline{a}\, b\, B^*$ – with $a \neq b$ – hence also $\overline{y}\, [(\overline{v}\, z\, u)\rho_B]\, x$ and the equality is also proved in this case.

7. Rationals in commutative monoids

$\boxed{\textbf{7.1}}$ *We say that a finite antichain X of an ordered set is* wide *if the cardinal of every antichain of E that contains X is bounded.*

(a) *Show that every antichain of $\mathbb{N}^2$ is wide.*

(b) *Give a necessary and sufficient condition for an antichain of $\mathbb{N}^k$, $k \geqslant 3$, to be wide.*

Solution: (a) Let $X = \{x_1, x_2, \ldots, x_p\}$ be an antichain in $\mathbb{N}^2$. For each i we write $x_i = (x'_i, x''_i)$ and X is indexed in such a way that $x'_1 \leqslant x'_2 \leqslant \cdots \leqslant x'_p$. For X to be an antichain, these inequalities must be strict: $x'_1 < x'_2 < \cdots < x'_p$, and the strict inverse inequalities

$x''_1 > x''_2 > \cdots > x''_p$ must hold. Thus, for all i in $[p]$, we have $i \leqslant x'_i$ and $p - i \leqslant x''_i$ and every antichain which contains X has at most k elements, with

$$k = \inf_{i \in [p]} (x'_i + x''_i) \ .$$

(b) Let X be an antichain in $\mathbb{N}^k$. A necessary condition for X to be wide is that, for every pair (i, j) of coordinates, there exists an element x of X in $\mathbb{N}^k_{i,j}$, the submonoid of $\mathbb{N}^k$ formed from elements all of whose coordinates are zero except possibly the coordinates i and j. Otherwise, we can add to X an arbitrarily long antichain in $\mathbb{N}^k_{i,j}$. This condition is also sufficient.

7.2 *Show*[97] *that every rational language contains a bounded rational language with the same commutative image (Proposition 7.4).*

see p. 326

Solution: The property is true for finite languages. As the union and product of bounded rationals are bounded rationals, it suffices to show that if the rational language L satisfies the property (we can then assume it is bounded) then so does L^*. Therefore, let

$$L = \bigcup_{i \in [N]} S_i \ ,$$

where N is an integer and each of the S_i can be written (omitting one level of indices)

$$S_i = u_0 \, v_1^* \, u_1 \, v_2^* \cdots u_n \, v_n^* \, u_{n+1} \ .$$

We set $K = S_1^* S_2^* \cdots S_N^*$ and we have $K \subseteq L^*$ and $K\alpha = L\alpha$ (here α is the canonical morphism from A^* to $A^{\oplus}$). It is enough to show the property for K. For each (non-empty) subset Q of $[N]$, we set

$$K_Q = S_{j_1}^+ S_{j_2}^+ \cdots S_{j_q}^+ \ ,$$

with $j_1 < j_2 < \cdots < j_q$ and $Q = \{j_1, j_2, \ldots, j_q\}$. We have

$$K = \bigcup_{Q \in \mathfrak{P}([N])} K_Q \ .$$

It thus suffices to show that, for all i, S_i^+ satisfies the property. Let $u = u_0 \, u_1 \cdots u_n \, u_{n+1}$; then

$$u^* S_i \subseteq S_i^+ \qquad \text{and} \qquad (u^* S_i)\alpha = (S_i^+)\alpha$$

and $u^* S_i$ is a bounded rational set.

7.3 *Explain how we can sensibly give the same name to the lexicographic order defined on A^* as to that defined on $\mathbb{N}^k$.*

Solution: Let $A_\$ = A \cup \{\$\}$, where \$ is a symbol which does not belong to A. We identify A^* with the subset of $A_\$^{\mathbb{N}}$ – that is, of maps from $\mathbb{N}$ to $A_\$$ – that are constant from the first index for which they take the value \$ and that only take values distinct from \$ for a finite number of indices.

Suppose that $A_\$$ is (totally) ordered in such a way that \$ is the greatest element. We can immediately verify that the lexicographic order defined on A^* is the trace of the lexicographic order defined on $A_\$^{\mathbb{N}}$ by the general method used to define a well ordering on the set of maps from one well ordered set to another.

[97] *Cf.* [146, Lemma 4].

7.4 (a) *Describe a well ordering on $A^{\oplus}$ that satisfies (7.3) and is not a lexicographic order, whatever order is chosen on A.*

(b) *Describe a well ordering on $A^{\oplus}$, that satisfies (7.3) and such that the set of minorants of every element is finite.*

Solution: The classic procedure of enumerating the elements of $\mathbb{N}^k$ in a zig-zag gives an order that satisfies the two conditions.

For example, if $k = 2$, $x = (x', x'')$ has rank $\mathsf{o}(x) = \frac{1}{2}(x' + x'')(x' + x'' + 1) + x'$ and x is less than y if and only if $\mathsf{o}(x) < \mathsf{o}(y)$.

7.5 ***Simplifiable commutative monoid.***

(a) *Let η be a congruence on M. For each m in M, we define the congruence η_m by*

$$\forall x, y \in M \qquad x \equiv y \bmod \eta_m \quad \Longleftrightarrow \quad x + m \equiv y + m \bmod \eta \ .$$

Verify

$$\forall m \in M \qquad \eta \subseteq \eta_m \tag{7.7}$$

$$\forall m \in M \qquad \eta \subseteq \xi \quad \Longrightarrow \quad \eta_m \subseteq \xi_m \tag{7.8}$$

$$\forall m, m' \in M \qquad \eta_{m+m'} = (\eta_m)_{m'} \tag{7.9}$$

Verify that the quotient M by η is simplifiable if and only if for all m in M, $\eta_m = \eta$. (We say in this case that η is simplifiable*).*

(b) *Prove:*

Proposition 7.7 *Let η be a congruence on a finitely generated monoid M. The family of congruences $\{\eta_m \mid m \in M\}$ contains exactly once simplifiable congruence and this is its maximum set.*

(c) *Prove:*

Proposition 7.8 *Let M be a commutative monoid. There exist a commutative group $K(M)$ and a morphism (of monoids) $\gamma\colon M \to K(M)$ such that every morphism from M to a commutative group factorises in terms of γ. The monoid M is simplifiable if and only if γ is injective.*

Solution: (a) A direct consequence of Corollary 7.6. *See p. 328*

(b) If there were no simplifiable congruences in the family $\{\eta_m \mid m \in M\}$, we could construct an infinitely ascending chain of congruences of M. Therefore let η_s be a simplifiable congruence. For all m in M, we deduce from $\eta \subseteq \eta_s$ that $\eta_m \subseteq (\eta_s)_m = \eta_s$: η_s is the maximum element, hence unique, of the family $\{\eta_m \mid m \in M\}$.

(c) *First method:* M is finitely generated. Following in the footsteps of the previous proof, we see that η_s is the finest of the congruences coarser than η and is simplifiable. In fact, if ξ is such a congruence, $\eta \subseteq \xi$ implies $\eta_s \subseteq \xi_s = \xi$.

If we apply this result to the identity congruence ι of M, we deduce that every morphism σ of M in a simplifiable monoid factorises uniquely through the canonical morphism of M on the quotient of M by the unique simplifiable congruence ι_s.

Second method: as indicated by the proposition, the result holds without requiring that M be finitely generated and follows from a classic algebraic construction.[98]

The *free Abelian group* with basis B can be identified with the set of formal sums of elements of B, with coefficients in $\mathbb{Z}$ and almost all zero. The group $K(M)$, called the *Grothendieck group* of M, is isomorphic to the quotient of the free Abelian group with basis M by the subgroup generated by the elements $[x + y] - [x] - [y]$, for all x and y in M.

[98] *Cf.* [145, p. 43 *et seq.*].

On the monoid $M \times M$, we define the relation τ by

$$(x,y) \equiv (x',y') \bmod \tau \quad \Longleftrightarrow \quad \forall z \in M \quad x + y' + z = x' + y + z \ .$$

We verify that τ is a congruence on $M \times M$ and that $K(M)$ is isomorphic to the quotient of $M \times M$ by τ. We can identify M with the set of pairs $(x,0)$, and the morphism σ is the composition of this injection with the canonical morphism from $M \times M$ to its quotient by τ.

If M is simplifiable, and only in this case, the congruence τ is defined by

$$(x,y) \equiv (x',y') \bmod \tau \quad \Longleftrightarrow \quad x + y' = x' + y \ ,$$

and we recognise there, for example, the classic construction of the integers from the naturals, or that of the rationals from the multiplicative monoid of the non-zero integers. The second part of Proposition 7.8 states that M is simplifiable if and only if it is the submonoid of a group, which is not true for non-commutative monoids.

If M is not simplifiable, the restriction of τ to the set of pairs $(x,0)$ – identified with M – is the upper bound of the family of congruences $\{\iota_m \mid m \in M\}$. This upper bound is itself a congruence ι_s in the case where M is finitely generated, which is what Proposition 7.7 shows.

7.8 *Give some examples of submonoids of $\mathbb{N}^2$ which are not finitely generated.*

Solution: (i) The set $Q = \{(0,0)\} \cup \big[(1,1) + \mathbb{N}^2\big]$ is a submonoid; a generating set of Q must contain the elements $(n,1)$ and $(1,n)$ for all positive integers n.

(ii) Let $r > 0$ be an *irrational* number. The set

$$P_r = \{(x,y) \in \mathbb{N}^2 \mid y \leqslant r\,x\}$$

is a submonoid which is not finitely generated.

cf. p. 332 **7.9** *Let S be an affine subset of M. Verify the following properties (Properties 7.2):*

(i) *For all x in M, $S - x$ is affine.*

(ii) *If N is a commutative monoid and $\alpha\colon N \to M$ is a morphism, $S\alpha^{-1}$ is an affine subset of N.*

(iii) *The intersection of two affine subsets is affine.*

Solution:

(i) For all s in $S - x$, $s + x$ is in S and $\{S - (s+x)\} = \{S - x\} - s$ is a submonoid.

(ii) Follows from Property 7.1, since $(S\alpha^{-1}) - t = (S - t\alpha)\alpha^{-1}$ and $t\alpha$ is in S if t is in $S\alpha^{-1}$.

(iii) For all s in $S \cap T$, $S - s$ and $T - s$ are submonoids, their intersection is a submonoid and $(S \cap T) - s = (S - s) \cap (T - s)$.

7.10 ***Stable submonoid.*** *We say that a finitely generated submonoid M of $\mathbb{N}^k$ is* stable *if it is the intersection $M = C \cap H$ of a closed cone C in $\mathbb{Q}_+^k$ and a subgroup H of $\mathbb{Z}^k$.*

Show that M is stable if and only if

$$\forall u \in M\,,\ \forall v \in \mathbb{N}^k \qquad u + v^{\oplus} \subset M \quad \Longrightarrow \quad v \in M \ . \tag{7.12}$$

Solution:[99] If $u + v^{\oplus} \subseteq M$, then u and $u + v$ are in M hence $v = (u+v) - u$ is in H. For all n in $\mathbb{N}$, $u + n\,v$ is in C hence also $\frac{1}{n}(u + n\,v)$ which tends to v; as C is closed, v is in C, hence in M.

Conversely, suppose that M satisfies (7.12). Let $H = \langle M \rangle = M - M$ be the subgroup of $\mathbb{Z}^k$ generated by M, and let $C = \mathbb{Q}_+ \cdot M = \{x \in \mathbb{Q}^k \mid \exists q \in \mathbb{N}\,,\ \exists m \in M \quad x = \frac{1}{q}\,m\}$, the cone generated by M; we have C is closed since M is finitely generated.

[99] Due to S. Lombardy (private communication).

Let x be in $C \cap H$; $x \in H$ hence $x = y - z$ with y and z in M; $x \in C$ hence there exists q in $\mathbb{N}$ such that $q\,x$ is in M. We verify that $(q\,z) + x^{\oplus} \subseteq M$. Every integer k can be written $k = n\,q + r$ with $r < q$ hence

$$(q\,z) + k\,x = (q\,z) + (n\,q + r)\,(y - z) = ((q - r)\,z) + r\,y + n\,(q\,x)$$

is in M and by (7.12) x is in M.

7.12 *Prove:*

Corollary 7.27 *Let $\beta\colon M \to N$ be a morphism on commutative monoids. If M is finitely generated and if X is a rational of N, $X\beta^{-1}$ is a rational of M.*

Solution: Let $\alpha\colon \mathbb{N}^k \to M$ be a morphism, surjective since M is finitely generated. Then $X\beta^{-1} = ((X\beta^{-1})\alpha^{-1})\alpha = (X(\alpha \circ \beta)^{-1})\alpha$ and $X(\alpha \circ \beta)^{-1}$ is rational by Proposition 7.25.

7.13 *Prove the Fatou property for commutative monoids:*

Corollary 7.28 *Let M' be a finitely generated submonoid of a commutative monoid M. A rational of M included in M' is a rational of M'.*

Solution: In the previous corollary, set $N = M$, $M = M'$ and β the canonical injection.

7.14 *Prove:*

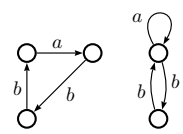
Corollary 7.29 *If X and Y are two rationals of a finite generated commutative monoid M, $Y - X$ is a rational of M.*

Solution: We consider the two morphisms: $\sigma\colon M \times M \to M$ with: $(z,t)\sigma = z + t$ and $\pi_2\colon M \times M \to M$ with: $(z,t)\pi_2 = t$. Then, $Y - X = \big[(Y\sigma^{-1}) \cap (X \times M)\big]\,\pi_2$.

8. Star height

8.1 *Give an example of a non-conformal (but surjective) morphism from one ball to another which has the same loop complexity.*

Solution : See adjacent figure.



8.2 *Show that the minimal automaton of L has the least loop complexity of all the deterministic automata that recognise a language L over A^*.*

Solution: The morphism that sends a deterministic automaton to the minimal automaton of the language that it recognises is a covering (Exer. 3.13), hence a conformal morphism, and we apply Theorem 8.2.

NOTES & REFERENCES

Authors who tackle the theory of automata with an algebraic perspective in mind are obviously fewer than those who have taken a 'naive' approach. In general they have done it in order to deal with parts of the 'theory of varieties' (see below).

J. Berstel, *Transductions and Context-Free Languages*. Teubner, 1979. [27]

S. Eilenberg, *Automata, Languages and Machines*, volume A. Academic Press, 1974. [80]

J. M. Howie, *Automata and Languages*. Oxford, 1991. [123]

G. Lallement, *Semigroups and Combinatorial Applications*. Wiley, 1979. [144]

J.-E. Pin, *Variétés de langages formels*. Masson, 1984. Translation: *Varieties of Formal Languages*, Oxford University Press, 1986. [185]

1. Automata and rational subsets

Jean Berstel [27] attributes to Eilenberg the idea of distinguishing the family of recognisables from that of rationals.

It seems that Theorem 1.1 appears for the first time in [84]; hidden in a footnote of this seminal but difficult article, it was then given many statements, usually in special cases. Eilenberg calls[100] the automata over a monoid *generalized M-automata* (not to be confused with what we here call generalised automata). The proof of Proposition 1.7 is that given by Eilenberg.

It also seems that the definition of unambiguous rational subsets is due to Eilenberg and Schützenberger [83]; Proposition 1.12 is implicit there (*cf.* [206]).

2. Actions and recognisable subsets

Everything that pertains to the definition of actions and their minimisation belongs to folklore. Theorem 2.3 combines two propositions usually attributed to Nerode and Myhill in the case of languages. Proposition 2.5 is due to J. McKnight [166], Corollary 2.20 is unanimously ascribed to G. Mezei (unpublished). Corollary 2.17 is due to C. Reutenauer [196]. Example 2.6 is taken from Eilenberg *op. cit.*, who attributes it to S. Winograd.

3. Morphisms and coverings

Of course, the notion of automata morphism is not new; it is particularly to be found in the literature that deals with automata under the name of *transition systems* (*cf.* [9] for example). Quotients are there called 'simulations'.

As already mentioned, the notion of a covering of an automaton is adapted from that defined for graphs by J. R. Stallings [239]. The definition of a Schützenberger covering and of the corresponding immersions and Corollary 3.14 came out of my own reading [211] of two articles by Schützenberger [225, 227].

4. Universal automaton

We have also already mentioned that the idea of factors, and the associated results, are due to J. H. Conway [66, Ch. 6]. I was inspired by an unpublished manuscript sent to me by Olivier Carton as to the syntactic virtues of factorisations and above all for the presentation of the universal automaton, which was not really elaborated in [66].

[100] See *op. cit.*, § VII.10

The method of construction of the universal automaton given in Subsection 4.3 was adapted[101] from the works of S. Lombardy [152]; the expansion of a deterministic automaton which is its starting point was introduced by R. Cohen and J. Brzozowski [64] in the context of the determination of the star height of certain languages, a subject whose proximity we came to understand in Section 8. The idea of covering the graph of a relation with maximal 2-factorisations was rediscovered by B. Courcelle, D. Niwinski and A. Podelski [67] who described the determinisation and minimisation of automata in terms of them, and applied them to *tree automata*. In [10], it is mentioned that the idea of constructing a 'universal' automaton from 2-factorisations of a language is also found in an unpublished report by C. Carrez [50].

5. The importance of being well ordered

The notion of a 'well quasi-ordering' seems to be due to Higman [118], who proved Theorem 5.2 with the aim of simplifying the original proof of Neumann [176] which we shall see in Chapter IV (Th. IV.4.7). This result is the starting point for an abundant literature on partial well orderings and related orderings, of which a critical overview is given in [140]. In the domain of word combinatorics and language theory, Higman's Theorem for the free monoid is often attributed to Haines [110]. The proof given here is that of Conway [66] who used an idea of Nash-Williams. *cf. p. 575*

Theorems 5.3 and 5.4 and their proofs are due to A. Ehrenfeucht, D. Haussler and G. Rozenberg [78].

6. Rationals in the free group

Proposition 6.2 was proved in [7] with an induction on the star height (*cf.* [27]) which is, in fact, how we proved Proposition 6.4. The proof given here is that of [97], where Proposition 6.4 is also stated.

The idea of proving the properties of Dyck and Shamir reductions together is attributed to Schützenberger by Jean Berstel who uses it in his work on rational transductions [27] with a formalism very close to that of simplifications. The proof of Proposition 6.13 is essentially Schützenberger's (and Chomsky's?) proof that the Dyck language is an *unambiguous* algebraic language in [61]; it is also found in [27].

Theorem 6.1 is due to M. Benois [24]; the proof that we give, which has the merit of being generalisable, is due to M. Fliess [89].

The references for the other results were given in the course of the text, and in the corresponding part of the solutions to the exercises.

7. Rationals in the commutative monoids

Lemma 7.1, due to L. E. Dickson [75, Lemma A], a simple but fundamental result, has often been (re)discovered. One tradition ascribes it to D. König. In his work *Theorie der endlichen und unendlichen Graphen* [136], published in 1936, D. König states it and attributes it to himself, in an earlier note.[102] The renown of this first treatise on graph theory explains the tradition more than the publication in the Magyar journal. L. Redei [193, p.62], less of a nationalist, restores the result to Dickson.

Theorem 7.3 (and Lemma 7.16, a vestige of the original proof) are due to S. Ginsburg and E. Spanier ([101], *cf.* also [99, Ch. 5 and Appendix]). For the rest, as already stated, I have followed Eilenberg and Schützenberger [83].

It should be mentioned in addition that the rational subsets of free commutative monoids are exactly the sets definable in Presburger arithmetic (*cf.* [102] and Remark 7.5); they are

[101] I am particularly indebted to S. Lombardy for his help with the writing of this section.

[102] Egy véssegi tétel és alkalmázasai, *Mathematikai és Fisikai Lapok* [135].

most often used in this form, especially in the community that deals with the verification and specification of systems. Another point which I did not deal with is the characterisation due to A. Muchnik of the same rational subsets of free commutative monoids; it is an essential step in a new proof of the Cobham–Semenov Theorem, and is presented clearly in [39].

8. Star height of group languages

This new presentation of the proof of McNaughton's Theorem [168] was done in collaboration with Sylvain Lombardy [155].

Theorem 8.2 is due to McNaughton [168], and the given version of the proof of Lemma 8.3 to S. Lombardy.

Theorem 8.3 states that the method – the universal automaton contains a sub-automaton of minimum loop complexity – holds for all *reversible* languages [154]. Thus extended, this result strictly includes those of R. Cohen [63] and K. Hashiguchi [114] on the star height of languages called *reset free*, which are the languages whose minimal automaton is reversible.[103]

Furthermore, Hashiguchi [117] has shown that the star height of rational languages is computable, but the algorithm that corresponds to his proof is of hopeless complexity, even for small automata. The problem of finding another proof of this beautiful result, using the universal automaton, or any other method, is open.[104]

*

'Variety theory' is covered in books by Almeida [6], Eilenberg [81], Howie [123], Lallement [144] and Pin [185]. The chapters by Perrin [183] and Pin [186] in the two handbooks of the area give references to the numerous works to which it has given rise.

[103] *Note added in translation:* The proof of this result is due to S. Lombardy and has recently been published in English [157].

[104] *Note added in translation:* We have already written twice (p. 49 and p. 158) that D. Kirsten gave a solution to this problem in [132].

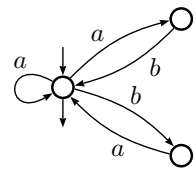
Chapter III

The pertinence of enumeration

Weighted automata

One last time we erase everything – but forget nothing – and start again. We have seen that automata are structures that accept, or do not accept, *words* (Chapter I) or, more generally, *elements of a monoid* M (Chapter II). The starting point of the approach which we shall develop in this chapter is to be *more precise* and to *count* the computations that make an element m of M be accepted by an automaton. That is, an automaton will no longer associate with each m the value 'true' or 'false' but a *number*, the number of successful computations of which m is the label, and will thus give us richer information.

Let us consider for example the automaton $\mathcal{A}_X$ in the adjacent figure. It is not difficult to convince ourselves not only that the language accepted by $\mathcal{A}_X$ is $\{a, ab, ba\}^*$ but also that the number of computations of $\mathcal{A}_X$ labelled by a word f in $\{a, b\}^*$ is exactly the number of ways in which f can be factorised into powers of a, ab and ba.



We shall not stop there. If we count the computations, then each computation, even each transition, can be ascribed a *coefficient*. Above all, this coefficient need not be an integer, but can be taken in any set in which we can reasonably compute; that is, any semiring $\mathbb{K}$. The behaviour of automata becomes what we can call *series over M with coefficients in* $\mathbb{K}$.

We shall thus embark on a generalisation that will go far beyond its starting point. First, by choosing as the coefficient set a semiring less impoverished than the Boolean semiring, we shall find ourselves with objects, series, blessed with a richer, more solid structure. We can then study them with the standard methods of algebra and topology, not necessarily complex, but powerful. Some fundamental problems, such as equivalence or minimisation of automata, will thereby be transformed.

Next, working over $\mathbb{K}$ and M, we shall capture a variety of situations, from the shortest paths in a graph to discrete event systems by way of the ambiguity of grammars and relations between words, which will seem completely to exhaust the possibilities available. Sometimes, the same situation will allow itself to be described in several different ways, and its properties will be seen in a new light from a different direction. In all these cases, the common mathematical framework will allow us to treat seemingly distinct properties together, reveal relationships and suggest methods.

Such a generality has a cost: somewhat heavy notation, some assumptions required to state and prove certain propositions which would be obvious in a simpler context, a degree of abstraction which may make us lose sight of the effectiveness and of the corresponding constructive nature of results that are proved at a global level by what could seem merely a formal game. Memories of the two preceding chapters, which will obviously feed into the general framework, and a certain redundancy in the propositions, will help the reader to keep up.

More fundamentally, the subject reveals a difficulty, which obliges the author to make a choice, or at least, to distinguish between certain situations. We will have automata whose behaviour will no longer be a subset of a monoid, but a *series*. To extend the results seen earlier, we shall require the set of series, which we shall write $\mathbb{K}\langle\langle M\rangle\rangle$, to be a semiring, like $\mathfrak{P}(M)$, and this is the first problem. To relate the behaviour of automata to 'rational operations' we shall have to define a star operation: this is the second problem. How, and with what assumptions, can we justify writing (if s is a series)

$$s^* = \sum_{n\in\mathbb{N}} s^n \ , \qquad \text{the natural counterpart of} \qquad P^* = \bigcup_{n\in\mathbb{N}} P^n \ ,$$

where P is a subset of a monoid? How can we ensure the properties of such an operation? In this introduction, we shall just sketch how to resolve this last problem.

Mathematics gives us a standard way of dealing with infinite sums: to put a *topology* on the set being considered and to use the notions of limit and continuity. In the current situation, the difficulty is that there is no solution – that is, a way to define a topology – that works for *all possible choices* of the semiring of coefficients $\mathbb{K}$ and of the basis monoid M.

There are at least two options. The first[1] is to restrict ourselves to the case where M is a free monoid A^*. It is then easy to define, traditionally, from the 'length of words' function, a topology on $\mathbb{K}\langle\langle A^*\rangle\rangle$, whatever $\mathbb{K}$ is, and to consider the elements of the latter all to be 'detached' from each other (a discrete topology on $\mathbb{K}$). The second is, contrariwise, to allow M to be arbitrary and deduce the topology over $\mathbb{K}\langle\langle M\rangle\rangle$ from the properties of $\mathbb{K}$ which must be chosen appropriately.

We shall take a position[2] which, though it gives us more general propositions, remains fundamentally close to the first: we shall retain the restriction on M and require that we can define the equivalent of a length function on its elements, but without M necessarily being a free monoid. We will just as well be able to define a topology on $\mathbb{K}\langle\langle M\rangle\rangle$, for every semiring $\mathbb{K}$, itself equipped with a topology that is not necessarily discrete. The gain in generality will be used, among others, in the study of relations and functions on words implemented by finite automata in the two following chapters. The description of this type of monoid, which we shall call *graded*, then of the topology on $\mathbb{K}\langle\langle M\rangle\rangle$, occupies the first section. We then present results on weighted

[1] This is Berstel and Reutenauer's choice in their monograph *Rational Series and Their Languages*, Springer, 1988 [31].

[2] In large measure it is that of Eilenberg [80] and of Salomaa and Soittola [218].

automata in this framework, in the two subsequent sections, which correspond to the two first sections[3] of Chapter II.

In Section 4, we shall restrict the above option to the case where M is a free monoid. This changes nothing about $\mathbb{K}$, which can already be arbitrary in the first four sections, but allows on one hand a different treatment of results already demonstrated and on the other to prove other results which are particular to series over the free monoid. We shall concentrate on two points particularly. First, we shall return to the *derivation of expressions*, seen in the first chapter. This is a good example of the introduction of weights showing itself to be illuminating and powerful. Next, if $\mathbb{K}$ is a field (not necessarily commutative), we shall describe the completely algebraic theory, due to Schützenberger, of the *minimisation of automata* via their matrix representation.

Complementarily, we begin by returning to the assumptions about the choice of $\mathbb{K}$ and M and describe a class of semirings $\mathbb{K}$ such that $\mathbb{K}\langle\langle M\rangle\rangle$ can be defined without restriction on M, to be able at least to replace the results described in the previous chapter for the *subsets* of an arbitrary monoid in series terms. This section is brief and serves as a pointer to future developments.[4] We shall then come back, in Section 6, to the study of rational subsets of a family of monoids (no more weights!) with the matrix techniques so commonly used in this chapter. We end with a short non-commutative linear algebra primer, whose aim is the definition of the rank of a matrix, even in the case when the coefficients are taken from a non-commutative field.

[3] In fact to the first *three*, since the study of $\mathbb{K}$-coverings is integral to Section 2.

[4] In particular regarding the relationship between pushdown automata and series on the free group, which is not dealt with in this volume.

Contents of the chapter

1. Formal power series on a graded monoid
2. $\mathbb{K}$-automata and $\mathbb{K}$-rational series
3. $\mathbb{K}$-representations and $\mathbb{K}$-recognisable series
4. $\mathbb{K}$-rational series on a free monoid

Deeper & broader

5. Series with coefficients in a continuous semiring
6. Rational subsets in free products
7. A non-commutative linear algebra primer

Solutions to the exercises

Notes & references

Recall[5] that a semiring $\mathbb{K}$ is a set equipped with two operations, *addition* and *multiplication*, which satisfy the following axioms.

• SA1 $\mathbb{K}$ is a commutative monoid for addition, written $+$, whose neutral element is $0_{\mathbb{K}}$ (or 0).

• SA2 $\mathbb{K}$ is a monoid (not necessarily commutative) for multiplication, written by a dot, or more often by simple juxtaposition, and whose neutral element is written $1_{\mathbb{K}}$ (or 1).[6]

• SA3 Multiplication is left and right distributive over addition:

$$\forall i, j, k \in \mathbb{K} \qquad i \cdot (j + k) = (i \cdot j) + (i \cdot k) \quad \text{and} \quad (i + j) \cdot k = (i \cdot k) + (j \cdot k) \ .$$

• SA4 The neutral element for addition is a zero for multiplication:

$$\forall k \in \mathbb{K} \qquad k \cdot 0_{\mathbb{K}} = 0_{\mathbb{K}} \cdot k = 0_{\mathbb{K}} \ .$$

A semiring is called *commutative* when its multiplication is a commutative operation. A *ring* is a semiring that is a (commutative) group for addition. A ring all of whose non-zero elements have a multiplicative inverse (that is, the ring without 0 is a *group* for multiplication) is called a *skew field* or *division ring*. Such a ring, if it is *commutative*, is a *field*. *cf. Note 0.23, p. 29*

In what follows, M is a monoid and $\mathbb{K}$ a semiring, *a priori* arbitrary.

1.1 Formal power series over M with coefficients in $\mathbb{K}$

Definition 1.1 A *power series*[7] *over M with coefficients in* $\mathbb{K}$ is any map from M to $\mathbb{K}$. The set of these series is written $\mathbb{K}\langle\langle M \rangle\rangle$.

If s is a power series, the image of an element m of M under s is written

$$\langle s, m \rangle \ , \qquad \text{rather than} \qquad (m)s \ ,$$

and is called the *coefficient* of m *in* s. □

The set of maps from M to $\mathbb{K}$ is usually written $\mathbb{K}^M$. Since $\mathbb{K}$ is a semiring, this set is canonically equipped with the structure of a semiring, independently of the monoid structure of M, which corresponds to *pointwise* addition and multiplication (*cf.* below). But for this set of maps, which we wish to generalise the notion of *polynomials*, we will consider *another semiring structure* which derives from the monoid structure of M. This justifies our using different notation and terminology from those used for maps.[8]

[5] From the start of Section 0.5, reproduced here for the convenience of the reader.

[6] It would have been more accurate to write $\oplus$ and $\otimes$ for addition and multiplication in $\mathbb{K}$ to show that they are not necessarily the usual addition and multiplication on numbers. I preferred not to encumber the notation, but vigilance is therefore required; *cf.* Example 1.1 below.

[7] As indicated in the titles of the section and subsection, the adjective 'formal' is often prepended, but it adds little. We also generally omit the word 'power'.

[8] Eilenberg (*op. cit.*) calls power series *sets with multiplicity* and keeps the notation $\mathbb{K}^M$. The other authors cited in the introduction use the term *power series*.

Example 1.1 Take $\mathbb{K} = \mathbb{N}$ and $M = A^*$ with $A = \{a, b\}$. If we relate each word f in A^* to the number $|f|_a$ of occurrences of the letter a in f, we define a power series s_1 in $\mathbb{N}\langle\langle A^*\rangle\rangle$:

$$\forall f \in A^* \qquad \langle s_1, f\rangle = |f|_a \ .$$

Likewise, let t_1 be the power series in $\mathbb{N}\langle\langle A^*\rangle\rangle$ defined by $\langle t_1, f\rangle = |f|_b$ for all f in A^*.

If we take $\mathbb{K} = \mathcal{M} = \langle \mathbb{N} \cup \{+\infty\}, \mathsf{min}, + \rangle$ *the same coefficients* define two power series s'_1 and t'_1 in $\mathcal{M}\langle\langle A^*\rangle\rangle$: $\langle s'_1, f\rangle = |f|_a$ and $\langle t'_1, f\rangle = |f|_b$, for all f in A^*. □

Example 1.2 We take $\mathbb{K} = \mathbb{B}$ and an arbitrary monoid M. Let P be an arbitrary subset of M. If we relate to each m in M the coefficient $1_{\mathbb{B}}$ or $0_{\mathbb{B}}$ according to whether m belongs to P or not, we obtain a power series in $\mathbb{B}\langle\langle M\rangle\rangle$ and have thus defined a bijection[9] between $\mathfrak{P}(M)$ and $\mathbb{B}\langle\langle M\rangle\rangle$. □

Example 1.3 Let M and N be arbitrary monoids. Set $\mathbb{K} = \mathbb{B}\langle\langle N\rangle\rangle$. A power series σ in $[\mathbb{B}\langle\langle N\rangle\rangle]\langle\langle M\rangle\rangle$ relates a subset of N to each m in M: there is a bijection between $[\mathbb{B}\langle\langle N\rangle\rangle]\langle\langle M\rangle\rangle$ and the set of relations from M to N, or $\mathbb{B}\langle\langle M \times N\rangle\rangle$. □

Operations on $\mathbb{K}\langle\langle M\rangle\rangle$

For all s and t in $\mathbb{K}\langle\langle M\rangle\rangle$, and all k in $\mathbb{K}$, we define the following operations:

(i) *'exterior' multiplication* (left and right) by a constant:

$$k\,s \quad \text{with} \quad \forall m \in M \quad \langle k\,s, m\rangle = k\langle s, m\rangle \quad \text{and}$$
$$s\,k \quad \text{with} \quad \forall m \in M \quad \langle s\,k, m\rangle = \langle s, m\rangle k \ ,$$

(ii) pointwise *addition*:

$$s + t \quad \text{with} \quad \forall m \in M \quad \langle s + t, m\rangle = \langle s, m\rangle + \langle t, m\rangle \ ,$$

(iii) and the *Cauchy product*:[10]

$$s\,t \quad \text{with} \quad \forall m \in M \quad \langle s\,t, m\rangle = \sum_{\substack{u,v \in M \\ uv = m}} \langle s, u\rangle\langle t, v\rangle \ . \qquad (*)$$

Addition makes $\mathbb{K}\langle\langle M\rangle\rangle$ a commutative monoid, whatever $\mathbb{K}$ and M; together with the two exterior multiplications, it makes $\mathbb{K}\langle\langle M\rangle\rangle$ a left and right *module* on $\mathbb{K}$.

cf. Sec. 7, p. 488

Example 1.1 (continued) For all f in A^* we have

$$\langle s_1 + t_1, f\rangle = |f| \quad \text{and} \quad \langle s'_1 + t'_1, f\rangle = \mathsf{min}\{|f|_a, |f|_b\} \ . \qquad \square$$

The Cauchy product is problematic: there could very well exist an m in M such that, for those u, v satisfying $u\,v = m$, the set of pairs (u, v) is infinite, and hence

[9] The reader may wonder why we do not denote series, like languages, by capital letters. We do not do so in order to emphasise that a power series is an *element* (of $\mathbb{K}\langle\langle M\rangle\rangle$) whereas a language is a *set* of words.

[10] It is called 'Cauchy product' to distinguish it from the two other products we shall describe later: the Hadamard product and scalar product.

there could exist power series such that the sum on the right hand side of $(*)$ is undefined. Thus, we cannot ensure, without further assumptions, that the Cauchy product is a total operation on $\mathbb{K}\langle\langle M\rangle\rangle$. We can overcome this difficulty in at least two ways. *cf. Exer. 1.3*

The first is to retreat: we no longer consider $\mathbb{K}\langle\langle M\rangle\rangle$ but only the set, written $\mathbb{K}\langle M\rangle$, of power series that have only a *finite number of non-zero coefficients*, which we call 'polynomials'. The sum in $(*)$ is infinite but only a finite number of terms are non-zero; the Cauchy product is defined on $\mathbb{K}\langle M\rangle$ and is in fact a semiring (a semialgebra on $\mathbb{K}$), a sub-semiring of $\mathbb{K}\langle\langle M\rangle\rangle$ when that is defined.

The second is to assume that $\mathbb{K}$ is *complete*: every sum, even if infinite, is defined on $\mathbb{K}$, and the Cauchy product of two series is defined for any M. This is the case, for example, if $\mathbb{K}$ is equal to $\mathbb{B}$, $\mathbb{B}\langle\langle M\rangle\rangle$, $\mathcal{N}$ or $\mathcal{M}$. We shall take this path in Section 5. *cf. Sec. 0.5* We should however immediately note that for every monoid M, $\mathfrak{P}(M)$ and $\mathbb{B}\langle\langle M\rangle\rangle$ are not only in bijection but are also *isomorphic* as semirings, and we will use both notations from here on, whichever is more convenient.

It seems more useful initially to be able to define weighted automata, and hence power series, *without restriction on* $\mathbb{K}$, and we are led in this case to make assumptions about M: we suppose for the rest of this section (and the three subsequent sections) that the monoids are *graded*, a condition that we shall describe in the next subsection and which allows the natural generalisation of the standard construction of formal power series of a single variable.[11] First, some more definitions.

Support of a series – characteristic series

The *support* of a power series s, written $\mathsf{supp}(s)$ or $\mathsf{supp}\, s$, is the set of elements of M which have a non-zero coefficient in s. We have already said that the *polynomials* are those series whose support is finite.

Example 1.1 (continued) We have $\mathsf{supp}\, s_1 = A^* a\, A^*$ and $\mathsf{supp}\, s_1' = A^*$ since 0 is *not* the zero of $\mathcal{M}$. □

Conversely, for each subset P of M ($\mathbb{K}$ being fixed, and in general equal to $\mathbb{N}$) there is a corresponding *characteristic series*, written $\underline{P}$, and defined as the series whose support is P and all of whose non-zero coefficients are $1_{\mathbb{K}}$:

$$\forall m \in M \qquad \langle \underline{P}, m\rangle = \begin{cases} 1_{\mathbb{K}} & \text{if} \quad m \in P \\ 0_{\mathbb{K}} & \text{otherwise} \end{cases}$$

We will also say (usually when $\mathbb{K} = \mathbb{N}$) that a series is *characteristic* or *unambiguous* if all its non-zero coefficients are $1_{\mathbb{K}}$; that is, if it is the characteristic series of its support.

[11] A fourth method exists, inspired by the first and third. It involves making an assumption about M (we require it to be an *ordered group*) and considering only a subset of $\mathbb{K}\langle\langle M\rangle\rangle$ (those power series whose support is well ordered). We shall describe and use this construction (equally standard, in fact) in Chapter IV.

cf. Sec. II.1.4, p. 228

This notation also leads to a concise (and useful) definition of unambiguous rational operations (that of the unambiguous star is postponed to the next section):

Properties 1.1 *Set $\mathbb{K} = \mathbb{N}$ and let P and Q be two subsets of a (graded) monoid M.*

(i) *The union $P \cup Q$ is unambiguous if and only if $\underline{(P \cup Q)} = \underline{P} + \underline{Q}$.*

(ii) *The product PQ is unambiguous if and only if $\underline{(PQ)} = \underline{P}\,\underline{Q}$.* ■

Remark 1.1 The monoid M is canonically a (multiplicative) submonoid of $\mathbb{K}\langle\langle M\rangle\rangle$: each m in M is identified with its characteristic polynomial $\underline{m}$. Likewise the semiring $\mathbb{K}$ is canonically a sub-semiring of $\mathbb{K}\langle\langle M\rangle\rangle$: each element k is identified with the polynomial $k\,\underline{1_M}$. □

Hadamard product

We will also use the product on series that is classically defined on maps in a monoid (here the multiplicative monoid of the semiring $\mathbb{K}$).

Definition 1.2 The *Hadamard product* of two power series s and t in $\mathbb{K}\langle\langle M\rangle\rangle$ is the series, written $s \odot t$, defined by

$$\forall m \in M \qquad \langle s \odot t, m\rangle = \langle s, m\rangle\langle t, m\rangle \ . \qquad \square$$

This product is commutative if $\mathbb{K}$ is commutative.

Example 1.1 (continued) It follows that $\langle s_1 \odot t_1, f\rangle = |f|_a\,|f|_b$ and $\langle s'_1 \odot t'_1, f\rangle = |f|$, for all f in A^*. □

The Hadamard product is the generalisation of set intersection to power series. We verify in particular that

$$\forall P, Q \subseteq M \qquad \underline{P} \odot \underline{Q} = \underline{(P \cap Q)} \ . \tag{1.1}$$

Scalar product

especially starting in Chapter IV

Finally, we will use the *scalar product* of two series s and t in $\mathbb{K}\langle\langle M\rangle\rangle$, written $\langle s, t\rangle$, and defined by the formula

$$\langle s, t\rangle = \sum_{m \in M} \langle s, m\rangle\langle t, m\rangle \ . \tag{1.2}$$

Clearly, the scalar product of two series s and t could easily be undefined, since the sum implied in (1.2) is potentially infinite. On the other hand, it is certainly defined if at least one of the two series s and t is a polynomial.

Remark 1.2 The fact that we have been writing the coefficient of a power series and the scalar product the same way is not a coincidence: the coefficient of m in s is the scalar product of s with $\underline{m}$, the characteristic series of m, which it is convenient to identify with m. □

Exercises

- 1.1 Verify that the Cauchy product indeed makes $\mathbb{K}\langle M\rangle$ a semiring, and $\mathbb{K}\langle\langle M\rangle\rangle$ too, if $\mathbb{K}$ is complete.

- 1.2 Recognise[12] the semirings $\mathbb{Z}\langle\mathbb{N}\rangle$, $\mathbb{Q}\langle\langle\mathbb{N}\rangle\rangle$, $\mathbb{R}\langle\langle\mathbb{N}^2\rangle\rangle$.

- 1.3 (a) Verify that the Cauchy product is not defined on $\mathbb{Q}\langle\langle\mathbb{Z}\rangle\rangle$.

  (b) Recognise $\mathbb{Q}_{\mathsf{wo}}\langle\langle\mathbb{Z}\rangle\rangle$, the subset[13] of $\mathbb{Q}\langle\langle\mathbb{Z}\rangle\rangle$ consisting of the power series whose support has a least element (in $\mathbb{Z}$).

1.4 Verify Property 1.1.

1.5 Verify Equation (1.1).

1.2 Graded monoids

For the Cauchy product to be always defined on $\mathbb{K}\langle\langle M\rangle\rangle$, independently of $\mathbb{K}$, it is necessary (and sufficient) that, for all m in M, the set of pairs (u, v) such that $u\,v = m$ is finite – we will say that m is *finitely decomposable*. However, making $\mathbb{K}\langle\langle M\rangle\rangle$ a semiring is not an end in itself: power series in $\mathbb{K}\langle\langle M\rangle\rangle$ will replace subsets of M as the behaviours of automata. If we want to extend to power series the characterisation of the behaviour of finite automata by rational operations – a fundamental theorem – then not only must sum and product be defined on the series, but so must the star operation, which implies an *infinite sum*. This forces us to have some sort of *topology* on $\mathbb{K}\langle\langle M\rangle\rangle$, to which we shall return in the next subsection.

cf. Sec. 2.4 infra

The construction of power series on A^*, which generalises that of series of one variable, shows that it is from the *length* of words in A^* that we build a topology on $\mathbb{K}\langle\langle A^*\rangle\rangle$. The existence of an *additive length* is the main assumption that we shall make about M. Returning to the initial problem, we then seek an additional condition that ensures that every element is finitely decomposable. For reasons of simplicity,[14] we assume that M is *finitely generated*. This solves the problem, while allowing us to deal with the cases that interest us.

Definition 1.3 Let M be a monoid. A function $\varphi\colon M \to \mathbb{N}$ is a *length* on M if:

(i) $\varphi(m)$ is *strictly* positive for all m other than 1_M ;

(ii) $\forall m, n \in M \qquad \varphi(m\,n) \leqslant \varphi(m) + \varphi(n)$.

We shall say that a length is a *gradation* if it is *additive*; that is, if:

(iii) $\forall m, n \in M \qquad \varphi(m\,n) = \varphi(m) + \varphi(n)$;

and that M is *graded* if it is equipped with a gradation.[15] □

cf. Sec. 0.4, p. 26

[12] In the way one recognises an old friend in the street.

[13] This is the notation that we will use in Sections IV.4.3 and IV.7

[14] Simplicity for the expression of the distance that defines the topology on $\mathbb{K}\langle\langle M\rangle\rangle$, and for the summability of locally finite families (see below). Note in any case that modulo certain additional precautions we can work with, for example, power series of $\mathbb{K}\langle\langle A^*\rangle\rangle$, even if A^* is not finitely generated.

[15] This definition, while not entirely standard, has the advantage of not contradicting the standard definition of the length of an element of the free group (as the length of the equivalent reduced word).

Every free monoid and every cartesian product of free monoids is graded. The definition implies that $\varphi(1_M) = 0$ and that a finite monoid, more generally a monoid that contains an idempotent other than the identity (for example, a zero), cannot be equipped with a gradation.

If M is graded, $M \setminus 1_M$ is a sub-semigroup; in other words, M has no divisors of the identity.

Proposition 1.1 *In a graded monoid of finite type, the number of elements whose length is less than an arbitrary given integer n is finite.*

Proof. Let G be a finite set of generators of M. The length of each generator is at least 1. Hence we deduce that if the length of m is not greater than n, m can be written as the product of at most n generators. There are at most $(\|G\|^{n+1} - 1)/(\|G\| - 1)$ such elements. ■

In other words, every element of a graded monoid M can only be written in a finite number of different ways as the product of elements of M other than 1_M.[16] We can deduce in particular:

Corollary 1.2 *In a graded monoid of finite type, every element is finitely decomposable.* ■

Corollary 1.3 *A graded monoid of finite type is countable.* ■

Note that a finite monoid is not graded, but that every element is nonetheless finitely decomposable. From Corollary 1.2, we deduce the proposition aimed at by Definition 1.3:

Proposition 1.4 *Let M be a graded monoid of finite type. If $\mathbb{K}$ is a semiring, then $\mathbb{K}\langle\langle M\rangle\rangle$, equipped with the Cauchy product, is one too and, what is more, a (left and right) semialgebra*[17] *on $\mathbb{K}$.* ■

In the following (up to Section 5), M is a graded monoid that is implicitly assumed to be of finite type. To simplify the notation and in imitation of the free monoid, we will write the length function as a pair of vertical bars: that is, $|m|$ rather than $\varphi(m)$.

Matrix of series and series of matrices

Let Q be a finite set; since $\mathbb{K}\langle\langle M\rangle\rangle$ is a semiring, so is $\mathbb{K}\langle\langle M\rangle\rangle^{Q\times Q}$. It is easy to verify a property which we shall use constantly: that this last is isomorphic to the semiring of power series on M with coefficients in $\mathbb{K}^{Q\times Q}$, the semiring of square matrices of dimension Q with entries in $\mathbb{K}$; that is:

[16] Eilenberg, *op. cit.*, calls a monoid with this property a *locally finite monoid*. This terminology inconveniently conflicts with another accepted meaning of the phrase: a monoid such that every finitely generated submonoid is finite (*cf.* [235]).

[17] If $\mathbb{K}$ is a ring, $\mathbb{K}\langle\langle M\rangle\rangle$ is even what is classically called a *graded algebra*, which is the origin of the terminology chosen for graded monoids.

Lemma 1.5 *For a graded monoid M, $\mathbb{K}\langle\langle M\rangle\rangle^{Q\times Q} \cong \mathbb{K}^{Q\times Q}\langle\langle M\rangle\rangle$.* ■

Example 1.4 Again we take $\mathbb{K} = \mathbb{N}$ and $M = A^*$. Let

$$m = \begin{pmatrix} a+b & b \\ 0 & a+b \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} a + \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} b \ ;$$

then
$$m^2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} a^2 + \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} a\,b + \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} b\,a + \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} b^2$$

$$= \begin{pmatrix} (a+b)^2 & a\,b + b\,a + 2\,b^2 \\ 0 & (a+b)^2 \end{pmatrix} \ .$$ □

Exercises

• 1.6 (a) Show that a graded monoid has no idempotents other than 1_M.

(b) Give an example of a graded monoid that is not finitely decomposable.

1.7 (a) Verify that the direct product of two graded monoids is graded.

(b) Let M be a graded monoid; show that the set $G = M \setminus M^2$ is a minimum generating set (that is, contained in every generating set of M). Let A be a set in bijection with G and α the morphism from A^* to M defined by this bijection. Show that the map α is of finite co-image (that is, the inverse image of every element of M is finite).

(c) Let $\mu\colon M \to N$ be a morphism between two graded monoids. Show that μ is of finite co-image if and only if $(1_N)\mu^{-1} = 1_M$.

• 1.8 Let A be a finite alphabet.

(a) Show that a monoid M presented by $M = \langle\, A\,;\, \{u_i = v_i \mid i \in I\}\,\rangle$ with $|u_i| = |v_i|$ for all i in I is a graded monoid.

(b) Show that a monoid M presented by $M = \langle\, A\,;\, u = 1_{A^*}\,\rangle$ is not a graded monoid, whatever the word u is.

(c) Give an example of a non-graded monoid M that is presented by $M = \langle\, A\,;\, u = v\,\rangle$ with $|u| > |v| > 0$.

1.9 Verify the isomorphism stated in Lemma 1.5.

1.10 Compute, with the method of Example 1.4, the square and cube of the matrix:

$$m = \begin{pmatrix} a+b & b \\ 0 & 2(a+b) \end{pmatrix} \ .$$

1.3 Topology on $\mathbb{K}\langle\langle M\rangle\rangle$

The concept of a graded monoid has allowed us to avoid infinite sums in the definition of the Cauchy product, but the definition of star, itself an infinite sum, forces us to confront the problem, with no possible escape. The classic method assumes that we know how to express, even before defining such sums, that a sequence $\{s_n\}_{n\in\mathbb{N}}$ of

power series of $\mathbb{K}\langle\langle M\rangle\rangle$ 'tends', or 'converges', to a series s; that is, that we have a *topology* on $\mathbb{K}\langle\langle M\rangle\rangle$. I do not assume that the reader has made even an elementary study of general topology; nor it is my intention to remedy this. Nevertheless, it is there that the answer is found.

Now is a good time to recall that $\mathbb{K}\langle\langle M\rangle\rangle = \mathbb{K}^M$ is the *set of maps* from M to $\mathbb{K}$, and that if we have a topology on $\mathbb{K}$, this set $\mathbb{K}^M$ is naturally equipped with the 'simple convergence topology':

$$s_n \ \textit{tends to} \ s \quad \text{if for all } m \text{ in } M, \quad \langle s_n, m\rangle \ \textit{tends to} \ \langle s, m\rangle,$$

also, that this topology coincides with the *product topology* on $\mathbb{K}^M$ and finally, that if $\mathbb{K}$ is a *metric space* (that is, its topology is defined by a distance) then $\mathbb{K}^M$ is also a metric space, since M is countable.

We have said everything, or nearly, and at the same time we have said nothing. We shall therefore reexamine the topology question using only the notion of *distance*, more in line with intuition. We shall explain how to define a distance between two power series under the assumption that M is graded, and, naturally, we shall say that a sequence $\{s_n\}_{n\in\mathbb{N}}$ tends to s if the distance between s_n and s tends to 0 when n tends to infinity.

The foregoing discussion was nevertheless worthwhile, as it made clear that *the basis of the topology on* $\mathbb{K}\langle\langle M\rangle\rangle$ *is the topology on* $\mathbb{K}$.

1.3.1 Distance

A *distance* on a set E is a map $\mathbf{d}$ which relates to every pair (x, y) of elements of E a *positive real number* $\mathbf{d}(x, y)$, called the *distance from x to y* (or *between x* and y), which satisfies the following properties:

- symmetry: $\mathbf{d}(x, y) = \mathbf{d}(y, x)$
- positivity: $\mathbf{d}(x, y) > 0$ if $x \neq y$ and $\mathbf{d}(x, x) = 0$
- triangular inequality: $\mathbf{d}(x, y) \leqslant \mathbf{d}(x, z) + \mathbf{d}(y, z)$

When this triangular inequality can be replaced by

- $\forall x, y, z \in E$ $\quad \mathbf{d}(x, y) \leqslant \mathsf{max}\{\mathbf{d}(x, z), \mathbf{d}(y, z)\}$,

the distance $\mathbf{d}$ is called *ultrametric*.

A sequence $\{x_n\}_{n\in\mathbb{N}}$ of elements of E *converges* to x, or *tends* to x, if the distance between x_n and x becomes arbitrarily small as n grows; that is, more formally, if for every positive real ε there exists an integer N such that for all n greater than N, the distance between x_n and x is less than ε:

$$\forall \varepsilon > 0 \quad \exists N \in \mathbb{N} \quad \forall n \geqslant N \qquad \mathbf{d}(x_n, x) \leqslant \varepsilon \ .$$

Such an element x is *unique*; it is called the *limit* of the series $\{x_n\}_{n\in\mathbb{N}}$ and we write

$$x = \lim_{n\to+\infty} x_n \ ,$$

or simply $x = \lim x_n$ if there is no ambiguity. We say that $\mathbf{d}$ *equips E with a topology.*

Remark 1.3 We can always assume that a distance is a real number less than or equal to 1. If that is not the case, then by taking

$$\mathbf{f}(x,y) = \mathsf{inf}\{\mathbf{d}(x,y), 1\} \ ,$$

we obtain a distance $\mathbf{f}$ on E that defines *the same topology*; that is, a distance for which *the same sequences* will converge to *the same limits*. □

Remark 1.4 Whatever E is, we can choose a trivial distance function which is 1 for every pair of distinct elements. This is equivalent to saying that two distinct elements are never 'close' to each other, and that the only convergent sequences are those that are *stationary* above a certain rank. We then say that E is equipped with the *discrete topology*. □

1.3.2 Distance on $\mathbb{K}\langle\langle M\rangle\rangle$

To construct a topology on $\mathbb{K}\langle\langle M\rangle\rangle$, we must start with the topology on $\mathbb{K}$. We are confronted with two situations which seem fundamentally different. The first is that of a semiring $\mathbb{K}$ such as $\mathbb{B}$, $\mathbb{N}$, $\mathbb{Z}$, $\mathcal{N}$, *etc.* whose elements are 'detached' from each other. The natural topology on these semirings is the discrete topology. The second is that of semirings such as $\mathbb{Q}$, $\mathbb{Q}_+$, $\mathbb{R}$, *etc.*, or even, later, $\mathbb{K}\langle\langle M\rangle\rangle$ itself, which can act as a semiring of coefficients for power series on another monoid; that is, semirings on which there is *a priori* a distance which can be arbitrarily small. We have just observed that even on these semirings one can always choose a discrete topology, but it is more satisfactory to preserve their 'native' topology.

By means of the definition of a distance and the topological notions derived from it, we could treat in the same way these two situations that intuitively seem radically different. We shall not do this immediately; we shall return to a uniform treatment starting in the next subsubsection.

Discrete topology on $\mathbb{K}$ The length function on M allows a natural definition of a *gap* $\mathbf{e}$ between two series s and t in $\mathbb{K}\langle\langle M\rangle\rangle$, which is the minimal length of an element of M for which s and t differ; that is,

$$\mathbf{e}(s,t) = \mathsf{min}\{n \in \mathbb{N} \mid \exists m \in M \quad |m| = n \quad \text{and} \quad \langle s,m\rangle \neq \langle t,m\rangle\} \ , \tag{1.3}$$

and we verify that the function $\mathbf{d}$ defined by:

$$\mathbf{d}(s,t) = 2^{-\,\mathbf{e}(s,t)} \tag{1.4}$$

is indeed a *distance* in the sense of the previous paragraph, bounded by 1, and that it is an *ultrametric distance*.

Remark 1.5 A more classical notion allowed by the length function on M is the *valuation* of a series: the valuation of s is the minimum of the lengths of the elements whose coefficient in s is non-zero; that is, the gap between s and the zero series:

$$\mathbf{v}(s) = \mathsf{min}\{|m| \mid \langle s,m\rangle \neq 0\} = \mathsf{min}\{|m| \mid m \in \mathsf{supp}\, s\} \ .$$

Conversely, and in the case where $\mathbb{K}$ *is a ring*, the gap between s and t is the valuation of the difference $s-t$. □

$\mathbb{K}$ is equipped with a distance c We first assume that the distance $\mathbf{c}$ between two elements of $\mathbb{K}$ is less than or equal to 1. The length function on M allows us to order the elements of M and we set

$$\mathbf{d}'(s,t) = \frac{1}{2}\sum_{n\in\mathbb{N}}\left(\frac{1}{2^n}\,\mathsf{max}\{\mathbf{c}(\langle s,m\rangle,\langle t,m\rangle)\;\big|\;|m|=n\}\right) \quad . \tag{1.5}$$

We then verify that $\mathbf{d}'$ is indeed a distance on $\mathbb{K}\langle\langle M\rangle\rangle$.

Whatever the distance $\mathbf{c}$ (bounded by 1) on $\mathbb{K}$, the distance $\mathbf{d}'$ between two series in $\mathbb{K}\langle\langle M\rangle\rangle$ which coincide for all m of length less than or equal to n is less than 2^{-n}. The two distances $\mathbf{d}$ and $\mathbf{d}'$ which we have just defined on $\mathbb{K}\langle\langle M\rangle\rangle$ – namely $\mathbf{d}$ from the gap (hence from the discrete topology on $\mathbb{K}$) and $\mathbf{d}'$ from $\mathbf{c}$ – are hence such that

$$\forall s,t\in\mathbb{K}\langle\langle M\rangle\rangle \qquad \mathbf{d}'(s,t)\leqslant\mathbf{d}(s,t)\leqslant 1 \quad . \tag{1.6}$$

If $\mathbf{c}$ is the distance associated with the discrete topology (that is, $\mathbf{c}(s,t)=1$ if $s\neq t$), it also follows that

$$\forall s,t\in\mathbb{K}\langle\langle M\rangle\rangle \qquad \frac{1}{2}\,\mathbf{d}(s,t)\leqslant\mathbf{d}'(s,t) \quad , \tag{1.7}$$

which means that *if* $\mathbb{K}$ *is equipped with the discrete topology* the two distances $\mathbf{d}$ and $\mathbf{d}'$ defined respectively by (1.4) and (1.5) are *equivalent*; that is, a sequence $\{s_n\}_{n\in\mathbb{N}}$ converges to s for one distance if and only if it converges to s for the other, and from now on there is no reason to distinguish the two constructions; the distance on $\mathbb{K}\langle\langle M\rangle\rangle$ will be denoted $\mathbf{d}$.

cf. Exer. 1.16 **The last word** The topology defined on $\mathbb{K}\langle\langle M\rangle\rangle$ by $\mathbf{d}$ is, as stated, the *simple convergence* topology; that is:

Property 1.2 *A sequence $\{s_n\}_{n\in\mathbb{N}}$ of series of $\mathbb{K}\langle\langle M\rangle\rangle$ converges to s if and only if for all m in M $\langle s_n,m\rangle$ tends to $\langle s,m\rangle$.* ■

Furthermore, choosing a topology on a semiring only really makes sense if the constituent operations of the semiring, addition and multiplication, are consistent with the topology – we say they are *continuous* – that is, if the limit of a sum (resp. of a product) is the sum (resp. the product) of the limits. We say in this case that not only is the semiring equipped with a topology, but that it is a *topological semiring.*
cf. Exer. 1.17 We easily verify that if $\mathbb{K}$ is topological, then so are $\mathbb{K}^{Q\times Q}$ and $\mathbb{K}\langle\langle M\rangle\rangle$; that is:

Property 1.3 *If $\mathbb{K}$ is a topological semiring, sum and product are* continuous *operations in $\mathbb{K}\langle\langle M\rangle\rangle$.* ■

In other words, if $\{s_n\}_{n\in\mathbb{N}}$ and $\{t_n\}_{n\in\mathbb{N}}$ are two convergent sequences of elements of $\mathbb{K}\langle\langle M\rangle\rangle$, we have

$$\mathsf{lim}(s_n + t_n) = (\mathsf{lim}\, s_n) + (\mathsf{lim}\, t_n) \quad \text{and} \quad \mathsf{lim}(s_n t_n) = (\mathsf{lim}\, s_n)\,(\mathsf{lim}\, t_n) \quad .$$

Note that the fact that the sequence $\{s_n + t_n\}_{n\in\mathbb{N}}$ or $\{s_n t_n\}_{n\in\mathbb{N}}$ converges says nothing about whether $\{s_n\}_{n\in\mathbb{N}}$ or $\{t_n\}_{n\in\mathbb{N}}$ converges or not.

1.3.3 Summable families

Let $\mathbb{T}$ be a semiring[18] equipped with a distance which makes it a topological semiring. We already know how to make precise the fact that an infinite sequence $\{t_n\}_{n\in\mathbb{N}}$ tends to a limit t when n tends to infinity. We must now give an equally precise meaning to the sum of an infinite family $\{t_i\}_{i\in I}$ and it turns out to be somewhat harder. The difficulty arises from the fact that we want a sort of *associativity–commutativity* extended 'to infinity' and hence ensure that the result and its existence does not depend on an arbitrary order existing on the set I of indices.

We shall therefore define an 'absolute' method of summability, and a family will be described as 'summable' if we can find an increasing sequence of finite sets of indices, a sort of 'hard core', such that not only do partial sums on these sets tend to a limit but above all that any sum on a finite set containing one of these kernels stays close to this limit. More precisely:

Definition 1.4 A family $\{t_i\}_{i\in I}$ of elements of $\mathbb{T}$ indexed by an arbitrary set I is called *summable* if there exists t in $\mathbb{T}$ such that, for all positive ε, there exists a finite subset J_ε of I such that, for all finite subsets L of I which contain J_ε, the distance between t and the sum of $\{t_i\}$ for i in L is less than ε; that is:

$$\exists t \in \mathbb{T}\,,\ \forall \varepsilon > 0\,,\ \exists J_\varepsilon \text{ finite}\,,\ J_\varepsilon \subset I\,,\ \forall L \text{ finite}\,,\ J_\varepsilon \subseteq L \subset I \qquad \mathbf{d}\left(\sum_{i\in L} t_i, t\right) \leqslant \varepsilon. \ \square$$

The element t thus defined is *unique* and is called the *sum* of the family $\{t_i\}_{i\in I}$. It is obviously equal to the usual sum if I is finite, and we write

$$t = \sum_{i\in I} t_i \quad .$$

From the definition of a summable family we easily deduce an associativity property restricted to finite groupings, but that repeats infinitely:

Lemma 1.6 *Let $\{t_i\}_{i\in I}$ be a summable family with sum t in $\mathbb{T}$. Let K be a set of indices and $\{J_k\}_{k\in K}$ a* partition *of I where all the J_k are* finite *(that is, $I = \bigcup_{k\in K} J_k$ and the J_k are pairwise disjoint). Set $s_k = \sum_{i\in J_k} t_i$. Then the family $\{s_k\}_{k\in K}$ is summable with sum t.* ■

[18] We have temporarily changed the symbol we use for a semiring on purpose: $\mathbb{T}$ will not only play the role of $\mathbb{K}$ but also of $\mathbb{K}\langle\langle M\rangle\rangle$ in what follows.

Returning to power series of $\mathbb{K}\langle\langle M\rangle\rangle$, we shall say that a family of series $\{s_i\}_{i\in I}$ is *locally finite* if for all m in M there is only a finite number of indices i such that $<s_i, m>$ is different from $0_{\mathbb{K}}$.

Property 1.4 *A locally finite family of power series is summable.*

This simple property is a good example of what the topological structure placed on $\mathbb{K}\langle\langle M\rangle\rangle$ imposes and adds. That we can *define a sum* for a locally finite family of series is trivial: pointwise addition is defined for each m, independently of any assumption about M. To say that the family is *summable* is to add extra information: it ensures that partial sums converge to the result of pointwise addition.

Proof of Property 1.4. Let $\{s_i\}_{i\in I}$ be a locally finite family in $\mathbb{K}\langle\langle M\rangle\rangle$ and s the series obtained by pointwise addition. For each integer n, let J_n be the set of indices of the series whose valuation is less than or equal to n. By assumption (that M is graded; that is why we need the assumption) J_n is finite. Let

cf. Prop. 1.1

$$t_n = \sum_{i\in J_n} s_i \ ,$$

and for all finite L which contain J_n, let

$$t_L = \sum_{i\in L} s_i \ .$$

The series t_n, t_L and s coincide for all the m of length less than or equal to n, and we have

$$\mathbf{d}(t_n, s) \leqslant 2^{-n} \quad \text{and} \quad \mathbf{d}(t_n, t_L) \leqslant 2^{-n} \ ,$$

from which follows $\mathbf{d}(t_L, s) \leqslant \mathbf{d}(t_n, s) + \mathbf{d}(t_n, t_L) \leqslant 2^{-(n-1)}$. ■

For all series s, the family of series $\{<s, m>m \mid m \in M\}$, where m is identified with its characteristic series, is locally finite, and we have

$$s = \sum_{m\in M} <s, m>m \ ,$$

cf. Exer. 1.18

which is the usual notation that is thus justified. We also deduce from this notation that $\mathbb{K}\langle M\rangle$ is *dense* in $\mathbb{K}\langle\langle M\rangle\rangle$.

Property 1.4 extends beyond locally finite families and generalises to a proposition which links the summability of a family of series and that of families of coefficients; its proof is left as an exercise.

cf. Exer. 1.19

Proposition 1.7 *A family $\{s_i\}_{i\in I}$ of $\mathbb{K}\langle\langle M\rangle\rangle$ is summable with sum s if and only if for each m in M the family $\{<s_i, m>\}_{i\in I}$ of elements of $\mathbb{K}$ is summable with sum $<s, m>$.* ■

1.3.4 Continuous morphisms

A morphism (on semirings) $\varphi\colon \mathbb{T} \to \mathbb{U}$ is a map that translates the structure of $\mathbb{T}$ into its image under $\mathbb{U}$; it is *continuous* if it is also consistent with the notion of limit. More precisely:

Definition 1.5 Let $\mathbb{T}$ and $\mathbb{U}$ be two semirings equipped respectively with distances $\mathbf{d}$ and $\mathbf{c}$ which make them two topological semirings. A morphism[19] $\varphi\colon \mathbb{T} \to \mathbb{U}$ is *continuous* if we have

$$\forall \varepsilon > 0\,,\ \forall x \in \mathbb{T}\,,\ \exists \eta > 0\,,\ \forall y \in \mathbb{T} \qquad \mathbf{d}(x,y) \leqslant \eta \Rightarrow \mathbf{c}(x\varphi, y\varphi) \leqslant \varepsilon \ . \tag{1.8}$$

Similarly, φ is continuous if and only if for all convergent sequences $\{t_n\}$ in $\mathbb{T}$, the sequence $t_n\varphi$ is convergent and $\lim\,(t_n\varphi) = (\lim t_n)\,\varphi$. It should be noted that in (1.8), η depends on ε *and* on x; if it could be chosen independently of x, we would say that the morphism was *uniformly continuous*.

The following properties, whose verification, a formality using the definitions, is left as an exercise, shows the robustness of the notion of continuity and the consistency of the distance chosen on $\mathbb{K}\langle\langle M\rangle\rangle$. The semirings $\mathbb{K}$, $\mathbb{L}$, $\mathbb{T}$ and $\mathbb{U}$ are assumed to be equipped with distances that make them topological semirings.

Property 1.5 *Let $\varphi\colon \mathbb{T} \to \mathbb{U}$ be a continuous morphism. If $\{t_i\}_{i\in I}$ is a summable family with sum t in $\mathbb{T}$, then the family $\{(t_i)\varphi\}_{i\in I}$ is a summable family in $\mathbb{U}$ with sum $t\varphi$.* ■

A morphism of semirings $\varphi\colon \mathbb{K} \to \mathbb{L}$ is extended canonically for every graded monoid M to a map $\varphi\colon \mathbb{K}\langle\langle M\rangle\rangle \to \mathbb{L}\langle\langle M\rangle\rangle$ defined for every s of $\mathbb{K}\langle\langle M\rangle\rangle$ by

$$\forall m \in M \qquad \langle s\varphi, m\rangle = \langle s, m\rangle\varphi \ .$$

It is easy to verify that it is a morphism of semirings. It then follows that:

Property 1.6 *If $\varphi\colon \mathbb{K} \to \mathbb{L}$ is a continuous morphism, so is the morphism $\varphi\colon \mathbb{K}\langle\langle M\rangle\rangle \to \mathbb{L}\langle\langle M\rangle\rangle$.* ■

A distance on $\mathbb{K}$ naturally induces a distance on $\mathbb{K}^{Q\times Q}$ by taking the maximum of the distances for each coefficient. *cf. Lem. 1.5, p. 385*

Property 1.7 *The canonical isomorphism between $\mathbb{K}\langle\langle M\rangle\rangle^{Q\times Q}$ and $\left[\mathbb{K}^{Q\times Q}\right]\langle\langle M\rangle\rangle$ is (bi-)continuous.*[20] ■

Exercises In the following exercises, $\mathbb{K}$ is a topological semiring and M a graded monoid.

1.11 Prove Lemma 1.6.

1.12 Let $\{t_i\}_{i\in I}$ be a summable family with sum t in $\mathbb{T}$ and s an element of $\mathbb{T}$. Show that the family $\{s\,t_i\}_{i\in I}$ is summable with sum $s\,t$.

---

[19] Beware! this is not the length.

[20] That is, this isomorphism and its inverse are continuous.

1.13 Calculate the series $\underline{A^*}\,\underline{f}\,\underline{A^*}$, for all f in A^*.

1.14 Verify that $\mathbf{d}$ is indeed an ultrametric distance on $\mathbb{K}\langle\langle M\rangle\rangle$ when $\mathbf{c}$ is the distance of the discrete topology or, more generally, an ultrametric distance.

• 1.15 Explain why the definition of the distance $\mathbf{d}'$ in (1.5) cannot be replaced by $\mathbf{f}(s,t) = \big(\mathsf{max}\{\mathbf{c}(\langle s,m\rangle,\langle t,m\rangle) \mid |m| = \mathbf{e}(s,t)\}\big)\, 2^{-\,\mathbf{e}(s,t)}$.

• 1.16 **Simple convergence.** Verify that the distance defined on $\mathbb{K}\langle\langle M\rangle\rangle$ is indeed such that a sequence $\{s_n\}_{n\in\mathbb{N}}$ of series of $\mathbb{K}\langle\langle M\rangle\rangle$ converges to s if and only if, for all m in M, $\langle s_n,m\rangle$ tends to $\langle s,m\rangle$ (Property 1.2).

1.17 Verify that $\mathbb{K}\langle\langle M\rangle\rangle$ is indeed a topological semiring.

1.18 Verify that every series of $\mathbb{K}\langle\langle M\rangle\rangle$ is the limit of a sequence of polynomials of $\mathbb{K}\langle M\rangle$ (that is, $\mathbb{K}\langle M\rangle$ is *dense* in $\mathbb{K}\langle\langle M\rangle\rangle$).

• 1.19 Prove Proposition 1.7.

• 1.20 **Uniform convergence.** Let $\mathbf{c}$ be a distance on $\mathbb{K}$. We say that a sequence $\{s_n\}_{n\in\mathbb{N}}$ of series of $\mathbb{K}\langle\langle M\rangle\rangle$ *converges uniformly* to s if all the sequences $\langle s_n,m\rangle$ tend *together* to $\langle s,m\rangle$; that is, if

$$\forall\varepsilon > 0\,,\ \exists N = N(\varepsilon)\in\mathbb{N}\,,\ \forall m\in M\,,\ \forall n \geqslant N \qquad \mathbf{c}(\langle s_n,m\rangle,\langle s,m\rangle) < \varepsilon\ .$$

We say that s is the *uniform limit* of the $\{s_n\}$ to mean that the sequence $\{s_n\}_{n\in\mathbb{N}}$ converges uniformly to s.

(a) Show that if the topology of $\mathbb{K}$ is discrete, the uniform limit of a sequence of polynomials is a polynomial.

(b) Suppose that $\mathbb{Q}$ is equipped with the usual distance. Characterise those series of $\mathbb{Q}\langle\langle M\rangle\rangle$ that are the uniform limit of polynomials of $\mathbb{Q}\langle M\rangle$.

1.21 **Morphisms of semirings of series.** Let $\varphi\colon \mathbb{K}\to\mathbb{L}$ be a morphism of semirings.

(a) Verify that the map $\varphi\colon \mathbb{K}\langle\langle M\rangle\rangle \to \mathbb{L}\langle\langle M\rangle\rangle$, defined for all series s of $\mathbb{K}\langle\langle M\rangle\rangle$ by $\langle s\varphi,m\rangle = \langle s,m\rangle\varphi$ for all m in M, is a morphism of semirings.

See Propty 1.6 (b) Verify that if φ is continuous, so is the induced morphism $\varphi\colon \mathbb{K}\langle\langle M\rangle\rangle \to \mathbb{L}\langle\langle M\rangle\rangle$.

2 $\mathbb{K}$-automata and $\mathbb{K}$-rational power series

We shall now generalise the notion of an automaton over a monoid to that of an automaton *over a semiring of series* and, as before, characterise its behaviour. It is no surprise to learn that the behaviour of an automaton is a power series obtained by a sequence of sum, product and star operations which we shall call a ($\mathbb{K}$-)rational series. It is simpler to start by defining these last and then to tackle automata.

In what follows, $\mathbb{K}$ is an arbitrary semiring, and M a graded monoid: $\mathbb{K}\langle\langle M\rangle\rangle$ is a semiring.

2.1 Star of a power series

We are ready to begin: $\mathbb{K}\langle\langle M\rangle\rangle$ has already been equipped with a topology essentially with the aim of being able to define the star operation. We start by considering the problem in arbitrary semirings and not only in the semirings of power series.

2.1.1 Star in a topological semiring

Let $\mathbb{T}$ be a semiring equipped with a distance which makes it a topological semiring; that is, such that addition and multiplication are continuous operations. Let t be an element of $\mathbb{T}$; it is possible for the family $\{t^n\}_{n\in\mathbb{N}}$ to be, or not to be, summable. That is the question, which depends on t, on $\mathbb{T}$, on the distance on $\mathbb{T}$, or on a combination of all these elements. If it is summable, we call its sum the 'star of t' and write it t^*: *cf. Note 18, p. 389*

$$t^* = \sum_{n\in\mathbb{N}} t^n \quad ,$$

and we shall simply say that 'the star of t is defined' or that 't^* is defined'. If it is not summable, we shall say that 't^* is undefined'.

For example, $(0_{\mathbb{T}})^* = 1_{\mathbb{T}}$ is defined for all $\mathbb{T}$; if $\mathbb{T} = \mathbb{Q}$, we have $(\frac{1}{2})^* = 2$ if $\mathbb{Q}$ is equipped with the natural topology, or undefined if the chosen topology is the discrete topology, while 1^* is not defined in either case.

Notation We shall also use the following notation: for every integer n,

$$t^{\leqslant n} = t^0 + t^1 + \cdots + t^n \qquad \text{and} \qquad t^{<n+1} = t^{\leqslant n} \quad .$$

Example 2.1 The star of an arbitrary element of $\mathbb{N}$ (equipped with the discrete topology) other than 0 is undefined. The semiring $\mathbb{N}^{n\times n}$ of square matrices of dimension n is also equipped with the discrete topology. The star of every *strict upper triangular* matrix m is defined, since $m^* = m^{\leqslant n}$. □ *cf. Exer. 0.6.5, p. 31*

If t^* is defined then

$$t^* = \lim_{n\to\infty} \sum_{k\leqslant n} t^k = \lim_{n\to\infty} t^{\leqslant n} \quad , \tag{2.1}$$

and it is often in this form that we shall evaluate the expressions in which star is implied.[21]

Remark 2.1 The case $\mathbb{T} = \mathfrak{P}(M) = \mathbb{B}\langle\langle M\rangle\rangle$, the semiring of subsets of an arbitrary monoid M (and hence not necessarily finitely generated) in which we have defined the star of an arbitrary subset as the generated submonoid, as also the case $\mathbb{T} = \mathcal{N}$ in which $x^* = +\infty$ for all x other than 0, are not covered by this method of definition: these semirings are equipped with the discrete topology and, usually, the $P^{\leqslant n}$ are not stationary sequences, and thus not convergent sequences. These cases are treated in Section 5. *cf. p. 470*

[21] Formally, Equation (2.1) holds only if we have already established that the family $\{t^n\}_{n\in\mathbb{N}}$ is summable, a property which is not implied *a priori* by the fact that the sequence $t^{\leqslant n}$ converges.

The case $\mathbb{T} = \mathbb{B}$, and hence $\mathbb{T} = \mathbb{B}^{Q \times Q}$, in which every infinite sum reduces to a finite sum, are on the contrary covered by this method of definition, as is $\mathbb{T} = \mathbb{B}\langle\langle M \rangle\rangle$ equipped with the distance $\mathbf{d}$ when M is graded. □

Lemma 2.1 *Let $\mathbb{T}$ be a topological semiring and t an element of $\mathbb{T}$ whose star is defined. We have the double equality*

$$t^* = 1_{\mathbb{T}} + t\,t^* = 1_{\mathbb{T}} + t^* t \quad . \tag{2.2}$$

Proof. We obviously have

$$t^{\leqslant n} = 1_{\mathbb{T}} + t\,t^{<n} = 1_{\mathbb{T}} + t^{<n} t \quad .$$

As $\lim t^{<n} = \lim t^{\leqslant n} = t^*$, and as *addition and multiplication are continuous operations* on $\mathbb{T}$, we obtain (2.2) by taking the limit of each side of the above equation. ■

The same reasoning implies that if t^* is defined, we have

$$t\,t^* = t^* t = \lim_{n \to \infty} t\,(t^{\leqslant n}) \qquad \text{and we write} \qquad t^+ = \sum_{n \in \mathbb{N}_*} t^n = t\,t^* = t^* t \quad .$$

Sec. I.4.2, p. 128 We can rewrite Lemma 2.1 using the notion of *rational identity*:

Corollary 2.2 *The rational expressions interpreted in a topological semiring satisfy the identities $\mathbf{U_l}$ and $\mathbf{U_r}$ (and hence the system denoted by $\mathbf{U}$).* ■

Remark 2.2 If $\mathbb{T}$ is a topological *ring*, and if the star of t is defined, (2.2) can be written $t^* - t\,t^* = t^* - t^* t = 1$ or $(1-t)\,t^* = t^*(1-t) = 1$ and so t^* is the *inverse* of $1 - t$.[22] Hence the classic identity

$$t^* = \frac{1}{1-t} = 1 + t + t^2 + \cdots \quad , \tag{2.3}$$

is justified in full generality. □

Propty 1.5, p. 391 As a result of the definitions and properties of the last section, we have:

Property 2.1 *Let $\mathbb{T}$ and $\mathbb{U}$ be two topological semirings, and $\varphi\colon \mathbb{T} \to \mathbb{U}$ a continuous morphism. Let t be in $\mathbb{T}$. If t^* is defined, then $(t\varphi)^*$ is also and we have $(t\varphi)^* = (t^*)\varphi$.* ■

(Note that $(t\varphi)^*$ can be defined without t^* being defined: this is the case for instance if φ is the unique possible morphism from $\mathbb{N}$ to $\mathbb{B}$.)

[22] Hence the name '*quasi-inverse of t*' sometimes given to the element t^*, and that of *quasi-invertible* given to those elements whose star is defined; *cf.* for example the work of Salomaa and Soittola, *op. cit.*, or that of Kuich and Salomaa [142].

2.1.2 Star of a proper series

By reference to polynomials and to power series in one or more unknowns, we give the name *constant term* of a series s of $\mathbb{K}\langle\langle M\rangle\rangle$ to the coefficient of the neutral element of M in s: $\mathsf{c}(s) = \langle s, 1_M\rangle$.[23]

A power series is called *proper* if its constant term is zero. The sum of two proper series is a proper series; the product of two proper series (indeed, the product of a proper series with any other series) is a proper series, *since M is graded.*

Property 2.2 *If s is proper, the family $\{s^n \mid n \in \mathbb{N}\}$ is locally finite.*

Proof. Since M is graded, each element m cannot be equal to the product of a number of elements that is greater than the length of m, and $\langle s^n, m\rangle = 0_{\mathbb{K}}$ for all n greater than or equal to $|m|$. ■

Furthermore, since a locally finite family of series is summable, we have:

Proposition 2.3 *The star of a proper series of $\mathbb{K}\langle\langle M\rangle\rangle$ is defined.* ■

Arden's Lemma, stated for languages, and the identities deduced from it, extend immediately to proper series. *Lem. I.2.9, p. 100*

Lemma 2.4 [**Arden II**] *Let s and t be two series of $\mathbb{K}\langle\langle M\rangle\rangle$; if s is a proper series, the equations*

$$X = s\,X + t \qquad (2.4)$$

and

$$X = X\,s + t \qquad (2.5)$$

each has a unique solution*: the series s^*t and $t\,s^*$ respectively.*

Proof. According to Property 1.3, $\mathbb{K}\langle\langle M\rangle\rangle$ is a topological semiring. By multiplying (2.2) on the left (resp. on the right) by t we obtain that s^*t (resp. $t\,s^*$) is a solution of (2.4) (resp. of (2.5)).

Conversely, if u is a solution of the equation $X = t + s\,X$, we have

$$u = t + s\,u \quad\Longrightarrow\quad u = t + s\,(t + s\,u) = t + s\,t + s^2u = \cdots = s^{<n}t + s^n u \ ,$$

for all integers n. Since s is proper, and multiplication is continuous, we have

$$\lim_{n\to+\infty} s^n = \lim_{n\to+\infty} s^n u = 0 \ ,$$

from which we deduce $\quad u = \lim_{n\to+\infty}(s^{<n}t) = (\lim_{n\to+\infty} s^{<n})\,t = s^*t\,.$ ■

From Lemmas 2.1 and 2.4, we then obtain, as in Propositions I.4.6 and I.4.7 *and with the same proofs*, a set of equalities which are also identities satisfied by the rational expressions interpreted by proper series of $\mathbb{K}\langle\langle M\rangle\rangle$. *cf. p. 130*

[23] A language L over A^* is a series of $\mathbb{B}\langle\langle A^*\rangle\rangle$ and we indeed recover the definition of $\mathsf{c}(L)$ given in Section I.2.3, p. 89, apart from the detail explained in note I.25.

Proposition 2.5 *Let s and t be two proper series of $\mathbb{K}\langle\langle M\rangle\rangle$; we have the equalities*

$$(s+t)^* = s^*(t\,s^*)^* = (s^*t)^*s^* \quad , \tag{S}$$

$$(s\,t)^* = 1 + s\,(t\,s)^*t \quad , \tag{P}$$

$$\forall n \in \mathbb{N} \qquad s^* = s^{<n}(s^n)^* \quad . \tag{$\mathbf{Z}_n$}$$

see p. 385 **Remark 2.3** It follows by Lemma 1.5 that a square matrix m of dimension Q with elements in $\mathbb{K}\langle\langle M\rangle\rangle$ is a proper series of $\mathbb{K}^{Q\times Q}\langle\langle M\rangle\rangle$ if all its elements are proper series; (we say in this case that m is proper), and hence that the identities **S**, **P** and $\mathbf{Z}_n$ are satisfied by proper matrices. □

2.1.3 Star of an arbitrary series

The star of an arbitrary power series, not necessarily proper, may or may not be defined. The following proposition allows us to tell the difference between the two cases. First, we make a timely definition, to avoid a difficulty.

Definition 2.1 I will say that a topological semiring is *strong* if the product of two summable families is a summable family; that is, if $\{s_i \mid i \in I\}$ and $\{t_j \mid j \in J\}$ are two summable families with sum s and t respectively, the family $\{s_i\,t_j \mid (i,j) \in I\times J\}$ is summable with sum $s\,t$. □

All the semirings which we shall consider are strong:[24] semirings equipped with the discrete topology, the sub-semirings of $\mathbb{C}^n$ (equipped with the natural topology), and the positive semirings. We then easily verify:

Property 2.3 *The semirings of matrices and the semirings of series on a graded monoid, with coefficients in a strong semiring, are strong.* ■

Let s be a series of $\mathbb{K}\langle\langle M\rangle\rangle$; the *proper part* of s is the proper series that coincides with s for all the elements m of M other than 1_M. It is convenient to write $s_0 = \mathsf{c}\,(s)$ for the constant term of s, and s_p for the proper part of s:

$$\langle s_\mathsf{p}, 1_M\rangle = 0_\mathbb{K} \qquad \text{and} \qquad \forall\, m \in M \setminus 1_M \qquad \langle s_\mathsf{p}, m\rangle = \langle s, m\rangle \quad ,$$

and we write $s = s_0 + s_\mathsf{p}$ (rather than $s = s_0\,1_M + s_\mathsf{p}$).

Proposition 2.6 *Let $\mathbb{K}$ be a strong topological semiring and M a graded monoid. Let s be a series of $\mathbb{K}\langle\langle M\rangle\rangle$, s_0 its constant term and s_p its proper part. Then s^* is defined if and only if s_0^* is defined and in this case we have*

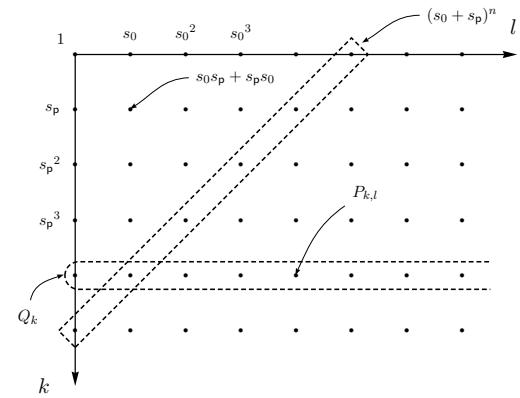
$$s^* = (s_0^*\,s_\mathsf{p})^*s_0^* = s_0^*(s_\mathsf{p}\,s_0^*)^* \quad . \tag{2.6}$$

Prop. 1.7, p. 390 *Proof.* The condition is necessary since $\langle s^n, 1_M\rangle = {s_0}^n$ and, if s^* is defined, the coefficients of 1_M in $\{s^n\}_{n\in\mathbb{N}}$ form a summable family.

[24] Definition 2.1 is timely in that it is sufficient to allow me to prove Proposition 2.6, at which we are aiming, without ruling out any semiring that I intend to consider. Indeed, the property that it states seems natural. I know no semiring which would not be strong (even if such could be imagined) and even less a semiring for which Proposition 2.6 would not be true.

Conversely, assume that $\{s_0^n\}_{n\in\mathbb{N}}$ is summable, with sum s_0^*. For all pairs of integers k and l, set

$$P_{k,l} = \sum_{\substack{i_0,i_1,\ldots,i_k\in\mathbb{N}\\ i_0+i_1+\cdots+i_k=l}} s_0^{i_0}\, s_{\mathsf{p}}\, s_0^{i_1}\, s_{\mathsf{p}} \cdots s_0^{i_{k-1}}\, s_{\mathsf{p}}\, s_0^{i_k} \quad .$$



By convention, set $P_{0,l} = s_0^l$ and $P_{k,0} = s_{\mathsf{p}}^k$. We verify by inspection that, for all integers n

$$s^n = (s_0 + s_{\mathsf{p}})^n = \sum_{l=0}^{l=n} P_{n-l,l} \quad . \tag{2.7}$$

By induction on k, we will show that the family

$$F_k = \{s_0^{i_0}\, s_{\mathsf{p}}\, s_0^{i_1}\, s_{\mathsf{p}} \cdots s_0^{i_{k-1}}\, s_{\mathsf{p}}\, s_0^{i_k} \mid i_0, i_1, \ldots, i_k \in \mathbb{N}\}$$

is summable in $\mathbb{K}\langle\langle M\rangle\rangle$, with sum

$$Q_k = (s_0^*\, s_{\mathsf{p}})^k\, s_0^* = s_0^*\, (s_{\mathsf{p}}\, s_0^*)^k \quad .$$

In fact, the hypothesis on s_0 ensures the property for $k = 0$, and also that the family $G = \{s_0{}^n\, s_{\mathsf{p}} \mid n \in \mathbb{N}\}$ is summable in $\mathbb{K}\langle\langle M\rangle\rangle$, with sum $s_0{}^* s_{\mathsf{p}}$. The family F_{k+1} is the product of the families G and F_k and the assumption that $\mathbb{K}$, and hence $\mathbb{K}\langle\langle M\rangle\rangle$, is strong gives us the conclusion. *cf. Exer. 1.12, p. 391*

Hence we deduce that, for each k, the family $\{P_{k,l} \mid l \in \mathbb{N}\}$ is summable, with sum Q_k. The family $\{Q_k \mid k \in \mathbb{N}\}$ is locally finite, hence summable, with sum *cf. Lem. 1.6, p. 389*

$$t = \sum_{k=0}^{\infty} Q_k = (s_0^*\, s_{\mathsf{p}})^*\, s_0^* = s_0^*\, (s_{\mathsf{p}}\, s_0^*)^* \quad .$$

We can now easily finish the proof by showing that the 'doubly indexed' family $\{P_{k,l} \mid k, l \in \mathbb{N}\}$ is summable, with sum t. Equation (2.7), and Lemma 1.6 once more, then ensure that the family $\{s^n \mid n \in \mathbb{N}\}$ is summable with sum t. ■ *cf. Exer. 2.5 infra*

In the same spirit as Remark 2.3, we note that (2.6) holds for every matrix m such that the star of its matrix of constant terms is defined. A particularly interesting case of this is where the matrix of constant terms is strict upper triangular.[25]

The definition of the star of a series also allows us to reformulate that of the *unambiguous star* of a subset of a monoid. *cf. Sec. II.1.4 and II.1.1*

Property 2.4 *Set $\mathbb{K} = \mathbb{N}$; the star of a subset P of a graded monoid M is* unambiguous *if and only if* $\underline{P^*} = (\underline{P})^*$. ■

Exercises

• 2.1 Let s and t be two proper series. Compute the star of the matrix $m = \begin{pmatrix} 0 & s \\ t & 0 \end{pmatrix}$.

[25] *Note added in translation:* This was an understatement of the French edition: this interesting case corresponds, among other things, to the *cycle-free series* of [142].

• 2.2 Compute $(a^*)^2$ and $(a^*)^3$ in $\mathbb{N}\langle\langle a^*\rangle\rangle$.

• 2.3 Let p be an integer; we define $a^{(p)}$ by induction: $a^{(0)} = a$ and $a^{(p+1)} = \left(a^{(p)}\right)^+$. Show that $a^{(p)} = (p\,a)^* a$. (We are still implicitly working in $\mathbb{N}\langle\langle a^*\rangle\rangle$.)

2.4 (a) Verify that the following semirings are strong: (i) the semirings equipped with the discrete topology; (ii) the sub-semirings of $\mathbb{C}$ (equipped with the natural topology); (iii) the positive semirings.

(b) Let $\mathbb{K}$ be a strong semiring. Show that $\mathbb{K}^{Q\times Q}$ and $\mathbb{K}\langle\langle M\rangle\rangle$ are strong, for all finite sets Q and every graded monoid M (Property 2.3).

2.5 Complete the proof of Proposition 2.6.

• 2.6 Suppose that $\mathbb{K}$ is a ring. Show that a series of $\mathbb{K}\langle\langle M\rangle\rangle$ is invertible if and only if its constant term is invertible.

2.7 Let $\{s_n\}_{n\in\mathbb{N}}$ be a sequence of proper series that tends to a limit s. Show that s is proper and that $\lim(s_n^*) = s^*$. What happens if we no longer assume the s_n to be proper?

2.8 Verify Property 2.4.

2.2 $\mathbb{K}$-rational series

We shall proceed in the same way as in the two preceding chapters: a characterisation of rational series 'from above' with the definition of rational operations then closed families, and an inductive characterisation 'from below', with the definition of weighted rational expressions. The subsection ends with a result that anticipates what follows and describes the star of a matrix.

2.2.1 $\mathbb{K}$-rational operations

The *rational operations* on $\mathbb{K}\langle\langle M\rangle\rangle$ are:

(i) the $\mathbb{K}$-*semialgebra operations*; that is:
- the two *exterior multiplications* by the elements of $\mathbb{K}$;
- *addition*;
- *product*;

(ii) the *star* operation, which is not everywhere defined.

Point (ii) leads us to tighten the notion of closure: a subset $\mathcal{E}$ of $\mathbb{K}\langle\langle M\rangle\rangle$ is *closed under star* if s^* belongs to $\mathcal{E}$ for every series s in $\mathcal{E}$ such that s^* is defined.

A subset of $\mathbb{K}\langle\langle M\rangle\rangle$ is *rationally closed* if it is closed under the rational operations; that is, if it is a sub-semialgebra of $\mathbb{K}\langle\langle M\rangle\rangle$ closed under the star operation. The intersection of any family of rationally closed subsets is rationally closed and thus the *rational closure* of a set $\mathcal{E}$ is the *smallest* rationally closed subset which contains $\mathcal{E}$, written $\mathbb{K}\mathrm{Rat}\,\mathcal{E}$.

Definition 2.2 A series of $\mathbb{K}\langle\langle M\rangle\rangle$ is $\mathbb{K}$-*rational* if it belongs to the rational closure of $\mathbb{K}\langle M\rangle$, the set of polynomials on M with coefficients in $\mathbb{K}$. The set of $\mathbb{K}$-rational series (over M with coefficients in $\mathbb{K}$) is written $\mathbb{K}\mathrm{Rat}\,M$. □

If the monoid M is implied by the context, we shall say $\mathbb{K}$-*rational series*, or just *rational series*, if $\mathbb{K}$ is also understood.

2.2.2 Rational $\mathbb{K}$-expressions

We can reuse, with adaptations, the material from Chapter I on rational expressions. In addition to the operators $\{0, 1, +, \cdot, *\}$, we define, for each k in $\mathbb{K}$, two unary operations, also written k. *cf. Def. I.4.1, p. 124*

Definition 2.3 A *weighted rational expression over M in* $\mathbb{K}$, or rational $\mathbb{K}$-expression over M, is any *well formed formula* obtained inductively in the following manner.

(i) 0, 1, and m, for all m in M, are rational expressions (the *atomic* expressions or formulas).

(ii) If E is a rational expression and k is in $\mathbb{K}$, then $(k\,\mathsf{E})$ and $(\mathsf{E}\,k)$ are rational expressions.

(iii) If E and F are rational expressions, then so are $(\mathsf{E}+\mathsf{F})$, $(\mathsf{E}\cdot\mathsf{F})$, and (E^*).

We write $\mathbb{K}\,\mathsf{RatE}\,M$ for the set of rational $\mathbb{K}$-expressions over M. □

Remarks 2.4 (i) We can restrict the atomic formulas, other than 0 and 1, to be elements g of any given generating set G without reducing the power of the definition. That is what we usually do when M is a free monoid A^*.

(ii) We could have considered the elements of $\mathbb{K}$ to be atoms and not operators, again without changing the power of the definition, and that would simplify somewhat some upcoming equations. The chosen way is however more consistent with the upcoming definition of the *derivation* of $\mathbb{K}$-expressions over A^*. □ *cf. Sec. 4.2, p. 443*

We define the *depth* of E, $\mathsf{d}(\mathsf{E})$, as in Chapter I:

$$\mathsf{d}(0) = \mathsf{d}(1) = \mathsf{d}(m) = 0 \ , \quad \text{for all } m \text{ in } M \ ,$$
$$\mathsf{d}((k\,\mathsf{E})) = \mathsf{d}((\mathsf{E}\,k)) = \mathsf{d}((\mathsf{E}^*)) = 1 + \mathsf{d}(\mathsf{E}) \ ,$$
$$\mathsf{d}((\mathsf{E}+\mathsf{F})) = \mathsf{d}((\mathsf{E}\cdot\mathsf{F})) = 1 + \mathsf{max}\big(\mathsf{d}(\mathsf{E}), \mathsf{d}(\mathsf{F})\big) \ .$$

Also as in Chapter I, the *constant term* of an expression E, $\mathsf{c}(\mathsf{E})$, is defined by induction on the depth of E; it is an element of $\mathbb{K}$, and no longer the function 1 or 0, computed by the following equations:

$$\mathsf{c}(1) = 1_{\mathbb{K}} \ , \qquad \mathsf{c}(0) = \mathsf{c}(m) = 0_{\mathbb{K}} \qquad \text{for all } m \text{ in } M \ ,$$
$$\mathsf{c}((k\,\mathsf{E})) = k\,\mathsf{c}(\mathsf{E}) \ , \quad \mathsf{c}((\mathsf{E}\,k)) = \mathsf{c}(\mathsf{E})\,k \ ,$$
$$\mathsf{c}((\mathsf{E}+\mathsf{F})) = \mathsf{c}(\mathsf{E}) + \mathsf{c}(\mathsf{F}) \ , \quad \mathsf{c}((\mathsf{E}\cdot\mathsf{F})) = \mathsf{c}(\mathsf{E})\ \mathsf{c}(\mathsf{F})$$
$$\text{and} \qquad \mathsf{c}((\mathsf{E}^*)) = \mathsf{c}(\mathsf{E})^* \qquad \text{if the left-hand side is defined in } \mathbb{K}.$$

A rational $\mathbb{K}$-expression is a *formula*, and this formula may represent a defined element of $\mathbb{K}\langle\langle M\rangle\rangle$ or not, the distinction between the two cases being made by the constant term, exactly as for the star of an arbitrary series and using that result. We shall say that an expression in $\mathbb{K}\,\mathsf{RatE}\,M$ is *valid* if its constant term is defined. *cf. Prop. 2.6*

cf. Def. I.4.2, p. 125

The power series *denoted by a valid expression* E, which we write $|\mathsf{E}|$, is defined by induction on the depth of E by the equations

$$|\mathsf{0}| = 0_{\mathbb{K}} \ , \quad |\mathsf{1}| = 1_{A^*} \ , \quad |m| = m \qquad \forall m \in M \ ,$$
$$|(k\,\mathsf{E})| = k\,|\mathsf{E}| \ , \quad |(\mathsf{E}\,k)| = |\mathsf{E}|\,k \ ,$$
$$|(\mathsf{E}+\mathsf{F})| = |\mathsf{E}|+|\mathsf{F}| \ , \quad |(\mathsf{E}\cdot\mathsf{F})| = |\mathsf{E}|\,|\mathsf{F}| \ , \qquad \text{and} \qquad |(\mathsf{E}^*)| = |\mathsf{E}|^* \ .$$

We verify both that these equations are well defined and that they are consistent, in the sense that the *constant term of the expression* E is the *constant term of the series* $|\mathsf{E}|$, *in parallel*, and *in the same induction*, using Proposition 2.6. In other words, to define $|\mathsf{E}|$, we shall also have proved:

cf. also Rem. I.4.4, p. 139

Property 2.5 *For all valid* $\mathbb{K}$*-expressions* E, $\mathsf{c}(\mathsf{E}) = \langle |\mathsf{E}|, 1_M \rangle$. ■

Example 2.2 Take $M = \{a,b\}^*$ and $\mathbb{K} = \mathbb{Q}$. The $\mathbb{Q}$-expression[26] $(a^* + (-1\,b^*))^*$ is valid, as is $\mathsf{E}_1 = (\frac{1}{6}\,a^* + \frac{1}{3}\,b^*)^*$ since $\mathsf{c}\left(\frac{1}{6}\,a^* + \frac{1}{3}\,b^*\right) = \frac{1}{2}$, and hence $\mathsf{c}(\mathsf{E}_1) = 2$ is defined; $(a^* + b^*)^*$ is not valid. □

The set of series denoted by valid $\mathbb{K}$-expressions is rationally closed, and every rationally closed subset of $\mathbb{K}\langle\langle M\rangle\rangle$ that contains every element of M (and thus $\mathbb{K}\langle M\rangle$) contains every series denoted by a valid $\mathbb{K}$-expression, which proves:

cf. also Prop. I.4.1, p. 126

Proposition 2.7 *A power series of* $\mathbb{K}\langle\langle M\rangle\rangle$ *is* $\mathbb{K}$*-rational if and only if it is denoted by a valid rational* $\mathbb{K}$*-expression over* M. ■

From this last result we deduce in particular:

Proposition 2.8 *Let* $\varphi\colon \mathbb{K} \to \mathbb{L}$ *be a continuous morphism (of topological semirings). The image under* $\varphi\colon \mathbb{K}\langle\langle M\rangle\rangle \to \mathbb{L}\langle\langle M\rangle\rangle$ *of a* $\mathbb{K}$*-rational series is an* $\mathbb{L}$*-rational series.* ■

cf. p. 127

We can also re-express what we wrote in Chapter I about rational expressions over a set of variables and their interpretations: an *interpretation* of a rational expression E is a map τ from $\mathrm{Var}\,\mathsf{E}$ to $\mathbb{K}\langle\langle M\rangle\rangle$. The *result* of the interpretation τ of E is defined by induction on the depth of E starting from the values $\mathsf{x}\tau$ for all x in $\mathrm{Var}\,\mathsf{E}$. The interpretation τ is a *valid interpretation of* E if each star which appears in the computation is well defined. We can therefore state in particular:

Proposition 2.9 *Let* $\mathcal{E}$ *be a family of elements of* $\mathbb{K}\langle\langle M\rangle\rangle$. *A series* s *belongs to* $\mathbb{K}\mathrm{Rat}\,\mathcal{E}$ *if and only if there exists a rational expression* F *and a valid interpretation of* F *with value in* $\mathcal{E}$ *whose result is* s. ■

[26] To improve legibility we have omitted some parentheses according to common usage (and the usual conventions for operators).

2.2.3 Star of a matrix

The following result shows that rational operations act 'rationally' on matrices. It is essential for the study of automata since it gives directly a 'weighted' version of the McNaughton–Yamada (MNY) algorithm and the state elimination method.

Proposition 2.10 *The coefficients of the star of a* proper *square matrix m with elements in $\mathbb{K}\langle\langle M\rangle\rangle$ belong to the rational closure of the coefficients of m.*

Proof. By induction on the size n of m. It is obvious if $n = 1$; we assume it holds for all k less than some $n \geqslant 2$. Let m be a proper square matrix of size n and let

$$m = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

be a block decomposition of m, where A and D are *square blocks*. Since m is proper, m^* is defined, by Lemma 1.5 (and Proposition 2.3). Let us make its block decomposition explicit, corresponding to that of m:

$$m^* = \begin{pmatrix} E & F \\ G & H \end{pmatrix} \quad .$$

By (2.2) (and Lemma 1.5), it follows that *cf. also Exer. 0.6.9*

$$m^* = \begin{pmatrix} E & F \\ G & H \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} \quad ,$$

which equation can be decomposed into a system of four others:

$$E = 1 + AE + BG \quad , \tag{2.8}$$

$$F = AF + BH \quad , \tag{2.9}$$

$$G = CE + DG \quad , \tag{2.10}$$

$$H = 1 + CF + DH \quad . \tag{2.11}$$

Lemma 2.4 applied to (2.9) and (2.10) gives

$$F = A^*BH \qquad \text{and} \qquad G = D^*CE \quad . \tag{2.12}$$

Substituting into (2.8) and (2.11), we obtain

$$E = (A + BD^*C)^* \qquad \text{and} \qquad H = (D + CA^*B)^* \quad .$$

which, substituted into (2.12), gives

$$m^* = \begin{pmatrix} (A + BD^*C)^* & A^*B(D + CA^*B)^* \\ D^*C(A + BD^*C)^* & (D + CA^*B)^* \end{pmatrix} \quad . \tag{2.13}$$

By the induction hypothesis, the coefficients of A^* and D^* belong respectively to the rational closure of the coefficients of A and D, hence of m, and hence also those of $(A + BD^*C)^*$ and of $(D + CA^*B)^*$. The same then applies to those of m^*. ∎

Remarks 2.5 (i) It should be noted that in order to be able to apply Lemma 2.4 and hence for the proof to be valid, the matrix m must be proper; it is not enough for m^* to be defined.

cf. Exer. I.2.16, p. 101

(ii) We can easily see that the procedure developed in the proof of Proposition 2.10 gives a matrix of rational expressions which does not coincide with any of the matrices computed by MNY algorithm, whatever order is taken on the set that indexes the rows of the matrix. However, there certainly exists a link between the two: I conjecture that for each method of guiding the induction in Proposition 2.10 (that is, every recursive splitting of m into blocks) and each coefficient of m^* obtained by this computation, there exists an order on the indices of m such that MNY algorithm applied according to this order gives the same rational expression for this coefficient.[27] □

Exercises

• 2.9 Compute the star of the matrix $m = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & j \end{pmatrix}$.

• 2.10 Compute the star of the matrix $m = \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}$.

2.11 Suppose $\mathbb{K}$ is a ring. Show that $\mathbb{K}\mathrm{Rat}\, M$ is the smallest sub-$\mathbb{K}$-algebra of $\mathbb{K}\langle\langle M\rangle\rangle$ which contains $\mathbb{K}\langle M\rangle$ that is closed under inversion (that is, which contains the inverse of every invertible series).

2.3 Weighted automaton in a semiring

Bis repetita placent.

Cicero

In accordance with the line we have taken up to now, we say that an *automaton over M with weight*[28] *in* $\mathbb{K}$, or $\mathbb{K}$-*automaton*[29] *over* M is a *graph* labelled with elements of $\mathbb{K}\langle\langle M\rangle\rangle$, associated with two maps from the set of vertices to $\mathbb{K}\langle\langle M\rangle\rangle$.

cf. Def. I.1.1, p. 51 and p. 219

We shall develop and complete this definition and reuse, with slight modifications, the terminology and definitions of Sections I.1 and II.1 based on the identification of a graph with its incidence matrix. The proofs we shall make will be performed systematically with matrix computations. However, we shall continue to use the graph representation and its vocabulary to aid intuition.

2.3.1 $\mathbb{K}$-automaton over M

Having chosen the semiring $\mathbb{K}\langle\langle M\rangle\rangle$, we can specify an automaton $\mathcal{A}$ over M with weights in $\mathbb{K}$ by the choice of the following:

[27] *Note added in translation:* Or at least equivalent modulo the system **U**.

[28] Or *with multiplicity* in $\mathbb{K}$.

[29] Or *weighted automaton* if $\mathbb{K}$ is understood or immaterial.

- A non-empty set Q of *states* of $\mathcal{A}$, also called the *dimension* of $\mathcal{A}$.
- An element E of $\mathbb{K}\langle\langle M\rangle\rangle^{Q\times Q}$, a square matrix of dimension Q with entries in $\mathbb{K}\langle\langle M\rangle\rangle$, called the *transition matrix* of $\mathcal{A}$; we can view each entry $E_{p,q}$ as the label of a unique edge which goes from state p to state q in the graph with vertices Q and we write $p \xrightarrow{x} q$, or $p \xrightarrow[\mathcal{A}]{x} q$, if $x = E_{p,q}$. (If $E_{p,q} = 0_{\mathbb{K}}$, we consider there to be *no* edge from p to q.)
- Two elements I and T of $\mathbb{K}\langle\langle M\rangle\rangle^{Q}$; that is, two functions I and T from Q to $\mathbb{K}\langle\langle M\rangle\rangle$: I is the *initial function* and T the *final function* of $\mathcal{A}$; they can also be seen as vectors of dimension Q: I is a *row vector* and T a *column vector*, called respectively the *initial vector* and *final vector* of $\mathcal{A}$.

The $\mathbb{K}$-automaton $\mathcal{A}$ is written, naturally enough, *cf. Def. I.1.3, p. 54*

$$\mathcal{A} = \langle\, I, E, T \,\rangle \ .$$

We shall use the familiar conventions to represent $\mathbb{K}$-automata graphically; the values of I labelling the incoming arrows and those of T the outgoing arrows.

Example 2.3 An $\mathbb{N}$-automaton over $\{a,b\}^*$ defined by

$$\mathcal{E}_1 = \left\langle \begin{pmatrix} a & 0 \end{pmatrix}, \begin{pmatrix} a+b & 2\,a \\ 0 & 3\,a+b \end{pmatrix}, \begin{pmatrix} 0 \\ 2\,b \end{pmatrix} \right\rangle$$

is shown in Figure 2.1 opposite.

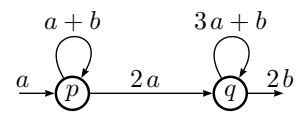


Fig. 2.1: The $\mathbb{N}$-automaton $\mathcal{E}_1$

A *path* in $\mathcal{A}$ is a sequence of transitions such that the source of each is the destination of the previous one; it can be written

$$c \ := \ p_0 \xrightarrow{x_1} p_1 \xrightarrow{x_2} p_2 \cdots \xrightarrow{x_n} p_n \ .$$

The *label*, or *result* of c, written $|c|$, is the *product* of the labels of the transitions of c. In the above case, $|c| = x_1 x_2 \cdots x_n$.

A *computation* in $\mathcal{A}$ is a path to which is added an arrow arriving at the source and one leaving from the destination, with their respective labels. The computation corresponding to the above path is hence

$$d \ := \ \xrightarrow{I_{p_0}} p_0 \xrightarrow{x_1} p_1 \xrightarrow{x_2} p_2 \cdots \xrightarrow{x_n} p_n \xrightarrow{T_{p_n}} \ .$$

The *label*, or *result*, of d, still written $|d|$, is the product of the label of the incoming arrow, that of the path, and that of the outgoing arrow, in that order; in our case, $|d| = I_{p_0} x_1 x_2 \cdots x_n T_{p_n}$.

Before continuing, let us assure ourselves that the definitions we have made for weighted automata are indeed a generalisation of the previous definitions.

(i) An automaton over A (Chapter I) is a $\mathbb{B}$-automaton over A^*; an automaton over M (Chapter II) is a $\mathbb{B}$-automaton over M.

(ii) To the subset I, or T, of Q we relate its characteristic function, also written I or T; this is a function from Q to $\mathbb{B}$, hence a function from Q to $\mathbb{B}\langle\langle M\rangle\rangle$ since $\mathbb{B}$ is identified with the sub-semiring of series whose support is 1_M.
(iii) The distinction between *path* and *computation*, which are often used as synonyms, may seem useless and smack of sophistry. But apart from the fact that it is consistent with our terminology – 'path' refers to 'graph' while 'computation' refers to 'automaton', and what distinguishes an automaton from a graph is precisely that initial and final states are taken into account – it was only introduced in order to make precise definitions that incorporate the generality that we have now allowed for I and T. In the majority of cases, the non-zero elements of I and T will be scalar (that is, elements of $\mathbb{K}$), usually equal to $1_{\mathbb{K}}$ – *cf.* (ii) above in particular – and the two notions will coincide.
(iv) Along the same lines, the notion of a *successful computation* has not really disappeared. A state p such that the component I_p is *non-zero* (that is, different from $0_{\mathbb{K}\langle\langle M\rangle\rangle}$) can be called *initial*, and a state where T_p is non-zero can be called *final*. We can then say that a computation is successful if its source is an initial state and its destination is a final state.

2.3.2 Behaviour of a $\mathbb{K}$-automaton

Definition 2.4 The *behaviour* of an automaton $\mathcal{A} = \langle I, E, T\rangle$ of finite dimension Q is defined if and only if for all p and q in Q the family of labels of paths with source p and destination q is summable. In this case, the family of labels of computations of $\mathcal{A}$ is summable and its sum is the behaviour of $\mathcal{A}$, written $|\mathcal{A}|$. We also say that $\mathcal{A}$ *accepts*, or *realises*,[30] the series $|\mathcal{A}|$. □

The description of the transitions of an automaton by a matrix is justified by the fact that a walk over a graph corresponds to a matrix multiplication. This is expressed by the following proposition,[31] which is a generalisation of Lemma 0.7.1.

cf. p. 34

Lemma 2.11 *Let $\mathcal{A} = \langle I, E, T\rangle$ be a $\mathbb{K}$-automaton over M of finite dimension. For every integer n, E^n is the matrix of the sums of the labels of paths of length n.*

Proof. By induction on n. The assertion is true for $n = 1$ (and also for $n = 0$ by convention). The definition of the $(n+1)$th power of E is

$$\forall p, q \in Q \qquad (E^{n+1})_{p,q} = \sum_{r\in Q} (E^n)_{p,r}\, E_{r,q} \ .$$

[30] But not *recognises*, as in Chapter I.
[31] For which the distinction between 'computation' and 'path' is necessary.

Every path of length $n+1$ is the concatenation of a path of length n with a path of length 1: that is, a single transition. We can therefore write[32]

$$\{c \mid c := p \underset{\mathcal{A}}{\longrightarrow} q \ , \ l(c) = n+1\} = \bigcup_{r \in Q} \{(d,e) \mid d := p \underset{\mathcal{A}}{\longrightarrow} r \ , \ l(d) = n \ , \ e := r \underset{\mathcal{A}}{\longrightarrow} q \in E\} \ ,$$

and hence

$$\begin{aligned} \sum \{|c| \mid c := p \underset{\mathcal{A}}{\longrightarrow} q \ , \ l(c) = n+1\} &= \sum_{r \in Q} \left(\{|d||e| \mid d := p \underset{\mathcal{A}}{\longrightarrow} r \ , \ l(d) = n \ , \ e := r \underset{\mathcal{A}}{\longrightarrow} q \in E\}\right) \\ &= \sum_{r \in Q} \left[\left(\sum \{|d| \mid d := p \underset{\mathcal{A}}{\longrightarrow} r \ , \ l(d) = n\} \right) E_{r,q} \right] \ . \end{aligned}$$

As $\sum\{|d| \mid d = p \underset{\mathcal{A}}{\longrightarrow} r \ , \ l(d) = n\} = (E^n)_{p,r}$ by the induction hypothesis, the lemma is proved. ■

Since the sum of the results of the *computations* of length n is equal by definition to the product $I \cdot E^n \cdot T$, and since the behaviour of $\mathcal{A}$ is equal to the sum of the results of the computations of all the lengths, it follows that: *cf. Lem. 1.6, p. 389*

Corollary 2.12 *Let $\mathcal{A} = \langle I, E, T \rangle$ be a $\mathbb{K}$-automaton of finite dimension whose behaviour is defined, then E^* is defined and we have*

$$|\mathcal{A}| = I \cdot E^* \cdot T \ . \qquad ■$$

2.3.3 Notes

The behaviour of an automaton was defined by returning to the essence of an 'automaton': a procedure for describing computations. With this definition, the behaviour of the two automata in Figure 2.2 (a) and (b) are not defined although in the first case the family $\{I \cdot E^n \cdot T\}_{n \in \mathbb{N}}$ is summable since all its terms are zero, and in the second E^* is defined since $E^2 = 0$. On the other hand, these examples also lead us to note that the transition between each pair of states p and q must be *unique*, and labelled $E_{p,q}$; otherwise we would be able to 'decompose' these entries in such a way that the family of labels of paths would no longer be summable.

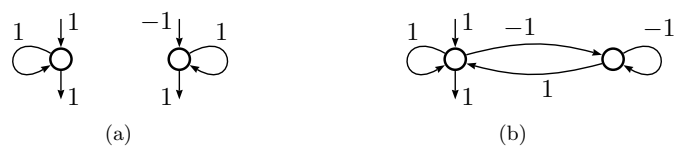


Figure 2.2: Two $\mathbb{Z}$-automata with behavioural problems

From Lemma 1.5 and Proposition 2.6 we deduce a sort of generalisation of the same Proposition 2.6, whose proof is left as an exercise.

[32] The length of a path c is here written $l(c)$.

Proposition 2.13 *Let $\mathbb{K}$ be a strong topological semiring and M a graded monoid. The behaviour of a $\mathbb{K}$-automaton over M, $\mathcal{A} = \langle I, E, T \rangle$ is defined if and only if the behaviour of the $\mathbb{K}$-automaton $\mathcal{A}_0 = \langle I, E_0, T \rangle$ is defined, where E_0 is the matrix of constant terms of entries of E, and in that case we have*

$$|\mathcal{A}| = I \cdot {E_0}^* \cdot (E_{\mathsf{p}} \cdot {E_0}^*)^* \cdot T \quad . \qquad \blacksquare$$

The example of Figure 2.2 (b) shows that it is not sufficient that ${E_0}^*$ be defined, nor even that E_0 be nilpotent[33] for the behaviour of $\mathcal{A}$ be defined. On the other hand, the behaviour of $\mathcal{A}$ is defined when E_0 is *strict upper triangular* since in this case the number of computations in $\mathcal{A}_0$ is *finite*.

Remark 2.6 To avoid difficulties later (for example, with the proof of Proposition 2.14), the definition of the behaviour of $\langle I, E, T \rangle$ must imply at least that the star of E is defined. I have adopted an even stricter definition of behaviour which seemed to me consistent with the previous definitions; but I am not sure that it is necessary, and it would perhaps be sufficient to consider the behaviour of $\langle I, E, T \rangle$ to be defined when the star of E is. $\square$

2.3.4 Some other definitions and examples

Definition 2.5 A $\mathbb{K}$-automaton over M, $\mathcal{A} = \langle I, E, T \rangle$, is *finite* if:

(i) the dimension of $\mathcal{A}$ is finite;

(ii) the coefficients of E, I and T are polynomials; that is, have finite support. $\square$

Example 2.4 The $\mathbb{N}$-automaton over $\{a, b\}^*$ defined by

$$\mathcal{B}_1 = \left\langle \begin{pmatrix} 1 & 0 \end{pmatrix}, \begin{pmatrix} a+b & b \\ 0 & a+b \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\rangle$$

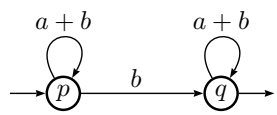


Fig. 2.3: The $\mathbb{N}$-automaton $\mathcal{B}_1$

cf. Exam. 1.1, p. 380

is shown in the adjacent figure. A simple calculation allows us to determine its behaviour:

$$\forall f \in A^* \qquad \langle |\mathcal{B}_1|, f \rangle = |f|_b \, ; \qquad \text{that is} \qquad |\mathcal{B}_1| = \sum_{f \in A^*} |f|_b \, f = t_1 \quad . \tag{2.14}$$

Example 2.5 Another $\mathbb{N}$-automaton over $\{a, b\}^*$ defined by

$$\mathcal{C}_1 = \left\langle \begin{pmatrix} 1 & 0 \end{pmatrix}, \begin{pmatrix} a+b & b \\ 0 & 2a+2b \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\rangle$$

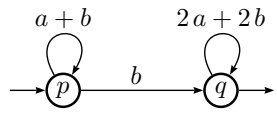


Fig. 2.4: The $\mathbb{N}$-automaton $\mathcal{C}_1$

is shown in the adjacent figure. We use the con-

cf. p. 52

vention of Example I.1.2 for the 'divider by 3': each word f of A^* is considered as a number written in binary, interpreting a as the digit 0 and b as the digit 1, and we

[33] That is, that there exists an n such that ${E_0}^n = 0$.

write $\overline{f}$ for the integer represented[34] by the word f. We can verify that $\overline{f}$ is correctly computed by $\mathcal{C}_1$ in the sense that

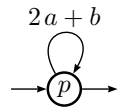
$$\forall f \in A^* \qquad \langle |\mathcal{C}_1|, f\rangle = \overline{f}\,; \qquad \text{that is} \qquad c_1 = |\mathcal{C}_1| = \sum_{f\in A^*} \overline{f}\, f \ . \tag{2.15}$$

The first few terms of c_1 are $\quad c_1 = b + a\,b + 2\,b\,a + 3\,b\,b + a\,a\,b + \cdots$ ■

Example 2.6 [To illustrate the case where M is not a free monoid, an $\mathbb{N}$-automaton over $\mathbb{N}^2$.] It improves legibility to write the operation of $M = \mathbb{N}^2$ multiplicatively and to represent the element (m, n) of $\mathbb{N}^2$ by $a^n b^m$: that is, to see $\mathbb{N}^2$ as the quotient of $\{a, b\}^*$ by the relation $a\,b = b\,a$.

The automaton $\mathcal{D}_1$, defined by

$$\mathcal{D}_1 = \left\langle \left(1\right), \left(2\,a + b\right), \left(1\right) \right\rangle$$



is shown in the adjacent figure.

It can be verified that the behaviour of $\mathcal{D}_1$ is: $\quad |\mathcal{D}_1| = \sum_{a^n b^m \in \mathbb{N}^2} 2^n\, \mathrm{C}^n_{n+m}\, a^n b^m$. □

Example 2.7 [To illustrate the case where $\mathbb{K}$ is *different* from $\mathbb{N}$: an $\mathcal{M}$-automaton[35] over $\{a, b\}^*$.] An $\mathcal{M}$-automaton $\mathcal{S}_1$ is shown in the figure below. The use of such automata is rather disconcerting at first. One must remember that $0 = 1_{\mathcal{M}}$ is the neutral element of *multiplication* in $\mathcal{M}$; we write it explicitly in Figure 2.5, although we do not mention the coefficients $1_{\mathbb{K}}$ explicitly in the preceding figures. Let us calculate for example the coefficient of $a\,b\,a$ in $|\mathcal{S}_1|$; this word is the label of two computations in $\mathcal{S}_1$:

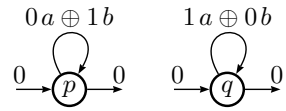


Fig. 2.5: The $\mathcal{M}$-automaton $\mathcal{S}_1$

$$\xrightarrow{0} p \xrightarrow{0\,a} p \xrightarrow{1\,b} p \xrightarrow{0\,a} p \xrightarrow{0} \qquad \text{and} \qquad \xrightarrow{0} q \xrightarrow{1\,a} q \xrightarrow{0\,b} q \xrightarrow{1\,a} q \xrightarrow{0} \ .$$

The coefficient of the first computation is the product (in $\mathcal{M}$, hence the sum) of the transition coefficients – that is 1 – while that of the second is 2; the coefficient of $a\,b\,a$ is the sum (in $\mathcal{M}$, hence the minimum) of these two coefficients: that is 1. Clearly the support of $|\mathcal{S}_1|$ is all of $\{a, b\}^*$ and the coefficient in $|\mathcal{S}_1|$ of an arbitrary word f of $\{a, b\}^*$ is $\mathsf{min}\{|f|_a, |f|_b\}$: that is, $|\mathcal{S}_1| = s'_1 + t'_1$. □ *cf. Exam. 1.1, p. 380*

[34]The reader may be irritated by this excess of caution and think it reasonable to view the characters 0 and 1 both as letters of an alphabet and as digits or numbers, as we did in the prologue. The justification of the encoding appears in this example: without it, the transition matrix of $\mathcal{C}_1$ would be written $\begin{pmatrix} 0+1 & 1 \\ 0 & 2\cdot 0 + 2\cdot 1 \end{pmatrix}$ which is rather illegible, despite the dot added to separate the digits from the coefficients; the digit 0 of the transition (p, p) must not be confused with the coefficient 0 of the transition (q, p).

[35]The 'tropical' semiring $\mathcal{M} = \langle \mathbb{N} \cup \{+\infty\}, \mathsf{min}, + \rangle$ (*cf.* Exam. 0.5.1, p. 28) appears naturally as the semiring of coefficients in several automata-theoretic problems.

Trim automaton

As for automata over A^*, or over M, we shall say that a state p of a $\mathbb{K}$-automaton is *accessible* (resp. *co-accessible*) if there exists a computation from an *initial state* – with the definition given above – to p (resp. a computation from p to a *final state*) whose label is non-zero. A state is *useful* if it is both accessible and co-accessible. An automaton is accessible (resp. co-accessible) if all its states are, and is *trim* if all its states are useful and co-accessible.

Note that if $\mathbb{K}$ has no divisor of zero, the determination of accessible (or co-accessible) states is identical in a $\mathbb{K}$-automaton and in a $\mathbb{B}$-automaton; in the contrary case, this same computation can be of arbitrary complexity, depending on $\mathbb{K}$.

Support of a $\mathbb{K}$-automaton – characteristic automaton

The *support automaton* of $\mathcal{A} = \langle I, E, T \rangle$, a $\mathbb{K}$-automaton over M, is the automaton over M, $\mathsf{supp}\,\mathcal{A}$, obtained by taking the support of each coefficient of I, E and T. Conversely, to each automaton over M, $\mathcal{A} = \langle Q, M, E, I, T \rangle$, we can associate ($\mathbb{K}$ being fixed, and in general equal to $\mathbb{N}$) its characteristic automaton defined as the $\mathbb{K}$-automaton whose support is $\mathcal{A}$ and all of whose non-zero coefficients are $1_{\mathbb{K}}$. The following properties should be obvious.

Property 2.6 *The support of the behaviour of $\mathcal{A}$, a $\mathbb{K}$-automaton over M, is contained in the behaviour of the support of $\mathcal{A}$. If $\mathbb{K}$ is* positive*, these two subsets of M are equal; that is,* $\mathsf{supp}\,|\mathcal{A}| = |\mathsf{supp}\,\mathcal{A}|$. ■

Property 2.7 *An automaton over A^* is unambiguous if and only if the behaviour of its characteristic $\mathbb{N}$-automaton is a characteristic series.* ■

Remark 2.7 We can also represent $\mathbb{K}$-automata, and particularly $\mathbb{K}$-automata over a free monoid, in such a way that if we take the support of the label of each transition we obtain the usual representation of the support automaton. For example, the $\mathbb{N}$-automata $\mathcal{B}_1$ and $\mathcal{C}_1$ above will be represented instead as in Figure 2.6. □

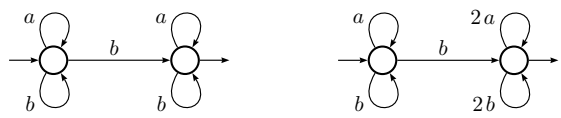


Figure 2.6: The $\mathbb{N}$-automata $\mathcal{B}_1$ and $\mathcal{C}_1$

Exercises

• 2.12 (a) Let E be a matrix with coefficients in $\mathcal{M}$. What do the coefficients of the matrix E^* represent?

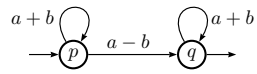
(b) Answer the same question for a matrix with coefficients in $\mathbb{P} = \langle \mathbb{N} \cup \{+\infty\}, \mathsf{max}, \mathsf{min} \rangle$.

2.13 Let E be an upper triangular matrix with elements in $\mathbb{K}$ such that the star of each element on the diagonal is defined. Show that the behaviour of an automaton $\mathcal{A} = \langle I, E, T \rangle$ is defined.

cf. also Exer. 2.10, p. 402

2.14 Verify the formulas given for the behaviours of the automata $\mathcal{B}_1$ and $\mathcal{D}_1$ (Examples 2.4 and 2.6). Compute the behaviour of the automaton $\mathcal{E}_1$ (Example 2.3).

• 2.15 Compute the behaviour of the automaton $\mathcal{D}_2$ in the adjacent figure, when $\mathcal{D}_2$ is
(a) considered as a $\mathbb{Z}$-automaton over $\{a, b\}^*$;
(b) considered as a $\mathbb{Z}$-automaton over $\mathbb{N}^2$.



• 2.16 Give an example of a $\mathbb{Z}$-automaton over $\{a\}^*$, $\mathcal{A} = \langle I, E, T \rangle$, whose behaviour is undefined and such that the family $\{I \cdot E^n \cdot T\}_{n \in \mathbb{N}}$ is nonetheless summable.

• 2.17 (a) Prove (2.15) (Example 2.5); that is, show that

$$\forall f \in A^* \qquad \langle |\mathcal{C}_1|, f \rangle = \overline{f} \ . \tag{2.15}$$

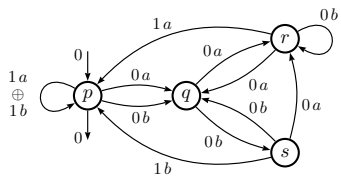
(b) Show that, for all n, there exists an $\mathbb{N}$-automaton $\mathcal{V}'_n$ with $n+1$ states which, for every f in A^*, 'computes' the nth power of the number represented by f:

$$\forall f \in A^* \qquad \langle |\mathcal{V}'_n|, f \rangle = \overline{f}^{\,n} \ .$$

• 2.18 Show that every automaton over A with spontaneous transitions is equivalent to an automaton without spontaneous transitions.

2.19 Prove Properties 2.6 and 2.7.

• 2.20 Show that the behaviour of the $\mathcal{M}$-automaton shown in the adjacent figure is a series whose support is all of $\{a, b\}^*$ and whose coefficients are less than or equal to 4.



2.4 The Fundamental Theorem of finite automata

One hesitates to say of a proposition, 'here is *the Fundamental Theorem*'. However, this seems justified for the one that follows: it states completely generally, at least under the current assumption that M is a graded monoid of finite type and $\mathbb{K}$ a strong semiring, that what one can 'do' with a finite automaton is precisely what one can 'do' with rational operations.

Theorem 2.1 *A power series of $\mathbb{K}\langle\langle M \rangle\rangle$ is rational if and only if it is the behaviour of some finite $\mathbb{K}$-automaton over M.*

We can make Theorem 2.1 both more precise and more general, closer to the properties used in the proof. For this we need to define a restricted class of $\mathbb{K}$-automata.

2.4.1 Proper automata – proper families of series

Definition 2.6 A $\mathbb{K}$-automaton over M, $\mathcal{A} = \langle I, E, T \rangle$, is *proper* if:

(i) the matrix E is *proper*;

(ii) the entries of I and T are *scalar*; that is, $I \in \mathbb{K}^{1\times Q}$ and $T \in \mathbb{K}^{Q\times 1}$. □

The automata in Examples 2.4 to 2.7 are proper. It follows from Proposition 2.13 that the behaviour of a proper automaton is well defined; the following result adds the converse.

Proposition 2.14 *Every $\mathbb{K}$-automaton $\mathcal{A}$ over M whose behaviour is defined is equivalent to a proper automaton whose entries, other than the scalar entries of the initial and final vectors, are linear combinations of proper parts of entries of $\mathcal{A}$.*

Proof. We first show that $\mathcal{A} = \langle I, E, T \rangle$ is equivalent to an automaton $\mathcal{B} = \langle J, F, U \rangle$ where the entries of J and U are scalar. We set

$$J = \begin{pmatrix} 1 & \boxed{\quad 0 \quad} & 0 \end{pmatrix} \quad , \quad F = \begin{pmatrix} 0 & \boxed{\quad I \quad} & 0 \\ \boxed{0} & \boxed{\quad E \quad} & \boxed{T} \\ 0 & \boxed{\quad 0 \quad} & 0 \end{pmatrix} \quad , \quad U = \begin{pmatrix} 0 \\ \boxed{0} \\ 1 \end{pmatrix} \quad . \tag{2.16}$$

Every path in $\mathcal{B}$ is a path or a computation in $\mathcal{A}$ and the behaviour of $\mathcal{B}$ is defined if and only if that of $\mathcal{A}$ is, and in that case E^* is defined. We verify by induction that, for every integer n greater than or equal to 2,

$$F^n = \begin{pmatrix} 0 & I \cdot E^{n-1} & I \cdot E^{n-2} \cdot T \\ 0 & E^n & E^{n-1} \cdot T \\ 0 & 0 & 0 \end{pmatrix} \quad . \tag{2.17}$$

We have $J \cdot U = J \cdot F \cdot U = 0$, $J \cdot F^{n+2} \cdot U = I \cdot E^n \cdot T$, hence $J \cdot F^* \cdot U = I \cdot E^* \cdot T$ and $\langle J, F, U \rangle$ is equivalent to $\mathcal{A}$.

Next, starting from an automaton $\mathcal{B} = \langle J, F, U \rangle$ whose initial and final vectors are scalar, we set

$$F = Z + G \quad ,$$

cf. Prop. 2.13, p. 406 where Z is the matrix of the constant terms of the entries of F (that is, $Z \in \mathbb{K}^{Q\times Q}$) and where G is that of the proper parts of the entries of F. The behaviour of $\mathcal{B}$ is defined if and only if the behaviour of the automaton $\langle J, Z, U \rangle$ is defined, and in *cf. Prop. 2.6, p. 396* this case Z^* is defined too. We then have $F^* = (Z^* \cdot G)^* \cdot Z^*$, whence

$$|\mathcal{B}| = I \cdot E^* \cdot T = J \cdot F^* \cdot U = J \cdot H^* \cdot V \quad ,$$

with $H = Z^* \cdot G$ and $V = Z^* \cdot U$. Since Z^* is an element of $\mathbb{K}^{Q\times Q}$, the entries of H are linear combinations (with coefficients in $\mathbb{K}$) of entries of G and the entries of V are scalar. ■

Remark 2.8 Recall that the behaviour of an automaton $\mathcal{A} = \langle I, E, T \rangle$ is defined in particular if the matrix of constant terms of the entries of E is a strict upper triangular matrix. □

Remark 2.9 If $\mathbb{K} = \mathbb{B}$, the matrix Z (in the second part of the proof) represents *the spontaneous transitions* of $\mathcal{A}$, and the automaton $\langle I, H, V \rangle$ is the automaton obtained by *backward closure*. If we set $I' = I \cdot Z^*$ and $H' = F \cdot Z^*$ the automaton $\langle I', H', U \rangle$ is the automaton obtained by *forward closure*. □

cf. Def. I.1.14, p. 80 and Exer. 2.18

Definition 2.7 We will say that a family of series is *proper* if it contains the proper part of each of its elements.[36] □

In particular, the polynomials are a proper family of $\mathbb{K}\langle\langle M \rangle\rangle$. We can thus state the promised generalisation of Theorem 2.1.

2.4.2 Statement and proof

Theorem 2.2 *Let $\mathcal{C}$ be a proper family of series of $\mathbb{K}\langle\langle M \rangle\rangle$. A series s of $\mathbb{K}\langle\langle M \rangle\rangle$ belongs to $\mathbb{K}\mathrm{Rat}\,\mathcal{C}$ if and only if s is the behaviour of a* proper *$\mathbb{K}$-automaton over M of finite dimension whose (non-scalar) entries are finite linear combinations of elements of $\mathcal{C}$.*

Proof. Let $\mathcal{A} = \langle I, E, T \rangle$ be a proper $\mathbb{K}$-automaton of finite size over M. Its behaviour is defined and

$$|\mathcal{A}| = I \cdot E^* \cdot T \quad .$$

The sufficient condition is a restatement of Proposition 2.10. It is therefore like a matrix and weighted version of the property proved in Chapter I by algorithm MNY and the state elimination method:[37] *a series accepted by a finite $\mathbb{K}$-automaton is $\mathbb{K}$-rational.*

cf. Prop. I.2.7, p. 94

For the necessary condition, we shall show that the family of behaviours of automata in question:

(a) contains the elements of $\mathcal{C}$ as well as $0_{\mathbb{K}}$ and $1_{\mathbb{K}}$;

(b) and is rationally closed.

We write $\mathcal{D}$ for the family of behaviours of proper $\mathbb{K}$-automata whose entries are linear combinations of elements of $\mathcal{C}$. We first show (a).

• $0_{\mathbb{K}}, 1_{\mathbb{K}} \in \mathcal{D}$. It follows from $0_{\mathbb{K}} = 0_{\mathbb{K}} \cdot (0_{\mathbb{K}})^* \cdot 0_{\mathbb{K}}$ and $1_{\mathbb{K}} = 1_{\mathbb{K}} \cdot (0_{\mathbb{K}})^* \cdot 1_{\mathbb{K}}$.

• $\mathcal{C} \subseteq \mathcal{D}$. We write s_{p} for the proper part of s and $s_0 = \langle s, 1_M \rangle$ for its constant term. Then

$$s = \begin{pmatrix} 1_{\mathbb{K}} & 0_{\mathbb{K}} \end{pmatrix} \cdot \begin{pmatrix} 0_{\mathbb{K}} & s_{\mathsf{p}} \\ 0_{\mathbb{K}} & 0_{\mathbb{K}} \end{pmatrix}^* \cdot \begin{pmatrix} s_0 \\ 1_{\mathbb{K}} \end{pmatrix} \quad . \tag{2.18}$$

[36] As opposed to all the series in the family being proper.

[37] Note that the above-mentioned algorithms are valid for K-automata and that, if they furnish a less elegant proof than that given by Proposition 2.10, the effective computations to which they lead are much more efficient (*cf.* Exercise 2.9).

We now show (b). Let $\langle I, E, T \rangle$ and $\langle J, F, S \rangle$ be two automata which satisfy the assumptions of construction of $\mathcal{D}$. We shall show that $\mathcal{D}$ is closed under $\mathbb{K}$-rational operations.

• Exterior multiplications. We have

$$k\,|\langle I, E, T \rangle| = |\langle k\,I, E, T \rangle| \qquad \text{and} \qquad |\langle I, E, T \rangle|\,k = |\langle I, E, T\,k \rangle| \ . \tag{2.19}$$

• Sum. We have $|\langle L, G, U \rangle| = |\langle I, E, T \rangle| + |\langle J, F, S \rangle|$, with

$$L = \begin{pmatrix} I & J \end{pmatrix} \ , \quad G = \begin{pmatrix} E & 0 \\ 0 & F \end{pmatrix} \ , \quad U = \begin{pmatrix} T \\ S \end{pmatrix} \ . \tag{2.20}$$

In fact

$$G^n = \begin{pmatrix} E^n & 0 \\ 0 & F^n \end{pmatrix} \ ,$$

hence $L \cdot G^* \cdot U = I \cdot E^* \cdot T + J \cdot F^* \cdot S$. If E and F satisfy the assumptions, then so does G.

• Product. We first set

$$L = \begin{pmatrix} I & 0 \end{pmatrix} \ , \quad G = \begin{pmatrix} E & T \cdot J \\ 0 & F \end{pmatrix} \ , \qquad \text{and} \qquad U = \begin{pmatrix} 0 \\ S \end{pmatrix} \ ; \tag{2.21}$$

$T \cdot J$ is a scalar $r \times s$ matrix if r is the dimension of E and s that of F. We then write $G = Z + H$, with

$$H = \begin{pmatrix} E & 0 \\ 0 & F \end{pmatrix} \qquad \text{and} \qquad Z = \begin{pmatrix} 0 & T \cdot J \\ 0 & 0 \end{pmatrix} \ .$$

The matrix Z is strict upper triangular (indeed, $Z^2 = 0$) and H is proper if E and F are. The behaviour of $\langle L, G, U \rangle$ is well defined.

It follows that for every positive integer n

$$G^n = \begin{pmatrix} E^n & R_n \\ 0 & F^n \end{pmatrix} \ , \qquad \text{with} \qquad R_n = \sum_{\substack{0 \leqslant p, q \leqslant n-1 \\ p+q=n-1}} E^p \cdot T \cdot J \cdot F^q \ , \tag{2.22}$$

and we deduce directly by (2.22) (implicitly using the assumption that $\mathbb{K}\langle\langle M \rangle\rangle$ is strong) that $|\langle L, G, U \rangle| = |\langle I, E, T \rangle|\,|\langle J, F, S \rangle|$. The automaton $\langle L, G, U \rangle$ is equivalent to $\langle L, G', U' \rangle$, where

$$G' = Z^* \cdot H = \begin{pmatrix} E & T \cdot J \cdot F \\ 0 & F \end{pmatrix} \ , \qquad \text{and} \qquad U' = Z^* \cdot U = \begin{pmatrix} T \cdot J \cdot S \\ S \end{pmatrix} \ . \tag{2.23}$$

The entries of G' are linear combinations of entries of H, hence of E and F, and $\mathcal{D}$ is closed under product.

• Star. We suppose that the star of the behaviour of the automaton $\langle I, E, T \rangle$ (of size r) is defined, hence that the star of its constant term $I \cdot T$ is defined (since E

is proper and M is graded). We construct an automaton of size $r+1$ in the following manner:

$$L = \begin{pmatrix} 1 & 0 \end{pmatrix} \quad , \quad G = \begin{pmatrix} I \cdot T & I \\ E \cdot T & E \end{pmatrix} \quad , \quad U = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad . \tag{2.24}$$

We set $G = Z + H$, with

$$H = \begin{pmatrix} 0 & 0 \\ E \cdot T & E \end{pmatrix} \qquad \text{and} \qquad Z = \begin{pmatrix} I \cdot T & I \\ 0 & 0 \end{pmatrix} \quad .$$

By assumption, H is proper (its entries are linear combinations of entries of E). As the star of $I \cdot T$ is defined, the behaviour of $\langle L, Z, U \rangle$ is defined, hence so is that of $\langle L, G, U \rangle$ (Proposition 2.13). *cf. Exer. 2.13*

The star of Z is written[38]

$$Z^* = \begin{pmatrix} (I \cdot T)^* & (I \cdot T)^* \cdot I \\ 0 & 1 \end{pmatrix} \quad .$$

If we set, for all integers l and k,

$$P_l^k = \sum_{\substack{0 \leqslant j_1, j_2, \ldots, j_k \\ j_1 + j_2 + \cdots + j_k = l}} I \cdot E^{j_1} \cdot T \cdot I \cdot E^{j_2} \cdot T \cdots I \cdot E^{j_k} \cdot T \quad , \tag{2.25}$$

we can verify that:

(i) for each k, the family $\{P_l^k \mid l \in \mathbb{N}\}$ is locally finite;

(ii) for each k and n, we have the identities:

$$\sum_{l=0}^{\infty} P_l^k = (I \cdot E^* \cdot T)^k \quad , \tag{2.26}$$

$$\sum_{l=0}^{n-1} P_l^{n-l} = L \cdot G^n \cdot U \quad , \tag{2.27}$$

from which we deduce that $L \cdot G^* \cdot U = (I \cdot E^* \cdot T)^*$ (by a similar computation to that in the proof of Proposition 2.6). The automaton $\langle L, G, U \rangle$ is equivalent to $\langle L, G', U' \rangle$, with *cf. p. 396*

$$G' = Z^* \cdot H = \begin{pmatrix} (I \cdot T)^* \cdot I \cdot E \cdot T & (I \cdot T)^* \cdot I \cdot E \\ E \cdot T & E \end{pmatrix} \quad ,$$

$$\text{and} \quad U' = Z^* \cdot U = \begin{pmatrix} (I \cdot T)^* \\ 0 \end{pmatrix} \quad . \tag{2.28}$$

The entries of G' are linear combinations of entries of H, hence of E, and $\mathcal{D}$ is closed under star. ■

[38] The coefficient 1 in Z^* implicitly represents the identity matrix of dimension equal to that of E.

2.4.3 Notes and corollaries

Not only does Theorem 2.2 imply Theorem 2.1, but it allows us to give a more precise version; since an arbitrary set of generators of M generates the semialgebra $\mathbb{K}\langle M\rangle$, we have:

Corollary 2.15 *A series of $\mathbb{K}\langle\langle M\rangle\rangle$ is $\mathbb{K}$-rational if and only if it is the behaviour of a proper $\mathbb{K}$-automaton over M of finite dimension whose entries are linear combinations (with coefficients in $\mathbb{K}$) of elements of a set of generators of M.* ■

Proposition 2.14 further gives us a slightly more general statement of the sufficient condition expressed in Theorem 2.2:

Corollary 2.16 *Let $\mathcal{A}$ be a $\mathbb{K}$-automaton whose (non-scalar) entries belong to a proper family of series $\mathcal{C}$. If the behaviour of $\mathcal{A}$ is defined, it belongs to $\mathbb{K}\mathrm{Rat}\,\mathcal{C}$.* ■

This last proposition may arouse some frustration.We have managed to free ourselves from the assumption that $\mathcal{A}$ is proper; why not from that that the *family* $\mathcal{C}$ is proper? After Theorem II.1.2, which our seemingly more general theorem should encompass, we should expect to be able to state that *a series s belongs to the $\mathbb{K}$-rational closure of a family of series $\mathcal{C}$ if and only if s is the defined behaviour of a $\mathbb{K}$-automaton over M of finite size whose entries belong to $\mathcal{C}\cup\mathbb{K}$*. However, this is not the case and if the result just quoted is valid, it is because it applies to the case where the semiring of weights is $\mathbb{B}$, which is special enough to entail further properties.

cf. p. 227

Consider for example (take $\mathbb{K}=\mathbb{Z}$ and $M=\{a\}^*$) the family $\mathcal{C}$ reduced to the single series[39] $s=-1+a$. The characteristic series $a=(s-s)^*+s$, and hence the characteristic series a^*, belong to $\mathbb{Z}\mathrm{Rat}\,\mathcal{C}$ but we would be hard put to find an automaton whose behaviour is defined and equal to a^*, and whose entries are either scalar or equal to s.

We must also return to the proof of Theorem 2.2, which not only proves the result but is also an effective means of constructing a $\mathbb{K}$-automaton from a $\mathbb{K}$-expression. The relationship with the constructions used in Chapter I for the proof of Proposition I.2.1 is obvious: Proposition 2.14 corresponds to the construction of normalised automata, and the matrix formulas (2.16), (2.20), (2.21), and (2.24) to constructions similar to those of §I.2.3.2.

cf. p. 87

We can continue the parallel and, as for automata over A^*, define *standard* $\mathbb{K}$-automata.

cf. Def. I.2.4, p. 92 and Sec. I.5.1, p. 145

Definition 2.8 A $\mathbb{K}$-automaton $\mathcal{A}=\langle I,E,T\rangle$ is *standard* if the initial vector I has a single non-zero coordinate i, equal to $1_{\mathbb{K}}$, and if this unique initial state i is not the destination of any transition whose label is non-zero. □

[39] $s=-1+a$ is a shorthand for $s=-1\cdot 1_{a^*}+1\cdot a$

In matrix terms, this means that $\mathcal{A}$ can be written

$$\mathcal{A} = \left\langle \begin{pmatrix} 1 & \boxed{\;0\;} \end{pmatrix}, \begin{pmatrix} 0 & \boxed{\;J\;} \\ \boxed{0} & \boxed{\;F\;} \end{pmatrix}, \begin{pmatrix} c \\ \boxed{U} \end{pmatrix} \right\rangle .$$

The definition does not forbid the initial state i from also being final; that is, the scalar c is not necessarily zero. If $\mathcal{A}$ is not only standard but also *proper*, c is the *constant term* of $|\mathcal{A}|$. The proof of Proposition 2.14 itself proves that: *cf. Exer. 2.24*

Proposition 2.17 *Every $\mathbb{K}$-automaton $\mathcal{A}$ over M whose behaviour is defined is equivalent to a* standard *proper automaton whose entries, other than the scalar entries of the initial and final vectors, are linear combinations of proper parts of the entries of $\mathcal{A}$.* ■

Again as in Chapter I, we prove: *cf. Sec. I.5.1, p. 145*

Proposition 2.18 *With every rational $\mathbb{K}$-expression E is canonically associated a standard $\mathbb{K}$-automaton $\mathcal{S}_{\mathsf{E}}$ of dimension $\ell(\mathsf{E}) + 1$.* ■

Furthermore, Properties 2.6 and 2.7, and the fact that every rational subset of A^* is recognised by an *unambiguous* automaton, imply the two following corollaries.

Corollary 2.19 *Let $\mathbb{K}$ be a* positive *semiring. The support of a $\mathbb{K}$-rational series over M is a rational subset of M.* ■

Corollary 2.20 *The characteristic series of a rational subset of A^* is an $\mathbb{N}$-rational series over A^*.* ■

Exercises

2.21 Prove the equalities (2.22), (2.25), (2.26) and (2.27).

2.22 Describe the matrix constructions (2.16), (2.20), (2.21), and (2.24) in terms of constructions on the graph representation of automata.

• 2.23 Using the above constructions, give a $\mathbb{Q}$-automaton over $\{a, b\}^*$ whose behaviour is equal to $\mathsf{E}_1 = ((\frac{1}{6}\, a^*) + (\frac{1}{3}\, b^*))^*$.

• 2.24 **Standard $\mathbb{K}$-automata.** (a) Write out the calculation that gives Proposition 2.17.

(b) Give matrix formulas equivalent to (2.16), (2.20), (2.21), and (2.24) which, applied to standard $\mathbb{K}$-automata, give standard $\mathbb{K}$-automata whose behaviours are respectively the product by a constant and the sum, product and star of the behaviours of the initial $\mathbb{K}$-automata. Deduce Proposition 2.18.

(c) Apply these formulas to construct a standard $\mathbb{Q}$-automaton whose behaviour is equal to $\mathsf{E}_1 = ((\frac{1}{6}\, a^*) + (\frac{1}{3}\, b^*))^*$.

2.25 Prove:

Proposition 2.21 *A rational subset of M is* unambiguous *if and only if it is realised by an automaton $\mathcal{A} = \langle Q, M, E, I, T \rangle$ over M whose characteristic $\mathbb{N}$-automaton realises a characteristic series.*

cf. p. 401 2.26 Find the error in the following reasoning: 'Equation (2.13) shows that when we divide a matrix m into a 2×2 matrix of blocks, we obtain for the star of m a rational expression of star height 2 in the blocks. By repeating the process by divisions that are as balanced as possible, we arrive at an expression of star height[40] $2\lceil \log_2 n\rceil$ (in the entries of m) where n is the dimension of m. Conclusion: every language recognised by an automaton of dimension n has a star height less than or equal to $2\lceil \log_2 n\rceil$.' (The conclusion is manifestly false.)

2.5 $\mathbb{K}$-coverings – $\mathbb{K}$-quotients

cf. Sec. II.3, p. 255 The idea of *automata coverings*, introduced and studied in the last chapter, seems at first glance perfectly suited to the treatment of weighted automata. In fact, if $\mathcal{A}$ and $\mathcal{B}$ are two automata, say over A^* to be more concrete, the existence of a covering $\varphi\colon \mathcal{B}\to\mathcal{A}$ implies not only that $\mathcal{B}$ is equivalent to $\mathcal{A}$, i.e. that $\mathcal{B}$ accepts the same subset of A^* as $\mathcal{A}$, but that there is a bijection between the successful computations of one and the other and hence that they are equivalent as $\mathbb{N}$-automata over A^*. In other words, they are equivalent *taking weights into account* with the natural assumption that the label of each of their transitions is given the coefficient $1_{\mathbb{N}}$.

We shall begin by observing that we need to give a more general definition to the notion of $\mathbb{K}$-covering to take account of these 'natural' cases of equivalence between $\mathbb{K}$-automata. We shall then make this explicit in matrix terms, which are most powerful in proofs, then we shall show how we can generalise the algorithm of minimisation by refinement of partitions to $\mathbb{K}$-automata. We shall end with an application to the construction of *characteristic* automata in anticipation of what we shall develop in the next section for series over the free monoid.

In what follows, M is a graded monoid and $\mathbb{K}$ a strong semiring, *a priori* arbitrary, even if in the examples we always take A^* as the monoid and $\mathbb{N}$ as the semiring of coefficients.

2.5.1 From coverings to $\mathbb{K}$-coverings

If, as we have just said, the idea of a covering is consistent with that of a weighted automaton, a moment of reflection and the observation of a few examples such as those of Figure 2.7 and Figure 2.8 show that it relates to a fairly naive conception of $\mathbb{K}$-automata.

In fact, and by definition, a covering of $\mathcal{B}$ over $\mathcal{A}$ induces a *bijection between the outgoing bouquets* of the states of $\mathcal{B}$ and those of $\mathcal{A}$; and this notion of outgoing bouquet has disappeared in $\mathbb{K}$-automata where we consider that there is only a single transition between any two states, labelled by a polynomial from $\mathbb{K}\langle M\rangle$ (or a series from $\mathbb{K}\langle\langle M\rangle\rangle$, since coverings are not restricted to finite automata).

A first attempt to mimic outgoing (and incoming) bouquets in a $\mathbb{K}$-automaton $\mathcal{A}$ is to break out each coefficient of $\mathcal{A}$ into a sum of monomials each of which labels a transition.

[40] $\log_2 n$ is the logarithm to the base 2 of n; $\lceil x\rceil$ is the smallest integer greater than or equal to the real x.

This is what we did implicitly for the $\mathbb{N}$-automata in Figure 2.7. But this decomposition into monomials does not take account of all the additive structure of $\mathbb{K}\langle M\rangle$. Let us see how this works in another example, which will also help us to understand the formal definition of $\mathbb{K}$-coverings that follows.

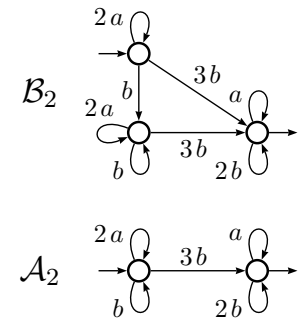


Fig. 2.7: A covering of ($\mathbb{K}$-)automata

Example 2.8 In anticipation of some automata that we shall see in the next section, we consider the two $\mathbb{N}$-automata $\mathcal{C}_2$ and $\mathcal{V}_2$ of Figure 2.8. There is an obvious morphism (the 'vertical' in the figure) from the *support automaton* of $\mathcal{C}_2$ to that of $\mathcal{V}_2$ which is not a covering (there is no bijection between $\mathsf{Out}_{\mathcal{C}_2}(j)$ and $\mathsf{Out}_{\mathcal{V}_2}(i)$), nor a co-covering (there is no bijection between $\mathsf{In}_{\mathcal{C}_2}(u)$ and $\mathsf{In}_{\mathcal{V}_2}(t)$).

These two $\mathbb{N}$-automata are however equivalent. The reason is that the *sum* of the transition labels which go from j to r and s, i.e. within *the set of states* whose image under φ is q, is equal to the label of the transition that goes from $i = j\varphi$ to q. This is what we shall now formalise. □

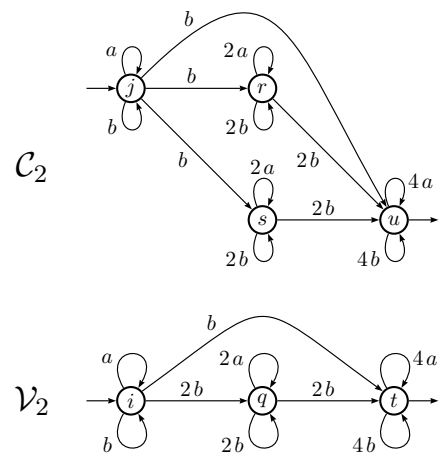


Figure 2.8: The 'vertical' is a morphism, not a covering

In what follows, $\mathcal{A} = \langle I, E, T\rangle$ and $\mathcal{B} = \langle J, F, U\rangle$ are two $\mathbb{K}$-automata over M, of dimension Q and R respectively.

Definition 2.9 A *surjective* map $\varphi\colon R \to Q$ determines a $\mathbb{K}$-covering of $\mathcal{B}$ over $\mathcal{A}$ – or makes $\mathcal{A}$ a $\mathbb{K}$-*quotient* of $\mathcal{B}$ – if, for each r in R and each q in Q, the *sum* of the transition labels that go from r to the *set of states* whose image under φ is q, is equal to the label of the transition that goes from $p = r\varphi$ to q; that is,

$$\forall p, q \in Q\,,\ \forall r \in p\varphi^{-1} \qquad E_{p,q} = \sum_{s \in q\varphi^{-1}} F_{r,s}\ , \tag{2.29}$$

and if the analogous conditions are satisfied by the initial and final vectors:

$$\forall q \in Q \qquad I_q = \sum_{s \in q\varphi^{-1}} J_s \quad (2.30) \qquad \text{and} \qquad \forall p \in Q\,,\ \forall r \in p\varphi^{-1} \qquad T_p = U_r\,. \quad (2.31)$$

Remark 2.10 The terminology used conceals a trap. We have defined $\mathbb{K}$-coverings, but the notion of $\mathbb{K}$-morphisms makes no sense while automata coverings are special cases of automata morphisms. For this reason, if the terms 'automaton' and '$\mathbb{B}$-automaton' are synonymous this natural identity does not extend to coverings, and a $\mathbb{B}$-covering is not a covering. By restricting Definition 2.9 to $\mathbb{K} = \mathbb{B}$, we see that a $\mathbb{B}$-covering is a *totally surjective morphism*, i.e. if $\mathcal{A}$ is a $\mathbb{B}$-*quotient* of $\mathcal{B}$, it is indeed a *quotient* of $\mathcal{B}$. The terminology is consistent with respect to quotients. □

cf. Def. II.3.2, p. 262

2.5.2 Matrix description

Another way to express that φ is a $\mathbb{K}$-covering of $\mathcal{B}$ over $\mathcal{A}$ rests more globally on transition matrices. We take the matrix F and form the sum of all the *columns* whose indices have the same image under φ. Then, φ is a $\mathbb{K}$-covering if, in the resultant matrix, two rows whose indices have the same image under φ are equal.

Example 2.8 (continued) If we write $\mathcal{C}_2 = \langle J_2, F_2, U_2 \rangle$ and $\mathcal{V}_2 = \langle I_2, E_2, T_2 \rangle$ (of dimension R_2 and Q_2 respectively), we have

$$F_2 = \underbrace{\begin{pmatrix} a+b & b & b & b \\ 0 & 2a+2b & 0 & 2b \\ 0 & 0 & 2a+2b & 2b \\ 0 & 0 & 0 & 4a+4b \end{pmatrix}}_{R_2} \left.\vphantom{\begin{pmatrix} a \\ 0 \\ 0 \\ 0 \end{pmatrix}}\right\} R_2$$

We form the sum of the median columns and obtain the matrix

$$\underbrace{\begin{pmatrix} a+b & 2b & b \\ 0 & 2a+2b & 2b \\ 0 & 2a+2b & 2b \\ 0 & 0 & 4a+4b \end{pmatrix}}_{Q_2} \left.\vphantom{\begin{pmatrix} a \\ 0 \\ 0 \\ 0 \end{pmatrix}}\right\} R_2$$

whose median rows are equal to that of

$$E_2 = \underbrace{\begin{pmatrix} a+b & 2b & b \\ 0 & 2a+2b & 2b \\ 0 & 0 & 4a+4b \end{pmatrix}}_{Q_2} \left.\vphantom{\begin{pmatrix} a \\ 0 \\ 0 \end{pmatrix}}\right\} Q_2$$

□

Definition 2.9 can then be rewritten in matrix form. To a surjective map $\varphi\colon R \to Q$ we relate the row-monomial $R \times Q$ matrix H_φ defined by

$$(H_\varphi)_{r,q} = \begin{cases} 1 & \text{if} \quad r\varphi = q \ ; \\ 0 & \text{otherwise} \ . \end{cases}$$

Since φ is surjective, each column of H_φ contains at least one 1.

From H_φ, we construct a matrix K_φ by transposing H_φ and by cancelling certain entries in such a way that K_φ is row monomial (with exactly one 1 per row); K_φ is not uniquely determined by φ but also depends on the choice of a 'representative' in each class for the map equivalence of φ.

Example 2.8 (continued) If φ_2 is the map from $\{j, r, s, u\}$ to $\{i, q, t\}$ such that $j\varphi_2 = i$, $u\varphi_2 = t$ and $r\varphi_2 = s\varphi_2 = q$, then

$$H_{\varphi_2} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad K_{\varphi_2} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} . \quad \square$$

Definition 2.10 Let $\mathcal{B} = \langle J, F, U \rangle$ be a $\mathbb{K}$-automaton of dimension R. A surjective map $\varphi\colon R \to Q$ is a $\mathbb{K}$-*covering* of $\mathcal{B}$ over $\mathcal{A} = \langle I, E, T \rangle$, a $\mathbb{K}$-automaton of dimension Q defined by

$$E = K_\varphi \cdot F \cdot H_\varphi \ , \qquad I = J \cdot H_\varphi \qquad \text{and} \qquad T = K_\varphi \cdot U \ , \tag{2.32}$$

if $\mathcal{B}$ satisfies the two equations

$$H_\varphi \cdot K_\varphi \cdot F \cdot H_\varphi = F \cdot H_\varphi \tag{2.33}$$

and

$$H_\varphi \cdot K_\varphi \cdot U = U \ . \tag{2.34}$$

Too many definitions can spoil things and we must show that Definitions 2.9 and 2.10 coincide.

The multiplication of an $X \times R$ matrix Z by H_φ *on the right* produces an $X \times Q$ matrix whose column q is the sum of the columns of Z with index s such that $s\varphi = q$. The multiplication of an $R \times X$ matrix Z by K_φ *on the left* produces a $Q \times X$ matrix whose row p is chosen from the rows of Z with index r such that $r\varphi = p$.

Hence the equivalence of the definition of $\mathcal{A}$ by (2.32) and by Equations (2.29), (2.30) and (2.31), from which we have removed the quantifier over r.

The multiplication of a $Q \times X$ matrix Z by H_φ *on the left* produces an $R \times X$ matrix whose row r is equal to the row with index $r\varphi$ of Z. Thus multiplication of an $R \times X$ matrix Z by $H_\varphi \cdot K_\varphi$ *on the left* comes down to replacing, for each p in Q, all the rows of Z with index r, such that $r\varphi = p$, by one of them chosen arbitrarily (the choice is fixed by the matrix K_φ).

We thus obtain the equivalence between (2.33) and (2.29), and between (2.34) and (2.31).

The definition of $\mathbb{K}$-coverings is conceived in the light of the following property:

Proposition 2.22 *Every $\mathbb{K}$-automaton is equivalent to each of its $\mathbb{K}$-coverings.*

Proof. If $\varphi\colon \mathcal{B} \to \mathcal{A}$ is a $\mathbb{K}$-covering, we have, for each n in $\mathbb{N}$

$$\begin{aligned} I \cdot E^n \cdot T &= J \cdot H_\varphi \cdot (K_\varphi \cdot F \cdot H_\varphi)^n \cdot K_\varphi \cdot U && \text{by (2.32)} \\ &= J \cdot (H_\varphi \cdot K_\varphi \cdot F)^n \cdot H_\varphi \cdot K_\varphi \cdot U \\ &= J \cdot F^n \cdot H_\varphi \cdot K_\varphi \cdot U && \text{by (2.33)} \\ &= J \cdot F^n \cdot U \ , && \text{by (2.34)} \end{aligned}$$

which ensures the equality $|\mathcal{B}| = J \cdot F^* \cdot U = I \cdot E^* \cdot T = |\mathcal{A}|$. ■

We cannot avoid being struck by the power of the matrix form of $\mathbb{K}$-coverings in this proof.

2.5.3 Co-$\mathbb{K}$-covering

As in Chapter II, the $\mathbb{K}$-covering has a dual notion, the *co-$\mathbb{K}$-covering*. Roughly, it involves summing the rows rather than the columns. More precisely, with the preceding notation, and using the matrix expression directly as our definition:

Definition 2.11 Let $\mathcal{B} = \langle J, F, U \rangle$ be a $\mathbb{K}$-automaton of dimension R. A surjective map $\varphi\colon R \to Q$ is a *co-$\mathbb{K}$-covering* of $\mathcal{B}$ over $\mathcal{A} = \langle I, E, T \rangle$, a $\mathbb{K}$-automaton of dimension Q defined by

$$E = H_\varphi^{\mathsf{t}} \cdot F \cdot K_\varphi^{\mathsf{t}} \ , \qquad I = J \cdot K_\varphi^{\mathsf{t}} \qquad \text{and} \qquad T = H_\varphi^{\mathsf{t}} \cdot U \tag{2.35}$$

if $\mathcal{B}$ satisfies the two equations

$$H_\varphi^{\mathsf{t}} \cdot F \cdot K_\varphi^{\mathsf{t}} \cdot H_\varphi^{\mathsf{t}} = H_\varphi^{\mathsf{t}} \cdot F \tag{2.36}$$

and

$$J \cdot K_\varphi^{\mathsf{t}} \cdot H_\varphi^{\mathsf{t}} = J \ . \tag{2.37}$$

In other words, φ is a co-$\mathbb{K}$-covering of $\mathcal{B}$ over $\mathcal{A}$ if, when in the matrix F we form the sum of all the *rows* whose indices have the same image under φ, we obtain a matrix in which two columns whose indices have the same image under φ are equal. We shall also say that $\mathcal{A}$ is a co-$\mathbb{K}$-quotient[41] of $\mathcal{B}$. With a proof that exactly resembles that of Proposition 2.22, we prove:

Proposition 2.23 *Every $\mathbb{K}$-automaton is equivalent to each of its co-$\mathbb{K}$-coverings.* ■

Example 2.8 (continued) We return to $\mathcal{C}_2 = \langle J_2, F_2, U_2 \rangle$:

$$F_2 = \begin{pmatrix} a+b & b & b & b \\ 0 & 2a+2b & 0 & 2b \\ 0 & 0 & 2a+2b & 2b \\ 0 & 0 & 0 & 4a+4b \end{pmatrix}$$

[41] Try not to spit.

whose two median rows we add, to give the matrix

$$\begin{pmatrix} a+b & b & b & b \\ 0 & 2a+2b & 2a+2b & 4b \\ 0 & 0 & 0 & 4a+4b \end{pmatrix}$$

whose columns r and s are equal to the median column of the matrix

$$E_2' = \begin{pmatrix} a+b & b & b \\ 0 & 2a+2b & 4b \\ 0 & 0 & 4a+4b \end{pmatrix}$$

which defines *another* $\mathbb{N}$-automaton $\mathcal{V}_2' = \langle\, I_2, E_2', T_2 \,\rangle$ equivalent to $\mathcal{C}_2$. □ *cf. also Exer. 2.17*

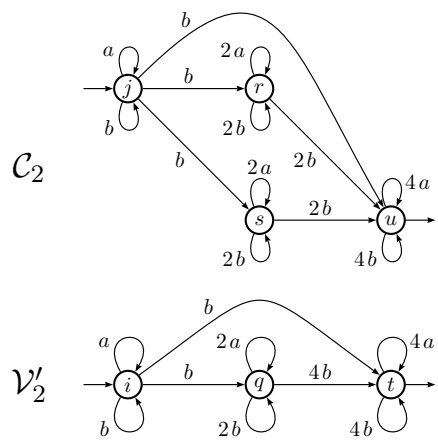


Figure 2.9: $\mathcal{C}_2$ is a co-$\mathbb{N}$-covering of $\mathcal{V}_2'$

Returning to our initial intuition, we can easily prove:

Property 2.8 *Let $\mathcal{A}$ and $\mathcal{B}$ be two automata over M and $\varphi\colon \mathcal{B} \to \mathcal{A}$ a covering (resp. a co-covering). Then for any $\mathbb{K}$, φ is a $\mathbb{K}$-covering (resp. a co-$\mathbb{K}$-covering) of the characteristic $\mathbb{K}$-automaton of $\mathcal{B}$ over the characteristic $\mathbb{K}$-automaton of $\mathcal{A}$.* ■

2.5.4 Minimal $\mathbb{K}$-quotient

If $\varphi\colon \mathcal{B} \to \mathcal{A}$ is a $\mathbb{K}$-covering, we also say, more correctly in fact, that $\mathcal{A}$ is a $\mathbb{K}$-quotient of $\mathcal{B}$. The following proposition states that every $\mathbb{K}$-automaton has a minimal $\mathbb{K}$-quotient.

Theorem 2.3 *Let $\mathcal{B}$ be a $\mathbb{K}$-automaton of finite dimension over M. Among all the $\mathbb{K}$-quotients of $\mathcal{B}$ (resp. among all the co-$\mathbb{K}$-quotients of $\mathcal{B}$), there exists one, unique up to isomorphism and effectively computable from $\mathcal{B}$, which has a minimal number of states and of which all these $\mathbb{K}$-automata are $\mathbb{K}$-coverings (resp. co-$\mathbb{K}$-coverings).*

Proof. A surjective map $\varphi\colon R \to Q$ defines a $\mathbb{K}$-covering $\varphi\colon \mathcal{B} \to \mathcal{A}$ if Equations (2.33) and (2.34) (which do not involve $\mathcal{A}$) are satisfied.

To prove the existence of a minimal $\mathbb{K}$-quotient, it suffices to show that if $\varphi\colon R \to Q$ and $\psi\colon R \to P$ are two maps that define $\mathbb{K}$-coverings, the map $\omega\colon R \to S$ also defines a $\mathbb{K}$-covering, where $\omega = \varphi \vee \psi$ is the map whose map equivalence is the upper bound of those of φ and ψ; that is, the finest equivalence which is coarser than the map equivalences of φ and ψ. In other words, there exist $\varphi'\colon Q \to S$ and $\psi'\colon P \to S$ such that $\omega = \varphi\varphi' = \psi\psi'$ and each class modulo $\omega = \varphi \vee \psi$ can be seen at the same time as a union of classes modulo φ and as a union of classes modulo ψ. It follows that

$$F \cdot H_\omega = F \cdot H_\varphi \cdot H_{\varphi'} = F \cdot H_\psi \cdot H_{\psi'} \quad , \tag{2.38}$$

and if two states p and r of R are congruent modulo ω, there exists q such that $p\varphi = q\varphi$ and $q\psi = r\psi$ (in fact a sequence of states q_i *etc.*). Rows p and q of $F \cdot H_\varphi$ are equal, and rows q and r of $F \cdot H_\psi$ are equal, hence, by (2.38), rows p and r of $F \cdot H_\omega$ are too.

To compute this minimal $\mathbb{K}$-quotient we can proceed by successive refinements of partitions, exactly as for the computation of the minimal automaton of a language from a deterministic automaton which recognises the language.

cf. § *I.3.3.3, p. 114*

In what follows the maps φ_i are identified with their map equivalences; the image is irrelevant. A state r of R is identified with the row vector of dimension R, characteristic of r and treated as such. For example, $r\varphi = s\varphi$ can be written $r \cdot H_\varphi = s \cdot H_\varphi$.

The maps φ_0 have the same map equivalence as U; that is,

$$r \cdot H_{\varphi_0} = s \cdot H_{\varphi_0} \Leftrightarrow r \cdot U = s \cdot U \quad ,$$

which can also be written

$$H_{\varphi_0} \cdot K_{\varphi_0} \cdot U = U \quad , \tag{2.39}$$

and the same equation holds for every map finer than φ_0. For each i, φ_{i+1} is finer than φ_i and, by definition, r and s are joint in φ_i (that is, $r \cdot H_{\varphi_i} = s \cdot H_{\varphi_i}$) and disjoint in φ_{i+1} if $r \cdot F \cdot H_{\varphi_i} \neq s \cdot F \cdot H_{\varphi_i}$. Let j be the index such that $\varphi_{j+1} = \varphi_j$ – that is, such that

$$r \cdot H_{\varphi_j} = s \cdot H_{\varphi_j} \quad \Longrightarrow \quad r \cdot F \cdot H_{\varphi_j} = s \cdot F \cdot H_{\varphi_j} \quad , \tag{2.40}$$

which can be rewritten

$$H_{\varphi_j} \cdot K_{\varphi_j} \cdot F \cdot H_{\varphi_j} = F \cdot H_{\varphi_j} \quad . \tag{2.41}$$

By (2.39) and (2.41), φ_j is a $\mathbb{K}$-covering.

Conversely, every $\mathbb{K}$-covering ψ satisfies (2.34) and is hence finer than φ_0. Then, for all i, if ψ is finer than φ_i it must also be finer than φ_{i+1}. In fact, if r and s are joint in ψ, it follows that $r \cdot H_\psi = s \cdot H_\psi$ and hence also $r \cdot H_{\varphi_i} = s \cdot H_{\varphi_i}$ since φ_i is coarser than ψ, and hence r and s are joint in φ_{i+1}: ψ is finer than φ_j which is thus the coarsest $\mathbb{K}$-covering. ∎

Remark 2.11 As we already said in Chapter II, the minimal automaton of a language and the minimal $\mathbb{K}$-quotient of a $\mathbb{K}$-automaton, computed with the *same* algorithm, are nevertheless fundamentally different: the first automaton is canonically associated with the language, the second with the $\mathbb{K}$-automaton we started from, and not with its behaviour. □

cf. Rem. II.3.6, p. 264
cf. also Fig. II.3.3, p. 262

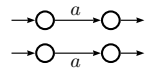
Remark 2.12 The notions of $\mathbb{K}$-covering and co-$\mathbb{K}$-covering, and the operations on rows and column of matrices which describe automata, bring to light a profound resemblance to the techniques of *state amalgamation* and *state splitting* for treating *dynamic symbolic systems*, for readers familiar with this subject. These links will be more precisely described in the notes section. □

cf. p. 519

Exercises

2.27 Verify Property 2.8.

• 2.28 Calculate all the $\mathbb{N}$-quotients and all the co-$\mathbb{N}$-quotients of the adjacent automaton.



• 2.29 **Characteristic automata.** We shall say that an $\mathbb{N}$-automaton $\mathcal{A} = \langle I, E, T \rangle$ (over an arbitrary graded monoid M) is *characteristic* if I, E and T are characteristic; that is, if each of their entries is a characteristic series.

(a) Prove:

Proposition 2.24 *Every $\mathbb{N}$-automaton over a graded monoid M is equivalent to a characteristic $\mathbb{N}$-automaton.*

(b) Compute the characteristic $\mathbb{N}$-automaton equivalent to the automaton $\mathcal{C}_1$ (*cf.* Example 2.5).

• 2.30 **Equisubtractive semirings.** A (commutative) monoid is called *equisubtractive* if for all p, q, r and s such that $p + q = r + s$ there exist x, y, z and t such that $p = x + y$, $q = z + t$, $r = x + z$ and $s = y + t$. A semiring is equisubtractive if it is so as a monoid for addition.

(a) Verify that if $\mathbb{K}$ is equisubtractive, then so are $\mathbb{K}\langle M\rangle$ and $\mathbb{K}\langle\langle M\rangle\rangle$.

(b) Verify that if $\mathbb{K}$ is equidivisible as a monoid for addition, the same is not necessarily true[42] of $\mathbb{K}\langle M\rangle$.

(c) Let $\mathbb{K}$ be an equisubtractive ring and $l_1, l_2, \ldots, l_n, k_1, k_2, \ldots, k_m$ elements of $\mathbb{K}$ such that

$$l_1 + l_2 + \cdots + l_n = k_1 + k_2 + \cdots + k_m \ .$$

Show that there exists an $n \times m$ matrix h such that the sum of the entries of each row i is equal to l_i and the sum of the entries of each column j is equal to k_j.

(d) Prove:

Proposition 2.25 *Let $\mathbb{K}$ be an* equisubtractive *semiring. If $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are three $\mathbb{K}$-automata over M such that $\mathcal{A}$ is a $\mathbb{K}$-covering of $\mathcal{C}$ and $\mathcal{B}$ is a co-$\mathbb{K}$-covering of $\mathcal{C}$, there exists a $\mathbb{K}$-automaton $\mathcal{D}$ which is both a co-$\mathbb{K}$-covering of $\mathcal{A}$ and a $\mathbb{K}$-covering of $\mathcal{B}$.* ■

[42] This is obviously the reason why we introduced the property of equisubtractivity.

3 $\mathbb{K}$-recognisable series

As in the last section, $\mathbb{K}$ denotes a strong semiring and M a graded monoid, *a priori* arbitrary. We shall now consider another family of series of $\mathbb{K}\langle\langle M\rangle\rangle$ (other than $\mathbb{K}\mathrm{Rat}\,M$), whose relationship to $\mathbb{K}\mathrm{Rat}\,M$ will be comparable to that between the family of recognisable subsets and that of rational subsets in $\mathfrak{P}(M)$.

3.1 $\mathbb{K}$-representations

A $\mathbb{K}$-*representation of* M *of dimension* Q is a *morphism* μ from M to the semiring of square matrices of dimension Q with entries in $\mathbb{K}$. By definition, in fact so that we can multiply the matrices, the dimension Q is *finite*. A $\mathbb{K}$-representation of M (of dimension Q) is also the name we give a *triple* (λ, μ, ν) where, as before,

$$\mu\colon M \longrightarrow \mathbb{K}^{Q\times Q}$$

is a morphism and where λ and ν are two vectors:

$$\lambda \in \mathbb{K}^{1\times Q} \qquad \text{and} \qquad \nu \in \mathbb{K}^{Q\times 1} \ ;$$

that is, λ is a *row* vector and ν a *column* vector, of dimension Q, with coefficients in $\mathbb{K}$. Such a representation defines a map from M to $\mathbb{K}$ by

$$\forall m \in M \qquad m \longmapsto \lambda \cdot m\mu \cdot \nu \ ;$$

that is, a *series* s:

$$s = \sum_{m\in M} (\lambda \cdot m\mu \cdot \nu)\, m \ .$$

A series s of $\mathbb{K}\langle\langle M\rangle\rangle$ is *realised*, or *recognised*, by the representation (λ, μ, ν). We also say that (λ, μ, ν) *realises*, or *recognises*, the series s.

Definition 3.1 A series of $\mathbb{K}\langle\langle M\rangle\rangle$ is $\mathbb{K}$-*recognisable* if it is recognised by a $\mathbb{K}$-representation. The set of $\mathbb{K}$-recognisable series over M is written $\mathbb{K}\mathrm{Rec}\,M$. □

Example 3.1 Take $\mathbb{K} = \mathbb{N}$ and $M = \{a, b\}^*$. Let $(\lambda_1, \mu_1, \nu_1)$ be the representation defined by

$$a\mu_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \ , \quad b\mu_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \ , \quad \lambda_1 = \begin{pmatrix} 1 & 0 \end{pmatrix} \quad \text{and} \quad \nu_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \ .$$

For all f in $\{a, b\}^*$, we verify that

$$\lambda_1 \cdot f\mu_1 \cdot \nu_1 = |f|_b \ , \tag{3.1}$$

cf. Exam. 2.4, p. 406 hence the series $t_1 = \sum_{f\in A^*} |f|_b\, f$ is $\mathbb{N}$-recognisable. □

cf. Sec. II.2.1, p. 234 **Remark 3.1** Definition 3.1 coincides, for $\mathbb{K} = \mathbb{B}$, with the definition of the *recognisable subsets* of a monoid (Definition II.2.1). If s is a $\mathbb{B}$-recognisable series over M,

realised by the representation (λ, μ, ν), then $\mu\colon M \to \mathbb{B}^{Q\times Q}$ is a morphism from M *to a finite monoid.* We set $R = \{r \in \mathbb{B}^{Q\times Q} \mid \lambda\cdot r\cdot\nu = 1_{\mathbb{B}}\}$: then the series s of $\mathbb{B}\langle\langle M\rangle\rangle$,

$$s = \sum_{m\in M} (\lambda\cdot m\mu\cdot\nu)\, m \ ,$$

can be seen as the *subset* $s = R\mu^{-1}$ of M. Proposition II.2.1 makes the link with Definition II.2.1.

Conversely, let P be a subset of M recognised by an action δ of M on S, with s_0 the distinguished element and R the subset of S such that $P = \{m \in M \mid s_0\cdot m \in R\}$. We define a $\mathbb{B}$-representation (λ, μ, ν) by

$$\forall m \in m \qquad m\mu_{s,t} = \begin{cases} 1 & \text{if} \quad t = s\cdot m \ , \\ 0 & \text{otherwise,} \end{cases}$$

$$\lambda_s = \begin{cases} 1 & \text{if} \quad s = s_0 \ , \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \nu_s = \begin{cases} 1 & \text{if} \quad s \in R \ , \\ 0 & \text{otherwise.} \end{cases}$$

The assumption that δ is an action implies that μ is a morphism from M to $\mathbb{B}^{Q\times Q}$, and the verification that $\lambda\cdot m\mu\cdot\nu = 1$ if and only if $s_0\cdot m \in R$ is immediate. □ *cf. Equat. (II.2.1), p. 232*

Note also that these definitions and the following two properties of $\mathbb{K}\mathrm{Rec}\, M$ do not involve multiplication in $\mathbb{K}\langle\langle M\rangle\rangle$, and are hence valid without even requiring that M be graded. *cf. Sec. 1.1*

Proposition 3.1 *Every finite linear combination, with coefficients in $\mathbb{K}$, of $\mathbb{K}$-recognisable series over M is a $\mathbb{K}$-recognisable series.*

Proof. Let s and t be two $\mathbb{K}$-recognisable series over M, respectively recognised by the $\mathbb{K}$-representations (λ, μ, ν) and (η, κ, ζ). For all k in $\mathbb{K}$ the series $k\,s$ is recognised by the representation $(k\,\lambda, \mu, \nu)$, the series $s\,k$ by the representation $(\lambda, \mu, \nu\,k)$, and the series $s + t$ by the representation (ε, π, ξ) defined by the following block decomposition:

$$\varepsilon = \begin{pmatrix} \lambda & \eta \end{pmatrix} \ , \quad m\pi = \begin{pmatrix} m\mu & 0 \\ 0 & m\kappa \end{pmatrix} \ , \quad \xi = \begin{pmatrix} \nu \\ \zeta \end{pmatrix} \ .$$

It is clear that π is indeed a representation of M. ■

In other words, $\mathbb{K}\mathrm{Rec}\, M$ is a (left and right) *sub-$\mathbb{K}$-module* of $\mathbb{K}\langle\langle M\rangle\rangle$.

Let $\mathbb{K}$ and $\mathbb{L}$ be two semirings and $\varphi\colon \mathbb{K} \to \mathbb{L}$ a morphism (of semirings) which extends to a morphism $\varphi\colon \mathbb{K}\langle\langle M\rangle\rangle \to \mathbb{L}\langle\langle M\rangle\rangle$ by $\langle s\varphi, m\rangle = \langle s, m\rangle\varphi$ for all s in $\mathbb{K}\langle\langle M\rangle\rangle$ and all m in M. It is immediate that if (λ, μ, ν) is a representation of the series s of $\mathbb{K}\langle\langle M\rangle\rangle$, then $(\lambda\varphi, \mu\varphi, \nu\varphi)$ is a representation of $s\varphi$. This is expressed by the following proposition: *cf. p. 391*

Proposition 3.2 *Let $\varphi\colon \mathbb{K} \to \mathbb{L}$ be a morphism of semirings. The image under φ of a $\mathbb{K}$-recognisable series over M is an $\mathbb{L}$-recognisable series over M.* ■

Corollary 3.3 *If $\mathbb{K}$ is a positive semiring, the support of a $\mathbb{K}$-recognisable series over M is a recognisable subset of M.* ■

cf. Sec. 3.3, p. 433

We could now prove that Kleene's Theorem for languages extends to a theorem for the series over a free monoid. We instead prefer first to pursue the study of $\mathbb{K}$-recognisable series over an arbitrary graded monoid.

Exercises

3.1 Verify Equation (3.1).

cf. Propty II.2.1, p. 235; for $\odot$, see Def. 1.2, p. 382.

• 3.2 Prove:

Property 3.1 *Let s be a $\mathbb{K}$-recognisable series over M and N a submonoid of M. Then, $s \odot \underline{N}$ is a $\mathbb{K}$-recognisable series over N.*

• 3.3 Show that a series s of $\mathbb{K}\langle\langle M\rangle\rangle$ is $\mathbb{K}$-recognisable if and only if there exists an integer n and a morphism $\mu\colon M \to \mathbb{K}^{n\times n}$ such that

$$\forall m \in M \setminus 1_M \qquad \langle s, m\rangle = (m\mu)_{1,n} \quad . \tag{3.2}$$

cf. Exer. 0.6.7, p. 32 for the definition of the trace of a matrix.

• 3.4 Show that a series s of $\mathbb{K}\langle\langle M\rangle\rangle$ is $\mathbb{K}$-recognisable if and only if there exists an integer n, a matrix z of $\mathbb{K}^{n\times n}$ and a morphism $\mu\colon M \to \mathbb{K}^{n\times n}$ such that

$$\forall m \in M \qquad \langle s, m\rangle = \mathsf{Tr}(z \cdot m\mu) \quad . \tag{3.3}$$

3.2 Products

The Cauchy product was tricky to define, necessitating the assumption of gradation. The two products that we shall now consider, the Hadamard and shuffle products, do not pose this sort of problem. These products are defined on general series, but it is their effect on recognisable series which will interest us, and we first define a product on *representations*.

3.2.1 Tensor product of $\mathbb{K}$-representations

The operation that will play the distinguished role for representations that the cartesian product played in the first two chapters (product of automata, product of actions) is the *tensor product*.

Definition 3.2 [**Tensor product of matrices**] Let X be a matrix of dimension $P \times Q$ and Y a matrix of dimension $R \times S$ (with entries in the same semiring $\mathbb{K}$). The *tensor product* of X by Y, written $X \otimes Y$, is a matrix of dimension $(P \times R) \times (Q \times S)$ defined by

$$\forall p \in P\,,\ \forall q \in Q\,,\ \forall r \in R\,,\ \forall s \in S \qquad X \otimes Y_{(p,r),(q,s)} = X_{p,q} Y_{r,s} \quad . \qquad \square$$

Example 3.2 Let $X = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ and $Y = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ be two matrices with entries in $\mathbb{N}$. Then

$$X \otimes Y = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 2 \end{pmatrix} \qquad \text{and} \qquad Y \otimes X = \begin{pmatrix} 1 & 0 & 2 & 0 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} \quad . \qquad \square$$

If $\mathbb{K}$ is *commutative*, the tensor product is also. The apparent difference between $X \otimes Y$ and $Y \otimes X$ above is due to the rows and columns being swapped.

3.2.2 Commutable sub-semirings

We shall need the tensor product to be commutative under more general assumptions. We shall say that two sub-semirings $\mathbb{I}$ and $\mathbb{L}$ of a non-commutative semiring $\mathbb{K}$ are *commutable* if every element of $\mathbb{I}$ commutes with every element of $\mathbb{L}$. For example, the *centre* of $\mathbb{K}$ and any sub-semiring of $\mathbb{K}$ are commutable. As another example, which is the one we shall use most, $1_{\mathbb{T}} \times \mathbb{T}$ and $\mathbb{T} \times 1_{\mathbb{T}}$ are two commutable sub-semirings[43] in $\mathbb{T} \times \mathbb{T}$.

Lemma 3.4 *Let X, Y, U and V be four matrices with entries in $\mathbb{K}$, respectively of dimension $P \times Q$, $P' \times Q'$, $Q \times R$ and $Q' \times R'$, and such that all the entries of Y commute with those of U. Then*

$$(X \otimes Y) \cdot (U \otimes V) = (X \cdot U) \otimes (Y \cdot V) \quad . \tag{3.4}$$

Proof. The 'dimension equation' of (3.4) is obviously satisfied. We then have, for all $\big((p,p'),(r,r')\big)$ in $(P \times P') \times (R \times R')$

$$\begin{aligned}
\big((X \otimes Y) \cdot (U \otimes V)\big)_{(p,p'),(r,r')} &= \sum_{(q,q') \in Q \times Q'} (X \otimes Y)_{(p,p'),(q,q')} \, (U \otimes V)_{(q,q'),(r,r')} \\
&= \sum_{(q,q') \in Q \times Q'} X_{p,q} \underbrace{Y_{p',q'} \, U_{q,r}}_{\text{since the entries of } Y \text{ and } U \text{ commute}} V_{q',r'} \\
&= \sum_{(q,q') \in Q \times Q'} X_{p,q} \, U_{q,r} \, Y_{p',q'} \, V_{q',r'} \\
&= \left(\sum_{q \in Q} X_{p,q} \, U_{q,r} \right) \left(\sum_{q' \in Q'} Y_{p',q'} \, V_{q',r'} \right) \\
&= (X \cdot U)_{p,r} \; (Y \cdot V)_{p',r'} \\
&= \big((X \cdot U) \otimes (Y \cdot V)\big)_{(p,p'),(r,r')}
\end{aligned}$$

■

[43] On the other hand, we shall not say that two matrices X and Y are *commutable* to mean that all the entries of X commute with those of Y; this would be too easily confused with the statement that the two matrices *commute*: that is, $XY = YX$.

We then obtain, with no difficulties:

Proposition 3.5 [**Tensor product of representations**] *Let $\mathbb{I}$ and $\mathbb{L}$ be two commutable sub-semirings of $\mathbb{K}$. Let M and N be two arbitrary monoids and $\mu\colon M \to \mathbb{I}^{Q\times Q}$ and $\kappa\colon N \to \mathbb{L}^{R\times R}$ two representations. The map*[44] *$\mu \otimes \kappa$, defined for all (m,n) in $M\times N$ by*

$$(m,n)\,[\,\mu \otimes \kappa] = m\mu \otimes n\kappa$$

is a representation of $M\times N$ in $\mathbb{K}^{(Q\times R)\times(Q\times R)}$.

Proof. For all (m,n) and (m',n') in $M\times N$, we have:

$$\begin{aligned}\big((m,n)\,[\,\mu\otimes\kappa]\big)\cdot\big((m',n')\,[\,\mu\otimes\kappa]\big) &= \big(m\mu\otimes n\kappa\big)\cdot\big(m'\mu\otimes n'\kappa\big)\\ &= \big(m\mu\cdot m'\mu\big)\otimes\big(n\kappa\cdot n'\kappa\big)\\ &= (m\,m')\mu\otimes(n\,n')\kappa = (m\,m',n\,n')\,[\,\mu\otimes\kappa]\quad,\end{aligned}$$

since, under the proposition's assumptions, all the entries of $m'\mu$ commute with those of $n\kappa$. ■

3.2.3 Hadamard product

cf. Sec. 1.1, p. 379

We have said that the Hadamard product is to series (sets with multiplicity) what intersection is to sets, which only really makes sense if the semiring of coefficients is commutative. In the same way that the recognisable subsets of an arbitrary monoid are closed under intersection, the recognisable series are closed under the Hadamard product. This is what the following result states.

Theorem 3.1 *Let $\mathbb{K}$ be a commutative semiring and M an arbitrary monoid. The Hadamard product of two $\mathbb{K}$-recognisable series over M is a $\mathbb{K}$-recognisable series over M.*

Under the more precise assumptions of Proposition 3.5, we can state a more general result (which we shall use in all its generality later):

Theorem 3.2 [**Schützenberger**] *Let $\mathbb{I}$ and $\mathbb{L}$ be two commutable sub-semirings of $\mathbb{K}$ and M be a monoid. The Hadamard product of an $\mathbb{I}$-recognisable series over M and a $\mathbb{L}$-recognisable series over M is a $\mathbb{K}$-recognisable series over M.*

More precisely, if (λ,μ,ν) recognises s and (η,κ,ζ) recognises t, then $s\odot t$ is recognised by $(\lambda\otimes\eta,\ \mu\otimes\kappa,\nu\otimes\zeta)$.

[44] In the hope of making it easier to read formulas and equations, I use the tensor product symbol in two sizes: large ($\otimes$) for a 'real' tensor product between two matrices, and small ($\otimes$) for a 'symbolic' tensor product between two morphisms or representations.

Proof. First note that, since the map $m \mapsto (m,m)$ is a morphism from M to $M \times M$, Proposition 3.5 implies that the map $m \mapsto m\mu \otimes m\kappa$ is also a morphism, and we also write it $\mu \otimes \kappa$.

Let s be a series over M recognised by the $\mathbb{I}$-representation (λ, μ, ν) and t be a series over M recognised by the $\mathbb{L}$-representation (η, κ, ζ). By definition we have, for all m in M,

$$\langle s \odot t, m \rangle = (\lambda \cdot m\mu \cdot \nu)(\eta \cdot m\kappa \cdot \zeta) = (\lambda \cdot m\mu \cdot \nu) \otimes (\eta \cdot m\kappa \cdot \zeta) \quad , \tag{3.5}$$

the second equality expressing the product of two coefficients of $\mathbb{K}$ as the tensor product of two 1×1 matrices. Under the assumptions of the theorem, we can apply Lemma 3.4 (three times) and obtain

$$\langle s \odot t, m \rangle = (\lambda \otimes \eta) \cdot (m\mu \otimes m\kappa) \cdot (\nu \otimes \zeta) = (\lambda \otimes \eta) \cdot (m[\mu \otimes \kappa]) \cdot (\nu \otimes \zeta) \quad .$$

Since, again according to these assumptions, $\mu \otimes \kappa$ is a $\mathbb{K}$-representation, the series $s \odot t$ is recognisable, recognised by the stated representation. ∎

3.2.4 Commutable recognisable series

To shorten our propositions and make them apply more generally than to series with coefficients in a commutative semiring, we shall say that two *recognisable* series s and t of $\mathbb{K}\langle\langle M \rangle\rangle$ are *commutable* to mean that the assumptions of Theorem 3.2 are satisfied; that is, that s is in $\mathbb{I}\mathrm{Rec}\, M$ and t in $\mathbb{L}\mathrm{Rec}\, M$, where $\mathbb{I}$ and $\mathbb{L}$ are two commutable sub-semirings of $\mathbb{K}$. Theorem 3.2 then says: *the Hadamard product of two commutable $\mathbb{K}$-recognisable series is $\mathbb{K}$-recognisable.*

We shall also say that two series s and t are commutable if $s \in \mathbb{I}\langle\langle M \rangle\rangle$ and $t \in \mathbb{L}\langle\langle M \rangle\rangle$. But this use is misleading, and we must be aware that the assumption that s and t are two commutable recognisable series is a condition on the *representations* of s and t, which is strictly stronger than a condition on the *coefficients* (unless the semirings $\mathbb{I}$ and $\mathbb{L}$ have the Fatou property). Thus, for example, an arbitrary series s of $\mathbb{K}\langle\langle M \rangle\rangle$ and a characteristic series t are commutable, but are not necessarily commutable recognisable series, even if s and t are recognisable.

3.2.5 Tensor product of series

We now assume that $\mathbb{K}$ is commutative. The two semirings $\mathbb{K}\langle\langle M \rangle\rangle$ and $\mathbb{K}\langle\langle N \rangle\rangle$ are canonically subalgebras of $\mathbb{K}\langle\langle M \times N \rangle\rangle$; the injection is induced by

$$m \mapsto (m, 1_N) \qquad \text{and} \qquad n \mapsto (1_M, n) \quad ,$$

for all m in M and all n in N. Modulo this identification, a product $(k\,m)(l\,n)$ is written $k\,l\,(m,n)$ and the extension by linearity of this notation gives the following definition.

Definition 3.3 Let s be in $\mathbb{K}\langle\langle M\rangle\rangle$ and t be in $\mathbb{K}\langle\langle N\rangle\rangle$. The *tensor product* of s and t, written $s \otimes t$, is the series of $\mathbb{K}\langle\langle M \times N\rangle\rangle$ defined by[45]

$$\forall (m,n) \in M \times N \qquad \langle s \otimes t, (m,n)\rangle = \langle s, m\rangle \langle t, n\rangle \ . \qquad \square$$

cf. Cor. II.2.20, p. 254

Proposition 3.6 *A series s of $\mathbb{K}\langle\langle M \times N\rangle\rangle$ is recognisable if and only if there exists a finite family $\{r_i\}_{i\in I}$ of series of $\mathbb{K}\mathrm{Rec}\, M$ and a finite family $\{t_i\}_{i\in I}$ of series of $\mathbb{K}\mathrm{Rec}\, N$ such that*

$$s = \sum_{i\in I} r_i \otimes t_i \ .$$

Proof. If r is recognised by the representation (λ, μ, ν), the map $(m,n) \mapsto m\mu$ is also a morphism and the series r' of $\mathbb{K}\langle\langle M \times N\rangle\rangle$ defined by $\langle r', (m,n)\rangle = \lambda \cdot m\mu \cdot \nu = \langle r, m\rangle$ is recognisable. Likewise, if $t \in \mathbb{K}\mathrm{Rec}\, N$, the series t' of $\mathbb{K}\langle\langle M \times N\rangle\rangle$ defined by $\langle t', (m,n)\rangle = \langle t, n\rangle$ is recognisable. Definition 3.3 shows that

$$r \otimes t = r' \odot t' \ ,$$

which is thus recognisable and Proposition 3.1 implies that the condition is sufficient.

Conversely, suppose that s is recognised by (λ, μ, ν), a representation of $M \times N$ of dimension Q. The map $\mu' \colon M \to \mathbb{K}^{Q\times Q}$ defined by $m\mu' = (m, 1_N)\mu$ is a morphism. For each q in Q, let r_q be the series defined by

$$\forall m \in M \qquad \langle r_q, m\rangle = \left[\lambda \cdot m\mu'\right]_q \ ,$$

which is a recognisable series of $\mathbb{K}\langle\langle M\rangle\rangle$. Likewise, $\mu'' \colon N \to \mathbb{K}^{Q\times Q}$ defined by $n\mu'' = (1_M, n)\mu$ is a morphism and t_q defined by

$$\forall n \in N \qquad \langle t_q, n\rangle = \left[n\mu'' \cdot \nu\right]_q \ ,$$

is a recognisable series of $\mathbb{K}\langle\langle N\rangle\rangle$. Since for all (m,n) of $M \times N$ we have

$$\lambda \cdot (m,n)\mu \cdot \nu = \sum_{q\in Q} \left[\lambda \cdot m\mu'\right]_q \left[n\mu'' \cdot \nu\right]_q \ ,$$

it follows that

$$s = \sum_{q\in Q} r_q \otimes t_q \ . \qquad \blacksquare$$

Chapter IV is devoted to the study of series on a product $M \times N$ (and particularly on $A^* \times B^*$) so we shall stop here for the moment.

3.2.6 Shuffle product

We now suppose that M is a free monoid A^* (and that $\mathbb{K}$ is still commutative; usually $\mathbb{K} = \mathbb{N}$ but that is not required). We can give a version with multiplicity of the shuffle product (over A^*) in the following manner.

cf. Def. I.1.7, p. 65

[45] Since m and n are each identified with their characteristic series, it would not be absurd to write $m \otimes n$ for (m,n), then $\mathbb{K}\langle\langle M \otimes N\rangle\rangle$ for $\mathbb{K}\langle\langle M \times N\rangle\rangle$, a notation which is used by certain authors, but which we do not adopt. The reader familiar with this area of algebra will have recognised that $\mathbb{K}\langle\langle M \times N\rangle\rangle$ is isomorphic to the tensor product $\mathbb{K}\langle\langle M\rangle\rangle \otimes_{\mathbb{K}} \mathbb{K}\langle\langle N\rangle\rangle$.

Definition 3.4 For all f and g in A^*, the *shuffle* of f and g, written $f \between g$, is a polynomial of $\mathbb{K}\langle A^*\rangle$ defined by induction on $|f|+|g|$ by

$$\begin{aligned} \forall f \in A^* \qquad & f \between 1_{A^*} = 1_{A^*} \between f = f \ , \\ \forall f,g \in A^*\,,\ \forall a,b \in A \qquad & f\,a \between g\,b = (f\,a \between g)\ b + (f \between g\,b)\ a \ . \end{aligned}$$

The shuffle is extended 'by linearity' to $\mathbb{K}\langle\langle A^*\rangle\rangle$; that is,

$$\forall s,t \in \mathbb{K}\langle\langle A^*\rangle\rangle \qquad s \between t = \sum_{f,g\in A^*} \langle s,f\rangle\langle t,g\rangle f \between g \ . \qquad \square$$

Example 3.3

$$\begin{aligned} a\,b \between a\,b &= 4\,a\,a\,b\,b + 2\,a\,b\,a\,b \ , \\ a\,b \between b\,a &= a\,b\,a\,b + 2\,a\,b\,b\,a + 2\,b\,a\,a\,b + b\,a\,b\,a \ , \\ (1_{a^*} + a) \between a^* &= [a^*]^2 \ . \end{aligned} \qquad \square$$

We verify that Definition 3.4 is consistent with that given in Chapter I for the shuffle of two words and of two languages, and that the shuffle is an associative product, commutative and continuous over $\mathbb{K}\langle\langle A^*\rangle\rangle$. We then prove, by induction on the length of words, the following characterisation of the shuffle of two words.

Proposition 3.7 *Let $\chi\colon A^* \to \mathbb{K}\langle A^*\times A^*\rangle$ be the morphism (on monoids) defined by $a\chi = (a, 1_{A^*}) + (1_{A^*}, a)$, for all a in A^*. It then follows that*

$$\forall h \in A^* \qquad h\chi = \sum_{f,g\in A^*} \langle f \between g, h\rangle (f,g) \ . \tag{3.6}$$

■

Theorem 3.3 *Let $\mathbb{K}$ be a commutative semiring. The shuffle of two $\mathbb{K}$-recognisable series on A^* is a $\mathbb{K}$-recognisable series.* *cf. Prop. I.1.7, p. 65*

Proof. Let s and t be $\mathbb{K}$-recognisable series on A^*, respectively recognised by the $\mathbb{K}$-representations (λ,μ,ν) and (η,κ,ζ). For all h in A^* the definition yields

$$\begin{aligned} \langle s \between t, h\rangle &= \sum_{f,g\in A^*} \big(\langle s,f\rangle\langle t,g\rangle\big)\langle f \between g, h\rangle \\ &= \sum_{f,g\in A^*} \big((\lambda\cdot f\mu\cdot\nu)\,(\eta\cdot g\kappa\cdot\zeta)\big)\langle f \between g, h\rangle \\ &= \sum_{f,g\in A^*} \big((\lambda\otimes\eta)\cdot\big((f,g)[\,\mu\otimes\kappa]\big)\cdot(\nu\otimes\zeta)\big)\langle f \between g, h\rangle \\ &= (\lambda\otimes\eta)\cdot\big((h\chi)[\,\mu\otimes\kappa]\big)\cdot(\nu\otimes\zeta) \qquad \text{by Proposition 3.7} \ . \end{aligned}$$

Since, by the theorem's assumptions, $\chi\circ[\,\mu\otimes\kappa]$ is a $\mathbb{K}$-representation, the series $s \between t$ is recognisable. ■

Exercises

3.5 Let $A = \{a_1, a_2, \ldots, a_n\}$ be an alphabet with n letters and $P(x_1, x_2, \ldots, x_n)$ a polynomial in the commutative variables $x_1, x_2, \ldots, x_n$ with coefficients in $\mathbb{Z}$. Show that the series

$$s = \sum_{f \in A^*} P(|f|_1, |f|_2, \ldots, |f|_n)\, f$$

is $\mathbb{Z}$-recognisable (over A^*).

3.6 **Shuffle product.** (a) Verify the assertions that follow Definition 3.4: that it is consistent with the definition of the shuffle of two words and of two languages, and that shuffle is an associative, commutative and continuous product over $\mathbb{K}\langle\langle A^* \rangle\rangle$.

(b) Prove Proposition 3.7.

3.7 **Infiltration product.** An analogous step to that followed for the shuffle gives a definition of an infiltration product with multiplicity: for all f and g in A^*, $f \uparrow g$ is a polynomial of $\mathbb{K}\langle A^* \rangle$ defined by induction on $|f| + |g|$ by

$$\begin{aligned} \forall f \in A^* \qquad & f \uparrow 1_{A^*} = 1_{A^*} \uparrow f = f \ , \\ \forall f, g \in A^*,\ \forall a, b \in A \qquad & f\,a \uparrow g\,b = (f\,a \uparrow g)\ b + (f \uparrow g\,b)\ a + \delta_{a,b}\,(f \uparrow g)\ a \ , \end{aligned}$$

where $\delta_{a,b}$ is a Kronecker delta: $\delta_{a,b} = 1_{\mathbb{K}}$ if $a = b$, and $0_{\mathbb{K}}$ otherwise. The infiltration is extended 'by linearity' to $\mathbb{K}\langle\langle A^* \rangle\rangle$; that is,

$$\forall s, t \in \mathbb{K}\langle\langle A^* \rangle\rangle \qquad s \uparrow t = \sum_{f, g \in A^*} \langle s, f \rangle \langle t, g \rangle\, f \uparrow g \ . \tag{3.7}$$

(a) Compute $a\,b \uparrow a\,b$, $a\,b \uparrow b\,a$ and $a\,b \uparrow a\,c$.

cf. Exer. I.1.17, p. 65 (b) Show consistency with the definition given in Chapter I.

(c) Show that infiltration is an associative, commutative and continuous product over $\mathbb{K}\langle\langle A^* \rangle\rangle$.

(d) Let $\psi \colon A^* \to \mathbb{K}\langle A^* \times A^* \rangle$ be the morphism (on monoids) defined by

$$\forall a \in A \qquad a\psi = (a, 1_{A^*}) + (1_{A^*}, a) + (a, a) \ .$$

Show that

$$\forall h \in A^* \qquad h\psi = \sum_{f, g \in A^*} \langle f \uparrow g, h \rangle (f, g) \ .$$

(e) Prove:

cf. Prop. I.1.8, p. 66 **Theorem 3.4** *Let $\mathbb{K}$ be a commutative semiring. The infiltration product of two $\mathbb{K}$-recognisable series on A^* is a $\mathbb{K}$-recognisable series.*

3.8 Verify that the shuffle and infiltration of two commutable series can be defined like that of two series with elements in a commutative semiring, and that Theorems 3.3 and 3.4 extend to commutable $\mathbb{K}$-recognisable series.

3.9 **Binomial coefficients.** The *(generalised) binomial coefficient* of f and g (in A^*), written $\binom{f}{g}$, is the number of distinct subwords of f equal to g. For example:

$$\binom{a\,b\,a\,a\,b}{a\,b} = 4 \ .$$

(a) Justify the terminology. Prove

$$\forall f, g \in A^* \qquad \langle g \between \underline{A^*}, f \rangle = \binom{f}{g} \ . \tag{3.8}$$

(b) Let $\mu\colon A^* \to \mathbb{N}\langle A^* \rangle$ be the morphism (on monoids) defined by

$$\forall a \in A \qquad a\mu = 1_{A^*} + a \ .$$

Prove

$$\forall f, g \in A^* \qquad \langle f\mu, g \rangle = \binom{f}{g} \ . \tag{3.9}$$

(c) We write k for the cardinal of the alphabet A, I for the identity matrix of dimension k and U for the column vector of dimension k all of whose entries are 1.

Let V be the column vector (resp. the row vector) of dimension k whose entries are the letters of A, arranged according to a given total order $\leqslant$. Verify that $V \otimes V \otimes \cdots \otimes V$, where the product has n factors, is the column vector (resp. the row vector) of dimension k^n whose entries are the words of length n, arranged in the lexicographic order induced by $\leqslant$.

(d) For all p, q in $\mathbb{N}$, we write C_q^p for the $k^p \times k^q$ *matrix* on $\mathbb{N}$ whose rows are indexed by A^p and columns by A^q (in the lexicographic order) and whose (f, g) entry is equal to the generalised binomial coefficient $\binom{f}{g}$. Justify the notation and prove the generalisation of Pascal's formula:

$$\mathrm{C}_q^p = \mathrm{C}_q^{p-1} \otimes U + \mathrm{C}_{q-1}^{p-1} \otimes I \ .$$

3.3 The Kleene–Schützenberger Theorem

As mentioned above, Kleene's Theorem for languages extends to a theorem for series over a free monoid. *cf. Th. I.2.1, p. 87*

Theorem 3.5 [**Kleene–Schützenberger**] *Let $\mathbb{K}$ be a semiring, and A a finite alphabet. A series of $\mathbb{K}\langle\langle A^* \rangle\rangle$ is $\mathbb{K}$-rational if and only if it is $\mathbb{K}$-recognisable. That is,*

$$\mathbb{K}\mathrm{Rec}\, A^* = \mathbb{K}\mathrm{Rat}\, A^* \ .$$

As for the proof of Kleene's Theorem, we shall prove the two inclusions one at a time:

$$\mathbb{K}\mathrm{Rec}\, A^* \subseteq \mathbb{K}\mathrm{Rat}\, A^* \qquad \text{and} \qquad \mathbb{K}\mathrm{Rat}\, A^* \subseteq \mathbb{K}\mathrm{Rec}\, A^* \ . \tag{3.10}$$

Each of the inclusions is obtained from the Fundamental Theorem together with the freeness of A^*. This is used in both cases by means of the following result. *cf. Th. 2.2, p. 411*

Lemma 3.8 *Let $\mathbb{K}$ be a semiring and A a finite alphabet. Let Q be a finite set and $\mu\colon A^* \to \mathbb{K}^{Q \times Q}$ a morphism. We set*

$$X = \sum_{a \in A} (a\mu)\, a \ .$$

Then, for all f in A^*, we have $\langle X^*, f \rangle = f\mu$.

Proof. The matrix X is a proper series of $\mathbb{K}^{Q\times Q}\langle\langle A^*\rangle\rangle$ and hence X^* is defined. We first prove, by induction on the integer n, that

$$X^n = \sum_{f\in A^n} (f\mu)\, f \ ,$$

an equality trivially verified for $n = 0$, and true by definition for $n = 1$. It follows that

$$X^{n+1} = X^n \cdot X = \Big(\sum_{f\in A^n} (f\mu)\, f)\Big) \cdot \Big(\sum_{a\in A} (a\mu)\, a)\Big) \quad = \sum_{(f,a)\in A^n\times A} (f\mu \cdot a\mu)\, f\, a$$
$$= \sum_{(f,a)\in A^n\times A} (fa)\mu\, fa \ = \sum_{g\in A^{n+1}} (g\mu)\, g \ ,$$

since, for each integer n, A^{n+1} is in bijection with $A^n \times A$ as A^* is freely generated by A. For the same reason, A^* is the *disjoint* union of the A^n, for n in $\mathbb{N}$, and it follows, for all f in A^*, that

$$<X^*, f> = <X^{|f|}, f> = f\mu \ . \qquad \blacksquare$$

Remark 3.2 The assumption in the statement of Lemma 3.8 that A is *finite* is necessary because the star operation in $\mathbb{K}\langle\langle A^*\rangle\rangle$ has only been defined when A^* is *finitely generated*. We could in fact have avoided this (for this lemma, not for Theorem 3.5) with some preparation. □

cf. Exer. 3.13

Proof of Theorem 3.5. Each of the two inclusions (3.10) is proved in the form of a property.

Property 3.2 *If A is finite, $\mathbb{K}$-recognisable series on A^* are $\mathbb{K}$-rational.*

Proof. Let (λ, μ, ν) be a representation which recognises a series s; that is, $<s, f> = \lambda \cdot f\mu \cdot \nu$, for all f in A^*. Let $\langle\, \lambda, X, \nu\,\rangle$ be the automaton defined by

$$X = \sum_{a\in A} (a\mu)\, a \ .$$

By Lemma 3.8, we have

$$s = \sum_{f\in A^*} (\lambda \cdot f\mu \cdot \nu)\, f = \lambda \cdot \Big(\sum_{f\in A^*} (f\mu)\, f\Big) \cdot \nu = \lambda \cdot X^* \cdot \nu \ .$$

By Proposition 2.10, the series s belongs to the rational closure of the entries of X. These entries are finite linear combinations of elements of A since A *is finite*: s belongs to $\mathbb{K}\mathrm{Rat}\, A^*$. $\blacksquare$

Property 3.3 *The $\mathbb{K}$-rational series on A^* are $\mathbb{K}$-recognisable.*

Proof. By Theorem 2.2, the series s is the behaviour of a *proper* finite $\mathbb{K}$-automaton $\langle\, I, X, T\,\rangle$ such that the entries of X are finite linear combinations of elements of A

(and those of I and T are scalar). We can therefore write $X = \sum_{a\in A}(a\mu)a$ where $a\mu$ is the matrix of coefficients of the letter a in X. By Lemma 3.8, we have

$$\forall f \in A^* \qquad \langle s, f\rangle = \langle I \cdot X^* \cdot T, f\rangle = I \cdot f\mu \cdot T \ ,$$

and the series s is recognised by the *representation* (I, μ, T). ■

The two inclusions (3.10) prove the theorem. ■

Example 1.1 (continued) Examples 2.4 and 3.1 have already shown that the series t_1 , $t_1 = \sum_{f\in A^*} |f|_b\, f$, is both $\mathbb{N}$-rational and $\mathbb{N}$-recognisable. □

By Theorem 3.1, we therefore deduce:

Corollary 3.9 *If $\mathbb{K}$ is a commutative semiring, the Hadamard product of two $\mathbb{K}$-rational series on A^* is a $\mathbb{K}$-rational series.* ■

3.3.1 Commutable rational series

We can extend the preceding result with the same method as that used for recognisable series. We shall say that two *rational* series s and t of $\mathbb{K}\langle\langle M\rangle\rangle$ are *commutable* if $s \in \mathbb{I}\mathrm{Rat}\, M$ and $t \in \mathbb{L}\mathrm{Rat}\, M$ where $\mathbb{I}$ and $\mathbb{L}$ are two commutable sub-semirings of $\mathbb{K}$. As for recognisable series, the terminology is delicate and we must remember that to say that two series are 'commutable rational' is strictly stronger than to say that they are rational and commutable. We can then state:

Corollary 3.10 *The Hadamard product of two commutable $\mathbb{K}$-rational series on A^* is a $\mathbb{K}$-rational series.* ■

The Hadamard product of series generalises the intersection of languages; the Hadamard product of two $\mathbb{K}$-rational series is obtained by a construction which is the generalisation of that which gives the intersection of two languages (namely, the cartesian product of two automata) which we can easily verify coincides with the construction consisting of forming the tensor product of the $\mathbb{K}$-representations of the series realised by the automata. Figure 3.1 illustrates Proposition 3.11, with two arbitrarily chosen $\mathbb{N}$-automata.

cf. Prop. I.1.4, p. 61

Proposition 3.11 *Let $\mathcal{A}' = \langle\, Q', A, E', I', T'\,\rangle$ and $\mathcal{A}'' = \langle\, Q'', A, E'', I'', T''\,\rangle$ be two proper $\mathbb{K}$-automata on A (the initial and final vectors are scalar and the transitions are labelled with letters equipped with coefficients). The Hadamard product of $|\mathcal{A}'|$ and $|\mathcal{A}''|$ is recognised by the $\mathbb{K}$-automaton written $\mathcal{A}' \odot \mathcal{A}''$ and defined by*

$$\mathcal{A}' \odot \mathcal{A}'' = \langle\, Q' \times Q'', A, E, I' \otimes I'', T' \otimes T''\,\rangle \ ,$$

where the set E of transitions is described by

$$E = \big\{\big((p', p''), (k'k'')\, a, (q', q'')\big) \;\big|\; (p', k'\, a, q') \in E' \quad \textit{and} \quad (p'', k''\, a, q'') \in E''\big\} \ . \ ■$$

The generalisation to series of Propositions II.2.6 (the intersection of a recognisable subset and a rational subset is rational) and II.2.5 (a recognisable subset of a finitely generated monoid is rational) requires some preparation and a little precaution. For the second, the assumption 'M is finitely generated' needs to be replaced by the stronger assumption that 'M is rationally enumerable', in other words, M must be an unambiguous rational subset of itself; i.e., $M \in \mathrm{URat}\, M$.

cf. Exer. II.1.13

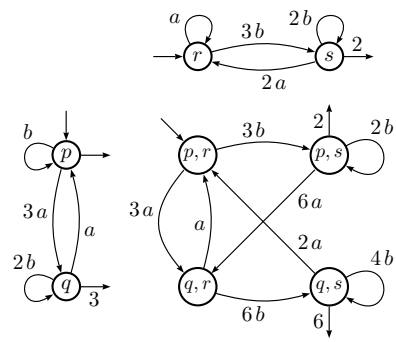


Figure 3.1: Hadamard product of two $\mathbb{N}$-automata

Proposition 3.12 *The Hadamard product of a $\mathbb{K}$-rational series and a $\mathbb{K}$-recognisable series which are commutable is a $\mathbb{K}$-rational series.*

cf. Rem. II.2.3

Hence we deduce:

Proposition 3.13 *If M is rationally enumerable, $\mathbb{K}\mathrm{Rec}\, M \subseteq \mathbb{K}\mathrm{Rat}\, M$.*

Proof. If M is rationally enumerable, M is an unambiguous rational subset of itself and $\underline{M}$ is a $\mathbb{K}$-rational series (Proposition 2.21). We have $s \odot \underline{M} = s$ for all s in $\mathbb{K}\langle\langle M \rangle\rangle$ and we apply Proposition 3.12. ∎

The proof of Proposition 3.12 requires some definitions which are better placed in the next chapter and we thus postpone it thither. In the same vein as Corollary 3.10 and Proposition 3.11, we also have:

cf. § IV.3.1.3, p. 561

cf. Exer. I.1.16 and Prop. I.1.7, p. 65.

Corollary 3.14 *The shuffle of two commutable $\mathbb{K}$-rational series on A^* is a $\mathbb{K}$-rational series.* ∎

cf. Exer. I.1.16, p. 181

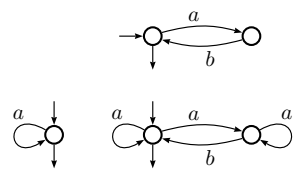
Proposition 3.15 *Let $\mathcal{A}' = \langle\, Q', A, E', I', T' \,\rangle$ and $\mathcal{A}'' = \langle\, Q'', A, E'', I'', T'' \,\rangle$ be two proper $\mathbb{K}$-automata over A (the initial and final vectors are scalar and the transitions are labelled by letters equipped with coefficients). The shuffle of $|\mathcal{A}'|$ and $|\mathcal{A}''|$ is realised by the $\mathbb{K}$-automaton written $\mathcal{A}' \between \mathcal{A}''$ and defined by*

$$\mathcal{A}' \between \mathcal{A}'' = \langle\, Q' \times Q'', A, E, I' \otimes I'', T' \otimes T'' \,\rangle \ ,$$

where the set E of transitions is described by

$$E = \big\{ \big((p', p''), k'\, a, (q', p'')\big) \;\big|\; (p', k'\, a, q') \in E' \quad \text{and} \quad p'' \in Q'' \big\}$$
$$\bigcup \big\{ \big((p', p''), k''\, a, (p', q'')\big) \;\big|\; p' \in Q' \quad \text{and} \quad (p'', k''\, a, q'') \in E'' \big\} \ . \quad ∎$$

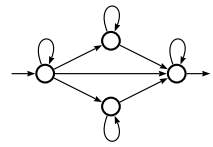
Example 3.4 The construction in the adjacent figure proves the identity[46] $a^* \between (a\,b)^* = (a\,(a^*)^2 b)^* a^*$.



Be aware that the computations are performed in $\mathbb{N}\langle\langle A^*\rangle\rangle$ and that $(a^*)^2 = \sum_{n\in\mathbb{N}}(n+1)\,a^n$. □

Exercises

• 3.10 Compute the behaviour of the $\mathbb{N}$-automaton over a^* in the adjacent figure (all the transitions are labelled a).

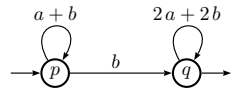


3.11 Prove Proposition 3.11.

• 3.12 Consider the $\mathbb{N}$-automaton $\mathcal{C}_1$ of Example 2.5, and use the conventions given there.

(a) Prove that the $\mathbb{N}$-automaton $\mathcal{C}_2$ of Example 2.8 'computes' $\bar{f}^2$; that is, is such that for all f, $\langle|\mathcal{C}_2|, f\rangle = \bar{f}^2$.

(b) Construct, using the method of Example 2.8, an $\mathbb{N}$-automaton $\mathcal{V}_3$ which has four states and which 'computes' $\bar{f}^3$.



The $\mathbb{N}$-automaton $\mathcal{C}_1$

(c) Generalise and construct, for each integer n and still with the same method, an $\mathbb{N}$-automaton $\mathcal{V}_n$ which has $n+1$ states and which computes $\bar{f}^n$:

$$\forall f \in A^* \qquad \langle|\mathcal{V}_n|, f\rangle = \bar{f}^n \quad .$$

(d) Compare with the automaton $\mathcal{V}'_n$ constructed in Exercise 2.17. *cf. p. 409*

3.13 Let A be an infinite countable alphabet (and $\mathbb{K}$ an arbitrary semiring). Prove that $\mathbb{K}\langle\langle A^*\rangle\rangle$ is a semiring.

(a) Define on $\mathbb{K}\langle\langle A^*\rangle\rangle$ a topology such that the star of a proper series is well defined (and satisfies the properties described in Section 2.1).

(b) Prove Lemma 3.8 without the assumption that A is finite.

(c) Let $\mathcal{C} = \mathfrak{P}(A)$, the family of (arbitrary) subsets of A. Show that

$$\mathbb{K}\mathrm{Rec}\, A^* = \mathbb{K}\mathrm{Rat}\, \mathcal{C} \quad .$$

3.14 Give a simple example of a monoid M and two $\mathbb{N}$-rational series on M whose Hadamard product is not a $\mathbb{N}$-rational series on M.

• 3.15 **Shuffle product.** (a) Prove Proposition 3.15.

(b) Prove the following identities:

(i) $a^* \between (-a)^* = 1$; (ii) $(1+a) \between a^* = (a^*)^2$; (iii) $(a\,b)^* \between (-a\,b)^* = (-4\,a^2 b^2)^*$.

(c) Spell out the construction of the *infiltration product* of two $\mathbb{K}$-automata $\mathcal{A}'$ and $\mathcal{A}''$ over A^*; that is, of the automaton $\mathcal{A}' \uparrow \mathcal{A}''$ such that $|\mathcal{A}' \uparrow \mathcal{A}''| = |\mathcal{A}'| \uparrow |\mathcal{A}''|$.

---

[46] *Note added in translation:* The computation of the behaviour of the automaton that is the result of the shuffle product should have been an exercise in Section 2. Hint: use identity **S**.

4 Series on a free monoid

Up to now we have developed the theory of series on the assumption that M is *graded* (so that we knew how to define star). In Section 3, we saw, with the Kleene–Schützenberger Theorem, that if M is free, then rational and recognisable series coincide. We will now require M to be a *free monoid* and ascertain all the consequences.

Without going over all the theory again, we can then give a different presentation, more algebraic, of some of the earlier results, and, continuing in this vein, obtain some fundamental properties which take their shape from the unique decomposition of the elements of the free monoid.

In the following, A is a finite alphabet, and $\mathbb{K}$ a strong semiring about which, to start with, we make no other assumption. In Section 4.2, we shall return to the problem of the construction of a finite automaton from a rational expression by the method of derivatives, where taking weights into account will give us remarkable power. In Section 4.3, we shall study the case where $\mathbb{K}$ is a *field*, which gives us a finished mathematical theory, and in particular an interpretation of the minimisation of automata.[47] The last subsection, which we could continue almost indefinitely, tackles various problems – in particular, *decision problems* – concerning the support of rational series on a free monoid and the coefficients of these series.

4.1 A characterisation of recognisable series

The principal notion that we shall use in this section is that of the *(left) quotient*[48] *of a series* which is a generalisation to series of the (left) quotient operation of a subset of a monoid (a free monoid in this case).

4.1.1 Quotients of series

The free monoid A^* *acts by quotient* on $\mathbb{K}\langle\langle A^*\rangle\rangle$: for all f in A^* and all series s in $\mathbb{K}\langle\langle A^*\rangle\rangle$, the series $f^{-1}s$ is defined by

$$f^{-1}s = \sum_{g\in A^*} \langle s, f\,g\rangle\, g\,;$$

that is, the coefficient of g in $f^{-1}s$ is equal to the coefficient of $f\,g$ in s:

$$\forall f, g \in A^* \qquad \langle f^{-1}s, g\rangle = \langle s, f\,g\rangle \ ,$$

and in particular

$$\forall f \in A^* \qquad \langle f^{-1}s, 1_{A^*}\rangle = \langle s, f\rangle \ . \tag{4.1}$$

[47] In Subsections 4.1 and 4.3, I essentially follow the presentation given by Jean Berstel and Christophe Reutenauer in their work *Rational Series and Their Languages* [31].

[48] Some authors speak of the *translation* of a series; I have preferred to use the same term as for languages.

If s is the characteristic series of a word g, then $f^{-1}g$ is the characteristic series of the word h such that $g = f\,h$; if $s = \underline{L}$ is the characteristic series of a language L, then $f^{-1}\underline{L}$ is the characteristic series of the language $f^{-1}L$: the definition is consistent with that of quotient which we gave in Chapter I. *cf. p. 62 and p. 112*

For a given f, the operation $s \mapsto f^{-1}s$ is *additive*:

$$f^{-1}(s+t) = f^{-1}s + f^{-1}t\,; \tag{4.2}$$

it *commutes with the exterior multiplications* of $\mathbb{K}$ on $\mathbb{K}\langle\langle A^*\rangle\rangle$:

$$f^{-1}(k\,s) = k\,(f^{-1}s) \qquad \text{and} \qquad f^{-1}(s\,k) = (f^{-1}s)\,k\,; \tag{4.3}$$

and it is *continuous*. These three properties ensure that the operation of quotient by f is entirely defined on $\mathbb{K}\langle\langle A^*\rangle\rangle$ by the operation of quotient by f on A^*.

Example 4.1 $(a\,b)^{-1}[a^2 + a\,b\,a^2 + 2a\,b\,a\,b + 3a\,b^2 + 2b] = a^2 + 2a\,b + 3b\,.$ □

Example 4.2 Let $t_1 = \sum_{f\in A^*} |f|_b\, f$, then $b^{-1}t_1 = \sum_{g\in A^*} (|g|_b + 1)\, g\,.$ □ *cf. Exam. 1.1, p. 380*

Considering Example 4.1 leads us directly to the following:

Property 4.1 *If p is a polynomial in $\mathbb{K}\langle A^*\rangle$, then for every word f in A^+, $f^{-1}p$ is also a polynomial, of lesser degree than p.* ■

Furthermore, this operation is, like the quotient of languages, a *(right) action* of A^* on $\mathbb{K}\langle\langle A^*\rangle\rangle$ since we have

$$\forall f, g \in A^*\,,\ \forall s \in \mathbb{K}\langle\langle A^*\rangle\rangle \qquad (f\,g)^{-1}s = g^{-1}(f^{-1}s)\ \ . \tag{4.4}$$

Finally, the fact that A^* is a *free* monoid enables us to prove:

Property 4.2 $\forall s \in \mathbb{K}\langle\langle A^*\rangle\rangle \qquad s = \langle s, 1_{A^*}\rangle + \sum_{a\in A} a\,(a^{-1}s)\ .$ ■

Remark 4.1 As for the quotient on languages, we can also define the *right* quotient of series:

$$\forall s \in \mathbb{K}\langle\langle A^*\rangle\rangle\,,\ \forall g \in A^* \qquad s\,g^{-1} = \sum_{f\in A^*} \langle s, f\,g\rangle\, f\ \ . \qquad \square$$

4.1.2 Stable modules

A language is rational (recognisable) if and only if the family of its quotients by the words of A^* is *finite*. The same property is not true, *stricto sensu*, for rational (recognisable) series. Let, for example, $s_2 = (2\,a)^*$, an $\mathbb{N}$-rational series on $\{a\}^*$; then, for all integers n, we have

$$(a^n)^{-1}s_2 = 2^n\, s_2$$

and all the quotients are distinct. Another example: $t_2 = (a^*)^2 = \sum_{k\in\mathbb{N}}(k+1)\,a^k$; for all integers n, we have

$$(a^n)^{-1}t_2 = \sum_{k\in\mathbb{N}}(n+k+1)\,a^k = t_2 + n\,a^*$$

cf. Exer. 2.2, p. 398

and all the quotients are again distinct. Despite their simplicity, these two examples exhibit the property that we seek: of course there are infinitely many quotients, but they can all be expressed as the linear combination of a finite number of suitably chosen series. We begin by interesting ourselves in this subsection in the *existence* of these series; in the next, we shall consider a procedure for *computing* them.

Definition 4.1 A subset U of $\mathbb{K}\langle\langle A^*\rangle\rangle$ is called *stable* if it is closed under quotient; that is, for all s in U and all f in A^*, $f^{-1}s$ is still in U. □

Theorem 4.1 *A series on A^* with coefficients in $\mathbb{K}$ is $\mathbb{K}$-recognisable if and only if it is contained in a stable submodule of $\mathbb{K}\langle\langle A^*\rangle\rangle$ of finite type.*

Proof. To allow us later to refer to parts of the proof, it is split into the following Properties 4.3 and 4.4, which show respectively that the condition is necessary and sufficient. ■

cf. Sec. 4.3, p. 451

The proof of these properties, and especially future developments, are greatly simplified by the following notations and conventions, which we call 'vectorial'.

Notation Let U be a (left) module on $\mathbb{K}$, most often $\mathbb{K}\langle\langle A^*\rangle\rangle$ or a submodule of $\mathbb{K}\langle\langle A^*\rangle\rangle$. If $T = \{t_1, t_2, \ldots, t_n\}$ is a set of n elements of U and if $x = (x_1, x_2, \ldots, x_n)$ is an element of $\mathbb{K}^n = \mathbb{K}^{1\times n}$, we write $x\cdot T$ for the element $x_1\,t_1 + x_2\,t_2 + \cdots + x_n\,t_n$ of U; that is, we view T as a column vector, an element of $U^{n\times 1}$. If X is a subset of $\mathbb{K}^n$, we write $X\cdot T$ for the set of linear combinations of the t_i with coefficients in X:

$$X\cdot T = \{x\cdot T \mid x\in X\}.$$

If m is a $r\times n$ matrix on $\mathbb{K}$, we write $m\cdot T$ for the set of r elements $u_i = m_i\cdot T$ where m_i is the ith row of m.

If U is indeed a submodule of $\mathbb{K}\langle\langle A^*\rangle\rangle$, and f is in A^*, we write $f^{-1}T$ for the set $\{f^{-1}t_1, f^{-1}t_2, \ldots, f^{-1}t_n\}$. These conventions give, with (4.2) and (4.3):

$$f^{-1}\,(x\cdot T) = x\cdot\left(f^{-1}T\right) \quad . \tag{4.5}$$

We also write $\langle T, f\rangle$ for the n-tuple $(\langle t_1, f\rangle, \langle t_2, f\rangle, \ldots, \langle t_n, f\rangle)$, an element of $\mathbb{K}^n = \mathbb{K}^{n\times 1}$; that is, we view $\langle T, f\rangle$ as a column vector. We have for example

$$\langle x\cdot T, f\rangle = x\cdot\langle T, f\rangle \quad \text{and, inversely, we will write} \quad T = \sum_{f\in A^*}\langle T, f\rangle f \quad . \tag{4.6}$$

Property 4.3 *Let s be a series recognised by a $\mathbb{K}$-representation (λ, μ, ν) of dimension Q and $T = \{t_p\}_{p\in Q}$ the set of series defined by*

$$\forall p\in Q \qquad t_p = \sum_{f\in A^*}(f\mu\cdot\nu)_p\, f \qquad \textit{or} \qquad T = \sum_{f\in A^*}(f\mu\cdot\nu)\, f \quad .$$

The left submodule U of $\mathbb{K}\langle\langle A^*\rangle\rangle$ generated by T is stable and contains s .

Proof. For all p in Q, we have *cf. also Exer. 4.2*

$$\forall f, g \in A^* \qquad \langle f^{-1}t_p, g\rangle = \langle t_p, f\,g\rangle = (f\mu \cdot g\mu \cdot \nu)_p$$
$$= \sum_{q\in Q} f\mu_{p,q}\,(g\mu \cdot \nu)_q = \sum_{q\in Q} f\mu_{p,q}\langle t_q, g\rangle$$

and hence

$$f^{-1}t_p = \sum_{q\in Q} [f\mu_{p,q}]\; t_q \quad . \tag{4.7}$$

which proves that the quotients of t_p are in U. An arbitrary element u of U is written $u = x \cdot T$ with x in $\mathbb{K}^{1\times n}$. By (4.5), all the quotients of u are in U; i.e., U is stable. Furthermore, for all f in A^*

$$\langle s, f\rangle = \lambda \cdot f\mu \cdot \nu = \sum_{p\in Q} \lambda_p\,(f\mu \cdot \nu)_p = \sum_{p\in Q} \lambda_p\langle t_p, f\rangle = \lambda \cdot \langle T, f\rangle \quad .$$

Thus, by (4.6), $s = \lambda \cdot T$ and U contains s. ■

Remark 4.2 With the vectorial notations, (4.7) can be written

$$\forall f \in A^* \qquad f^{-1}T = f\mu \cdot T \quad , \tag{4.8}$$

from which we deduce

$$\forall f \in A^* \qquad f^{-1}s = \lambda \cdot f\mu \cdot T \quad . \tag{4.9}$$

Property 4.4 *Every series s that belongs to a* stable *left submodule U of $\mathbb{K}\langle\langle A^*\rangle\rangle$, which is finitely generated by a family $T = \{t_p\}_{p\in Q}$, is recognisable. Furthermore, it is realised by a $\mathbb{K}$-representation $(\lambda^{(s)}, \mu, \nu)$ of dimension Q and whose morphism μ and vector ν are determined, independently of s, by T.*

Proof. If U contains s, there exists[49] a $\lambda^{(s)}$ in $\mathbb{K}^n$ such that $s = \lambda^{(s)} \cdot T$, whence, by (4.5), $f^{-1}s = \lambda^{(s)} \cdot f^{-1}T$ for all f in A^*. If U is stable, there exists, for each a in A, a matrix $a\mu$ in $\mathbb{K}^{Q\times Q}$ whose pth row is a sequence of coordinates of $a^{-1}t_p$ in T:

$$\forall p \in Q \qquad a^{-1}t_p = \sum_{q\in Q} a\mu_{p,q}\, t_p \quad , \qquad \text{that is,} \qquad a^{-1}T = a\mu \cdot T \quad . \tag{4.10}$$

Unless $\mathbb{K}$ is a field and T a basis of the vectorial subspace U, the matrix $a\mu$ defined by (4.10) is not unique.

We again write $\mu\colon A^* \to \mathbb{K}^{Q\times Q}$ for the morphism which extends the map defined on A by (4.10), and, by induction on the length of words, we compute

$$(f\,a)^{-1}T = a^{-1}(f^{-1}T) = a^{-1}(f\mu \cdot T) = f\mu \cdot (a^{-1}T) = f\mu \cdot (a\mu \cdot T) = (f\,a)\mu \cdot T \quad ,$$

and hence

$$\forall f \in A^* \qquad f^{-1}T = f\mu \cdot T \quad . \tag{4.11}$$

[49] This is not a λ 'to the power of s', but a λ 'which depends on s'.

Let ν be the (column) vector of dimension Q with coefficients in $\mathbb{K}$ defined by

$$\nu_q = \langle t_q, 1_{A^*} \rangle \ ; \quad \text{that is,} \qquad \nu = \langle T, 1_{A^*} \rangle \ .$$

It then follows, for all f in A^*, that

$$\begin{aligned} \langle s, f \rangle &= \langle f^{-1}s, 1_{A^*} \rangle = \langle \lambda^{(s)} \cdot f\mu \cdot T, 1_{A^*} \rangle \\ &= \lambda^{(s)} \cdot f\mu \cdot \langle T, 1_{A^*} \rangle = \lambda^{(s)} \cdot f\mu \cdot \nu \ . \end{aligned}$$

The series s is recognisable. ■

Property 4.3 makes explicit which are the 'suitably chosen series' that generate the stable module that contains s. Note also that an immediate consequence of Theorem 4.1 is that *every finite linear combination of* $\mathbb{K}$*-recognisable series is a* $\mathbb{K}$*-recognisable series.*

cf. Prop. 3.1, p. 425

4.1.3 The Kleene–Schützenberger Theorem revisited

The characterisation of $\mathbb{K}$-recognisable series by Theorem 4.1 also enables us to prove that $\mathbb{K}$-rational series on A^* are recognisable, more directly than by the Fundamental Theorem of finite $\mathbb{K}$-automata (Theorem 2.1) and Lemma 3.8. We start by describing the quotients of a product of two series and of the star of a series.

see p. 409 and p. 433

Lemma 4.1 *For all s and t in $\mathbb{K}\langle\langle A^* \rangle\rangle$ and all a in A, we have*

$$a^{-1}(s\,t) = \left(a^{-1}s\right) t + \langle s, 1_{A^*} \rangle \left(a^{-1}t\right) \ . \tag{4.12}$$

Proof. For all f in A^* we have

$$\begin{aligned} \langle a^{-1}s\,t, f \rangle = \langle s\,t, a\,f \rangle &= \sum_{g\,h = a\,f} \langle s, g \rangle \langle t, h \rangle \\ &= \langle s, 1_{A^*} \rangle \langle t, a\,f \rangle + \sum_{g'h' = f} \langle s, a\,g' \rangle \langle t, h' \rangle \\ &= \langle s, 1_{A^*} \rangle \langle a^{-1}t, f \rangle + \sum_{g'h' = f} \langle a^{-1}s, g' \rangle \langle t, h' \rangle \\ &= \langle s, 1_{A^*} \rangle \langle a^{-1}t, f \rangle + \langle \left(a^{-1}s\right) t, f \rangle \ . \end{aligned}$$

■

Lemma 4.2 *For all proper s in $\mathbb{K}\langle\langle A^* \rangle\rangle$ and all a in A, we have*

$$a^{-1}\left(s^*\right) = \left(a^{-1}s\right) s^* \ . \tag{4.13}$$

Proof. Starting from the equation $s^* = 1_{A^*} + s\,s^*$, we deduce (4.13) from (4.12) since $a^{-1}(1_{A^*}) = 0$ (trivial) and $\langle s, 1_{A^*} \rangle = 0$ by assumption. ■

Proof of the inclusion $\mathbb{K}\mathrm{Rat}\,A^* \subseteq \mathbb{K}\mathrm{Rec}\,A^*$ [Alternative] First observe that every polynomial p is recognisable: for all f whose length is greater than the degree of p, we have $f^{-1}p = 0$. Thus, the set of quotients of p, $\{f^{-1}p \mid f \in A^*\}$, is finite (and by definition it is stable): it generates in $\mathbb{K}$ a stable module which contains p (since $1_{A^*}^{-1}p = p$): p is recognisable.

Now let s and t be two $\mathbb{K}$-recognisable series and U and V be the two stable modules of finite type which respectively contain them. As we have already said, $s+t$ is recognisable since $U+V$ is a stable module of finite type which contains it.

The module $W = U\,t + V$ contains $s\,t$ and is of finite type since U and V are. It is stable by Lemma 4.1: $s\,t$ is recognisable.

First let us suppose that the series s is proper and consider the module $X = \mathbb{K}1_{A^*} + U\,(s^*)$; it is of finite type, and it contains $s^* = 1_{A^*} + s\,s^*$. It is stable, since for all u in U (and all a in A), the series $a^{-1}\,(k1_{A^*} + u\,s^*)$ can be written, by Lemma 4.2, as

$$\begin{aligned} a^{-1}\,(k1_{A^*} + u\,s^*) = a^{-1}\,(u\,s^*) &= \left(a^{-1}u\right)\,s^* + \langle u, 1_{A^*}\rangle\left(a^{-1}(s^*)\right) \\ &= \left(a^{-1}u\right)\,s^* + \langle u, 1_{A^*}\rangle\left(a^{-1}s\right)\,s^* \ , \end{aligned}$$

and is thus in X. Now let s be a series whose star is defined: $s = s_0 + s_{\mathsf{p}}$ where s_{p} is the proper part of s and s_0 its constant term; the star of s_0 is defined in $\mathbb{K}$. All the quotients of s_{p} are quotients of s and $U' = \mathbb{K}\,s_{\mathsf{p}} + U$ is a stable module of finite type which contains s_{p}. We end as before by considering the module $X' = \mathbb{K}\,s_0 + U'\,(s^*)$. We have thus proved that the $\mathbb{K}$-recognisable series contain the polynomials and are closed under the rational operations: they contain the $\mathbb{K}$-rational series. ■

cf. Prop. 2.6, p. 396

Underlying this last proof there is an effective procedure for computing a $\mathbb{K}$-representation of a series s from a rational expression that denotes s. We shall consider a more efficient method in the next subsection.

Exercises

4.1 Verify (4.2), (4.3) and (4.4).

• 4.2 Prove (4.8) directly (that is, without using (4.7)) using vectorial notation.

• 4.3 Give a proof of Theorem 3.2 using Theorem 4.1.

4.4 Do the same for Theorems 3.3 and 3.4.

4.2 Derivation of rational $\mathbb{K}$-expressions

The derivation of rational $\mathbb{K}$-expressions, which is to expressions what the quotient is to language and series, enables us to associate a $\mathbb{K}$-automaton with a $\mathbb{K}$-expression. This is therefore yet another proof of the inclusion $\mathbb{K}\mathrm{Rat}\,A^* \subseteq \mathbb{K}\mathrm{Rec}\,A^*$; it is the one which corresponds to the most direct algorithm. In Chapter I the derivation enabled us, by taking into account the *idempotence* of the union of languages (and also its associativity and commutativity), to construct *finite automata* from an expression. Addition in $\mathbb{K}$ in general is no longer idempotent. It will be nonetheless possible to construct a *finite* $\mathbb{K}$-*automaton* from a $\mathbb{K}$-expression by using a suitably modified derivation and by placing ourselves in a suitable universe. This universe is the set of polynomials of $\mathbb{K}$-expressions.

cf. Def. I.4.2, p. 445

cf. Th. I.4.3 and Prop. I.4.11, p. 143 and Th. I.5.1 and Prop. I.5.4, p. 152

4.2.1 Polynomials of $\mathbb{K}$-expressions

cf. Def. 2.3, p. 399 Recall that a rational $\mathbb{K}$-expression over A^* is a *well formed formula* obtained inductively from the letters of A, the constants $\{0, 1\}$, the operators $\{+, \cdot, *\}$ and the unary operators 'multiplication on the left by k' and 'multiplication on the right by k', both written k, for each k in $\mathbb{K}$. The set of rational $\mathbb{K}$-expressions over A^* is written $\mathbb{K}\,\mathsf{RatE}\,A^*$.

The *depth* of an expression E, $\mathsf{d}(\mathsf{E})$, the *constant term* of E, $\mathsf{c}(\mathsf{E})$, and the *series denoted* by a *valid* expression E, $|\mathsf{E}|$, were defined in Section 2.2. We also define the *literal length* of an expression E, written $\ell(\mathsf{E})$, as the number of occurrences of letters of A in E (*e.g.* $\ell((1/3\,(a^*)+1/6\,(b^*))^*) = 2$). We observe that $\ell((\mathsf{E}^*)) = \ell((k\,\mathsf{E})) = \ell(\mathsf{E})$ and that $\ell((\mathsf{E}+\mathsf{F})) = \ell((\mathsf{E}\cdot\mathsf{F})) = \ell(\mathsf{E}) + \ell(\mathsf{F})$.

cf. p. 399

As in the case of classical expressions (and of languages) two valid expressions are called *equivalent* if they denote the same series. The following identities are obviously satisfied by $\mathbb{K}$-expressions:

$$0\cdot\mathsf{E} \equiv \mathsf{E}\cdot 0 \equiv 0 \;,\quad 0+\mathsf{E} \equiv \mathsf{E}+0 \equiv \mathsf{E} \;,\quad 1\cdot\mathsf{E} \equiv \mathsf{E}\cdot 1 \equiv \mathsf{E} \tag{$\mathbf{T}$}$$

$$\left.\begin{array}{r} (k\,0) \equiv (0\,k) \equiv 0 \;,\quad (1\,k) \equiv (k\,1) \;, \\ (0_{\mathbb{K}}\,\mathsf{E}) \equiv (\mathsf{E}\,0_{\mathbb{K}}) \equiv 0 \;,\quad (1_{\mathbb{K}}\,\mathsf{E}) \equiv (\mathsf{E}\,1_{\mathbb{K}}) \equiv \mathsf{E} \;. \end{array}\right\} \tag{$\mathbf{T_K}$}$$

The identities ($\mathbf{T}$) are the 'trivial' identities already touched on for expressions over A^*; the identities ($\mathbf{T_K}$) are their natural analogue for weighted expressions.

We can see these identities as rewriting rules and it must be clear that each $\mathbb{K}$-expression is equivalent to a unique 'reduced' $\mathbb{K}$-expression which can be computed in time proportional to the length of the starting expression.

For the rest of this subsection, addition in $\mathbb{K}$, which is *associative* and *commutative*, is written $\oplus$ to distinguish it from the $+$ operator in expressions. The addition induced on $\mathbb{K}\langle\langle A^*\rangle\rangle$ is also written $\oplus$. The set of left linear combinations of $\mathbb{K}$-expressions with coefficients in $\mathbb{K}$, or polynomials of $\mathbb{K}\langle\mathbb{K}\,\mathsf{RatE}\,A^*\rangle$, is a left $\mathbb{K}$-module on $\mathbb{K}$: the addition $\oplus$ is commutative and the left exterior multiplication is distributive over $\oplus$, which gives the identities

$$\mathsf{k}\,\mathsf{E} \oplus k'\,\mathsf{E}' \equiv k'\,\mathsf{E}' \oplus k\,\mathsf{E} \qquad \text{and} \qquad k\,\mathsf{E} \oplus k'\,\mathsf{E} \equiv \big[k \oplus k'\big]\;\mathsf{E} \;. \tag{$\mathbf{B_K}$}$$

We next define a sort of multiplication on $\mathbb{K}\langle\mathbb{K}\,\mathsf{RatE}\,A^*\rangle$ so we can transform an expression of the form $(\mathsf{P}\cdot\mathsf{F})$ or $(\mathsf{P}\,k)$, where P is its polynomial, into a polynomial of expressions (which contains neither addition nor product on $\mathbb{K}$ as operators) by the following identities (these are no longer 'rational identities' but formulas which allow us to define this product):

$$[k\,\mathsf{E}]\cdot\mathsf{F} \equiv k\,\big(\mathsf{E}\cdot\mathsf{F}\big) \;,\qquad \big([k\,\mathsf{E}]\,k'\big) \equiv k\,\big(\mathsf{E}\,k'\big) \;, \tag{$\mathbf{C_K}$}$$

$$\big(\big[\mathsf{E}\oplus\mathsf{E}'\big]\;k\big) \equiv (\mathsf{E}\,k) \oplus (\mathsf{E}'\,k) \quad \text{and} \quad \big(\big[\mathsf{E}\oplus\mathsf{E}'\big]\cdot\mathsf{F}\big) \equiv (\mathsf{E}\cdot\mathsf{F}) \oplus (\mathsf{E}'\cdot\mathsf{F}) \;. \tag{$\mathbf{D_K}$}$$

We verify easily that these formulas are valid in the sense that they are consistent with interpretation: the series denoted by the left hand sides and right hand sides are

equal. We must also note the subtlety of the notations: $[k\,\mathsf{E}]$ and $k\,\mathsf{E}$ are monomials of $\mathbb{K}\langle\mathbb{K}\,\mathsf{RatE}\,A^*\rangle$, while $(k\,\mathsf{E})$ is a $\mathbb{K}$-expression in $\mathbb{K}\,\mathsf{RatE}\,A^*$. Note also that this multiplication on $\mathbb{K}\langle\mathbb{K}\,\mathsf{RatE}\,A^*\rangle$ is *not associative*.

4.2.2 $\mathbb{K}$-derivatives of a $\mathbb{K}$-expression

Definition 4.2 Let E be in $\mathbb{K}\,\mathsf{RatE}\,A^*$ and a in A. The $\mathbb{K}$-*derivative* of E with respect to a, written $\frac{\partial}{\partial a}\,\mathsf{E}$, is a *linear combination of* $\mathbb{K}$-*expressions* with coefficients in $\mathbb{K}$, defined recursively by the following formulas:

$$\frac{\partial}{\partial a}\,0 = \frac{\partial}{\partial a}\,1 = 0 \ ,$$

$$\forall a \in A \qquad \frac{\partial}{\partial a}\,a = 1 \ , \qquad \frac{\partial}{\partial a}\,b = 0 \quad \forall b \in A\,,\ b \neq a \ ,$$

$$\frac{\partial}{\partial a}(k\,\mathsf{E}) = k\,\frac{\partial}{\partial a}\,\mathsf{E} \ , \tag{4.14}$$

$$\frac{\partial}{\partial a}(\mathsf{E}\,k) = \left(\left[\frac{\partial}{\partial a}\,\mathsf{E}\right]\,k\right) \ , \tag{4.15}$$

$$\frac{\partial}{\partial a}(\mathsf{E}+\mathsf{F}) = \frac{\partial}{\partial a}\,\mathsf{E} \oplus \frac{\partial}{\partial a}\,\mathsf{F} \ , \tag{4.16}$$

$$\frac{\partial}{\partial a}(\mathsf{E}\cdot\mathsf{F}) = \left(\left[\frac{\partial}{\partial a}\,\mathsf{E}\right]\cdot\mathsf{F}\right) \oplus \mathsf{c}\,(\mathsf{E})\ \frac{\partial}{\partial a}\,\mathsf{F} \ , \tag{4.17}$$

$$\frac{\partial}{\partial a}(\mathsf{E}^*) = \mathsf{c}\,(\mathsf{E})^*\ \left(\left[\frac{\partial}{\partial a}\,\mathsf{E}\right]\cdot(\mathsf{E}^*)\right) \ . \tag{4.18}$$

The $\mathbb{K}$-derivatives of polynomials of $\mathbb{K}$-expressions are defined by linearity:

$$\frac{\partial}{\partial a}\left[\bigoplus_{i\in I} k_i\,\mathsf{E}_i\right] = \bigoplus_{i\in I} k_i\ \frac{\partial}{\partial a}\,\mathsf{E}_i \ . \tag{4.19}$$

□

Implicitly, the polynomials of expressions are reduced by the trivial identities; for example, if $\frac{\partial}{\partial a}\,\mathsf{E} = 1$, we will have

$$\frac{\partial}{\partial a}(\mathsf{E}\cdot\mathsf{F}) = \mathsf{F} \oplus \mathsf{c}\,(\mathsf{E})\ \frac{\partial}{\partial a}\,\mathsf{F} \quad \text{and not} \quad \frac{\partial}{\partial a}(\mathsf{E}\cdot\mathsf{F}) = (1\cdot\mathsf{F}) \oplus \mathsf{c}\,(\mathsf{E})\ \frac{\partial}{\partial a}\,\mathsf{F} \ .$$

In addition, the recursion implied by (4.15), (4.17) and (4.18) must be interpreted with the help of the identities $\mathbf{C_K}$ and $\mathbf{D_K}$.

Equations (4.14), (4.18) and (4.19) help to extend derivation to $\mathbb{K}$-expressions naturally. We insist on the fact that, as a consequence of the forms chosen for the $\mathbb{K}$-derivative of a sum (Equation (4.16)) and that of a product (Equation (4.16)), the $\mathbb{K}$-derivative of a $\mathbb{K}$-expression *is not* a $\mathbb{K}$-expression but a *polynomial of* $\mathbb{K}$-*expressions*. Here is the key[50] to the upcoming finiteness results.

[50]The adaptation to weighted expressions of V. Antimirov's construction on 'usual' expressions, given in Sec. I.5.2, p. 149.

The derivative of an expression with respect to a *word* f is defined by induction on the length of f:

$$\forall f \in A^+ \,,\ \forall a \in A \qquad \frac{\partial}{\partial f\, a}\,\mathsf{E} = \frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}\,\mathsf{E}\right) \quad . \tag{4.20}$$

Hence we deduce, by induction on the *length of* g, that derivation is indeed an *action* of A^* on $\mathbb{K}\langle\mathbb{K}\,\mathsf{RatE}\,A^*\rangle$:

Lemma 4.3 $\forall f, g \in A^+ \qquad \dfrac{\partial}{\partial f\, g}\,\mathsf{E} = \dfrac{\partial}{\partial g}\left(\dfrac{\partial}{\partial f}\,\mathsf{E}\right).$ ■

cf. Exam. 2.2, p. 400

Example 4.3 Returning to the valid $\mathbb{Q}$-expression $\mathsf{E}_1 = \left((\frac{1}{6}\,a^* + \frac{1}{3}\,b^*)^*\right)$, we set: $\mathsf{F}_1 = (\frac{1}{6}\,a^* + \frac{1}{3}\,b^*)$. It follows that:

$$\begin{aligned}
\frac{\partial}{\partial a}\,\mathsf{E}_1 &= \frac{\partial}{\partial a}\,({\mathsf{F}_1}^*) = 2\,\frac{\partial}{\partial a}\left(\tfrac{1}{6}\,a^*\right)\cdot {\mathsf{F}_1}^* \oplus 2\,\frac{\partial}{\partial a}\left(\tfrac{1}{3}\,b^*\right)\cdot {\mathsf{F}_1}^* = \tfrac{1}{3}\,(a^*\cdot {\mathsf{F}_1}^*) \quad ,\\
\frac{\partial}{\partial b}\,\mathsf{E}_1 &= 2\,\frac{\partial}{\partial b}\left(\tfrac{1}{3}\,b^*\right)\cdot {\mathsf{F}_1}^* = \tfrac{2}{3}\,(b^*\cdot {\mathsf{F}_1}^*) \quad , \qquad \text{(as we learned above)}\\
\frac{\partial}{\partial aa}\,\mathsf{E}_1 &= \tfrac{1}{3}\,\frac{\partial}{\partial a}\,(a^*\cdot {\mathsf{F}_1}^*) = \tfrac{1}{3}\left(\frac{\partial}{\partial a}\,a^*\right)\cdot {\mathsf{F}_1}^* \oplus \tfrac{1}{3}\,\mathsf{c}\,(a^*)\,\frac{\partial}{\partial a}\,({\mathsf{F}_1}^*)\\
&= \tfrac{1}{3}\,(a^*\cdot {\mathsf{F}_1}^*) \oplus \tfrac{1}{9}\,(a^*\cdot {\mathsf{F}_1}^*) = \tfrac{4}{9}\,(a^*\cdot {\mathsf{F}_1}^*) \quad ,\\
\frac{\partial}{\partial ab}\,\mathsf{E}_1 &= \tfrac{1}{3}\,\frac{\partial}{\partial b}\,(a^*\cdot {\mathsf{F}_1}^*) = \tfrac{1}{3}\left(\frac{\partial}{\partial b}\,a^*\right)\cdot {\mathsf{F}_1}^* \oplus \tfrac{1}{3}\,\mathsf{c}\,(a^*)\,\frac{\partial}{\partial b}\,({\mathsf{F}_1}^*) = \tfrac{2}{9}\,(b^*\cdot {\mathsf{F}_1}^*) \quad ,\\
\frac{\partial}{\partial ba}\,\mathsf{E}_1 &= \tfrac{2}{3}\,\frac{\partial}{\partial a}\,(b^*\cdot {\mathsf{F}_1}^*) = \tfrac{2}{3}\left(\frac{\partial}{\partial a}\,b^*\right)\cdot {\mathsf{F}_1}^* \oplus \tfrac{2}{3}\,\mathsf{c}\,(b^*)\,\frac{\partial}{\partial a}\,({\mathsf{F}_1}^*) = \tfrac{2}{9}\,(a^*\cdot {\mathsf{F}_1}^*) \quad ,\\
\frac{\partial}{\partial bb}\,\mathsf{E}_1 &= \tfrac{2}{3}\,\frac{\partial}{\partial b}\,(b^*\cdot {\mathsf{F}_1}^*) = \tfrac{2}{3}\left(\frac{\partial}{\partial b}\,b^*\right)\cdot {\mathsf{F}_1}^* \oplus \tfrac{2}{3}\,\mathsf{c}\,(b^*)\,\frac{\partial}{\partial b}\,({\mathsf{F}_1}^*)\\
&= \tfrac{2}{3}\,(b^*\cdot {\mathsf{F}_1}^*) \oplus \tfrac{4}{9}\,(b^*\cdot {\mathsf{F}_1}^*) = \tfrac{10}{9}\,(b^*\cdot {\mathsf{F}_1}^*) \quad .
\end{aligned}$$

□

The first property to prove is consistency between the $\mathbb{K}$-derivation of expressions and the quotient of series.

Theorem 4.2 $\forall \mathsf{E} \in \mathbb{K}\,\mathsf{RatE}\,A^*\,,\ \forall f \in A^+ \qquad \left|\dfrac{\partial}{\partial f}(\mathsf{E})\right| = f^{-1}\,|\,\mathsf{E}\,|\,.$

We begin by showing, still by induction on the *length of* f, the following properties:

Proposition 4.4 *For all* $\mathbb{K}$*-expressions* E *and* F *in* $\mathbb{K}\,\mathsf{RatE}\,A^*$*, for every word* f *in* A^+*, and for all* k *in* $\mathbb{K}$*, we have:*

(i) $\dfrac{\partial}{\partial f}(k\,\mathsf{E}) = k\,\dfrac{\partial}{\partial f}\,\mathsf{E}\,;$ (ii) $\dfrac{\partial}{\partial f}(\mathsf{E}\,k) = \left(\left[\dfrac{\partial}{\partial f}\,\mathsf{E}\right]\,k\right);$

(iii) $\dfrac{\partial}{\partial f}(\mathsf{E}+\mathsf{F}) = \dfrac{\partial}{\partial f}\,\mathsf{E} \oplus \dfrac{\partial}{\partial f}\,\mathsf{F}\,;$

(iv) $$\frac{\partial}{\partial f}(\mathsf{E}\cdot\mathsf{F}) = \left[\frac{\partial}{\partial f}\mathsf{E}\right]\cdot\mathsf{F} \oplus \left[\bigoplus_{\substack{f=g\,h\\ g\in A^*,\, h\in A^+}} \mathsf{c}\left(\frac{\partial}{\partial g}\mathsf{E}\right)\ \frac{\partial}{\partial h}\mathsf{F}\right] ;$$

v) $$\frac{\partial}{\partial f}(\mathsf{E}^*) = \bigoplus_{\substack{f=g_1 g_2\cdots g_n\\ g_1,g_2,\ldots,g_n\in A^+}} \mathsf{c}(\mathsf{E})^*\,\mathsf{c}\left(\frac{\partial}{\partial g_1}\mathsf{E}\right)\ \mathsf{c}(\mathsf{E})^*\,\mathsf{c}\left(\frac{\partial}{\partial g_2}\mathsf{E}\right)\cdots$$
$$\cdots\mathsf{c}(\mathsf{E})^*\,\mathsf{c}\left(\frac{\partial}{\partial g_{n\text{-}1}}\mathsf{E}\right)\ \mathsf{c}(\mathsf{E})^*\ \left[\frac{\partial}{\partial g_n}\mathsf{E}\right]\cdot(\mathsf{E}^*)\ . \qquad \blacksquare$$

Proposition 4.5 *For all $\mathbb{K}$-expressions E, we have:*

$$\forall f\in A^+ \qquad \langle |\mathsf{E}|, f\rangle = \mathsf{c}\left(\frac{\partial}{\partial f}\mathsf{E}\right)\ .$$

Proof. We omit the details of the proof, by induction on the *depth* of the expression E and based on the equalities of Proposition 4.4. The one delicate point is the case E^* when $|\mathsf{E}|$ is not a proper series; we write it in the form: *cf. Prop. 2.6, p. 396*

$$|\mathsf{E}|^* = (\mathsf{c}(\mathsf{E})\ 1_{A^*} \oplus |\mathsf{E}|_{\mathsf{p}})^* = (\mathsf{c}(\mathsf{E})^*\ |\mathsf{E}|_{\mathsf{p}})^*\,\mathsf{c}(\mathsf{E})^*\ . \qquad \blacksquare$$

Proof of Theorem 4.2. For all words f and g in A^+, it follows that:

$$\langle f^{-1}|\mathsf{E}|, g\rangle = \langle |\mathsf{E}|, f\,g\rangle = \mathsf{c}\left(\frac{\partial}{\partial f\,g}\mathsf{E}\right) = \mathsf{c}\left(\frac{\partial}{\partial g}\left[\frac{\partial}{\partial g}\mathsf{E}\right]\right) = \langle \left|\frac{\partial}{\partial f}\mathsf{E}\right|, g\rangle\ . \qquad \blacksquare$$

4.2.3 Derived terms

We are now able to state the principal result.

Theorem 4.3 *Let E be in $\mathbb{K}\,\mathsf{Rat}\mathsf{E}\,A^*$. There exists an integer n, $n \leqslant \ell(\mathsf{E})$, and n $\mathbb{K}$-expressions $\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_n$ such that for all letters a of A, there exist n coefficients, $k_1^{(a)}, k_2^{(a)}, \ldots, k_n^{(a)}$, and n^2 coefficients $\{z_{i,j}^{(a)}\}_{i,j\in[n]}$ in $\mathbb{K}$, such that*

(i) $$\frac{\partial}{\partial a}\mathsf{E} = \bigoplus_{i\in[n]} k_i^{(a)}\,\mathsf{K}_i\ ;$$ (ii) $$\forall i\in[n] \qquad \frac{\partial}{\partial a}\mathsf{K}_i = \bigoplus_{j\in[n]} z_{i,j}^{(a)}\,\mathsf{K}_j\ .$$

We shall call these expressions K_i, the derived terms *of E.*

This proposition and its proof below are the perfect generalisation of Theorem I.5.2 and its proof; to be more precise, the latter are a special case[51] of the former when $\mathbb{K} = \mathbb{B}$, with the simplifications due to the particularities of calculations in $\mathbb{B}$. This proposition also corresponds to the most natural procedure for computing the derived terms of E. We derive systematically with respect to the letters of A *cf. p. 150*

[51] We could logically have used just the former, but it would have been a pity to make the construction of the derived term automaton in the Boolean case depend on the most general definition of weighted automata.

first E, then the $\mathbb{K}$-expressions which we obtain in this manner, until the obtained set stops growing.

Examples 4.3 (continued) The above calculations show that the derived terms of E_1 are $(a^* \cdot \mathsf{F}_1{}^*)$ and $(b^* \cdot \mathsf{F}_1{}^*)$. □

Proof of Theorem 4.3. By induction on the depth of the expression E (not on its literal length). The proposition is obviously true for $\mathsf{0}$ and $\mathsf{1}$ and for $\mathsf{E} = a$, for each a in A.

(a) Equation (4.14) implies that if it is true for E, it is true for $(k\,\mathsf{E})$, $k \in \mathbb{K}$. The derived terms of $(k\,\mathsf{E})$ are the same as those of E. Equation (4.15), and $\mathbf{C_K}$ and $\mathbf{D_K}$, show that if K is a derived term of E, then $(\mathsf{K}\,k)$ is a derived term of $(\mathsf{E}\,k)$.

(b) If it is true for E and F, it is true for $(\mathsf{E}+\mathsf{F})$. Indeed, we have

$$\frac{\partial}{\partial a}(\mathsf{E}+\mathsf{F}) = \frac{\partial}{\partial a}\,\mathsf{E} \oplus \frac{\partial}{\partial a}\,\mathsf{F} = \bigoplus_{i\in[n]} k_i^{(a)}\,\mathsf{K}_i \oplus \bigoplus_{p\in[s]} l_p^{(a)}\,\mathsf{L}_p \ , \tag{4.21}$$

using the obvious notation. The set of derived terms of $(\mathsf{E}+\mathsf{F})$ is the union of those of E and F and this set satisfies the theorem.

(c) If it is true for E and F, it is true for $(\mathsf{E}\cdot\mathsf{F})$. In fact, we obtain

$$\frac{\partial}{\partial a}(\mathsf{E}\cdot\mathsf{F}) = \left(\left[\frac{\partial}{\partial a}\,\mathsf{E}\right]\cdot\mathsf{F}\right) \oplus \mathsf{c}\,(\mathsf{E})\ \frac{\partial}{\partial a}\,\mathsf{F} = \bigoplus_{i\in[n]} k_i^{(a)}\,(\mathsf{K}_i\cdot\mathsf{F}) \oplus \bigoplus_{p\in[s]} \left(\mathsf{c}\,(\mathsf{E})\ l_p^{(a)}\right)\ \mathsf{L}_p \ .$$

The set of derived terms of $(\mathsf{E}\cdot\mathsf{F})$ is the union of the set $\{(\mathsf{K}_i\cdot\mathsf{F})\}_{i\in[n]}$ and of $\{\mathsf{L}_p\}_{p\in[s]}$ and we verify that, for each i in $[n]$ and each a in A, we have

$$\frac{\partial}{\partial a}(\mathsf{K}_i\cdot\mathsf{F}) = \bigoplus_{j\in[n]} z_{i,j}^{(a)}\,(\mathsf{K}_j\cdot\mathsf{F}) \oplus \bigoplus_{p\in[s]} \left(\mathsf{c}\,(\mathsf{K}_i)\ l_p^{(a)}\right)\ \mathsf{L}_p \ . \tag{4.22}$$

(d) If it is true for E, it is true for (E^*). We have

$$\frac{\partial}{\partial a}(\mathsf{E}^*) = \mathsf{c}\,(\mathsf{E})^*\ \left(\left[\frac{\partial}{\partial a}\,\mathsf{E}\right]\cdot(\mathsf{E}^*)\right) = \bigoplus_{i\in[n]} \left(\mathsf{c}\,(\mathsf{E})^*\ k_i^{(a)}\right)\ (\mathsf{K}_i\cdot\mathsf{E}^*) \ .$$

The set of derived terms of (E^*) is $\{(\mathsf{K}_i\cdot\mathsf{E}^*)\}_{i\in[n]}$. We verify that for each i in $[n]$ and each a in A we have

$$\frac{\partial}{\partial a}(\mathsf{K}_i\cdot\mathsf{E}^*) = \bigoplus_{j\in[n]} z_{i,j}^{(a)}\,(\mathsf{K}_j\cdot\mathsf{E}^*) \oplus \bigoplus_{j\in[n]} \left(\mathsf{c}\,(\mathsf{K}_i)\ \mathsf{c}\,(\mathsf{E})^*\ k_j^{(a)}\right)\ (\mathsf{K}_j\cdot\mathsf{E}^*) \ . \tag{4.23}$$

■

Remark 4.3 To deduce from (4.21) that the set of derived terms of $(\mathsf{E}+\mathsf{F})$ is the *union* of those of E and F and not the concatenation of their sequences (we are performing *a priori* symbolic transformations), we use the identities $\mathbf{B_K}$. Likewise, the other mention of union in (c), and Equations (4.22) and (4.23), as well as the yet-to-come Equation (4.24), rely on $\mathbf{B_K}$, $\mathbf{C_K}$ and $\mathbf{D_K}$. □

Corollary 4.6 *Let* E *be in* $\mathbb{K}\,\mathsf{Rat}\mathsf{E}\,A^*$*. There exists an integer* m*,* $m \leqslant \ell(\mathsf{E})$ *, and* m $\mathbb{K}$*-expressions* K_1*,* K_2*, …,* K_m *such that for all words* f *in* A^+*, there exist* m *coefficients in* $\mathbb{K}$*,* $k_1^{(f)}$*,* $k_2^{(f)}$*, …,* $k_m^{(f)}$*, such that*

$$\frac{\partial}{\partial f}\,\mathsf{E} = \bigoplus_{i=1}^{i=m} k_i^{(f)}\,\mathsf{K}_i \quad .$$

Proof. By induction on the length of f. If f is a letter, the result reduces to (i) of Theorem 4.3. For each f in A^+ and each a in A, it follows that

$$\begin{aligned}\frac{\partial}{\partial f\,a}\,\mathsf{E} &= \frac{\partial}{\partial a}\left(\frac{\partial}{\partial f}\,\mathsf{E}\right) = \bigoplus_{i\in[n]} k_i^{(f)}\,\frac{\partial}{\partial a}\,\mathsf{K}_i \\ &= \bigoplus_{i\in[n]} k_i^{(f)}\left[\bigoplus_{j\in[n]} z_{i,j}^{(a)}\,\mathsf{K}_j\right] = \bigoplus_{j\in[n]}\left[\bigoplus_{i\in[n]}\left(k_i^{(f)}\,z_{i,j}^{(a)}\right)\right]\mathsf{K}_j \quad . \end{aligned} \tag{4.24}$$

■

4.2.4 The derived term automaton

We can see the statement of Theorem 4.3 itself as the definition of an automaton. Let $P'_{\mathsf{E}} = \{\mathsf{K}_1, \mathsf{K}_2, \ldots, \mathsf{K}_n\}$ be the set of derived terms of E, K_0 the expression E itself and P_{E} the union of P'_{E} and K_0 (that is, $P_{\mathsf{E}} = P'_{\mathsf{E}}$ if E is a derived term of itself). The *derived term automaton of* E is the $\mathbb{K}$-automaton $\mathcal{A}_{\mathsf{E}} = \langle\, P_{\mathsf{E}}, A, Z, I, T\,\rangle$ defined by

$$I_{\mathsf{K}_i} = \begin{cases} 1_{\mathbb{K}} & \text{if} \quad \mathsf{K}_i = \mathsf{K}_0 \ , \\ 0_{\mathbb{K}} & \text{otherwise} \ , \end{cases} \qquad Z_{\mathsf{K}_i,\mathsf{K}_j} = \bigoplus_{a\in A} z_{i,j}^{(a)}\,a \ , \qquad T_{\mathsf{K}_j} = \mathsf{c}\,(\mathsf{K}_j) \ ,$$

where the $z_{i,j}^{(a)}$ were defined in Theorem 4.3.

Theorem 4.4 *Let* E *be in* $\mathbb{K}\,\mathsf{Rat}\mathsf{E}\,A^*$*. The series recognised by the derived term automaton of* E *is equal to the series denoted by* E*:* *cf. Prop. I.4.11, p. 143*

$$|\mathcal{A}_{\mathsf{E}}| = |\,\mathsf{E}\,| \quad .$$

Proof. The definition of the $\mathbb{K}$-automaton $\mathcal{A}_{\mathsf{E}}$ is equivalent to that of the $\mathbb{K}$-representation (I, ζ, T) where I and T are the two vectors of dimension n defined above (with $n = \|P_{\mathsf{E}}\|$) and where $\zeta\colon A^* \to \mathbb{K}^{n\times n}$ is the morphism defined by:

$$\forall a \in A\,,\ \forall i,j \in [n] \qquad (a)\zeta_{i,j} = z_{i,j}^{(a)} \quad .$$

The series recognised by (I, ζ, T) (and hence by $\mathcal{A}_{\mathsf{E}}$) is, by definition

$$|\mathcal{A}_{\mathsf{E}}| = \bigoplus_{f\in A^*} \big(I\cdot(f)\zeta\cdot T\big)\,f \quad .$$

Equation (4.24) proves directly, by induction on the length of f, that

$$\forall f \in A^+\,,\ \forall i \in [n] \qquad \big(I\cdot(f)\zeta\big)_i = k_i^{(f)} \quad , \tag{4.25}$$

from which, by Proposition 4.5

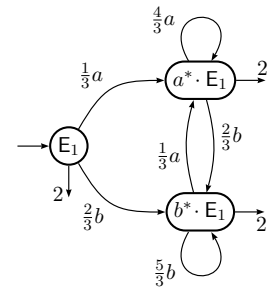
$$\forall f \in A^+ \qquad \langle |\mathcal{A}_{\mathsf{E}}|, f\rangle = \bigoplus_{i\in[n]} k_i^{(f)}\, \mathsf{c}\,(\mathsf{K}_i) = \langle |\,\mathsf{E}\,|, f\rangle \ .$$

The case of the empty word is treated by Property 2.5, which completes the proof. ■

Examples 4.3 (continued) The adjacent figure shows the derived term automaton of

$$\mathsf{E}_1 = \big((\tfrac{1}{6}\,a^*) + (\tfrac{1}{3}\,b^*)\big)^* \ .$$

which follows from the computations of p. 446. □

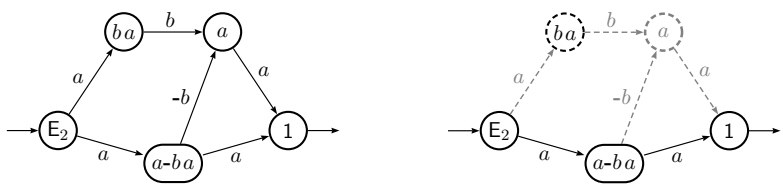


Remark 4.4 It can happen that for some i in $[n]$ all the coefficients $k_i^{(f)}$, defined in Corollary 4.6 and computed above by (4.25), are uniformly zero, for all f in A^*. We shall say in this case that the corresponding derived term K_i is a *phantom derived term*. Consider for example the expression $\mathsf{E}_2 = (a\,b\,a + (a\,(a - b\,a)))$. Then

$$\frac{\partial}{\partial a}\,\mathsf{E}_2 = b\,a \oplus (a - b\,a) \ , \qquad\qquad \frac{\partial}{\partial b}\,\mathsf{E}_2 = 0 \ ,$$
$$\frac{\partial}{\partial aa}\,\mathsf{E}_2 = \frac{\partial}{\partial a}\,b\,a \oplus \frac{\partial}{\partial a}(a - b\,a) = 1 \ ,$$
$$\frac{\partial}{\partial ab}\,\mathsf{E}_2 = \frac{\partial}{\partial b}\,b\,a \oplus \frac{\partial}{\partial b}(a - b\,a) = a \oplus ((-1_{\mathbb{K}})\,a) = 0 \ .$$

The derived terms of E_2 are $b\,a$, $(a - b\,a)$, a and 1 ; the preceding computations show that the coefficients of a are always zero: the expression $a = \frac{\partial}{\partial b}\,b\,a$ is a phantom derived term of E_2. □

see bibliographic notes



(a) before ... (b) ... and after removal of the phantom derived term

Figure 4.1: The derived term automaton of $\mathsf{E}_2 = (a\,b\,a + (a\,(a - b\,a)))$

If K_i is a phantom derived term, this state can be removed from the representation (I, ζ, T) without changing the recognised series. Now trimming the $\mathbb{K}$-automaton resulting from this deletion does not change the series recognised either, so we have just proved:

Proposition 4.7 *Let* E *be in* $\mathbb{K}\,\mathsf{Rat}\mathsf{E}\,A^*$ *and* $\mathcal{A}_{\mathsf{E}}$ *its derived term automaton. The* $\mathbb{K}$*-automaton obtained from* $\mathcal{A}_{\mathsf{E}}$ *by deleting the states that correspond to the phantom derived terms and trimming the result recognises the same series* $s_{\mathsf{E}} = |\,\mathsf{E}\,|$. ■

Example 4.4 Figure 4.1 shows the derived term automaton of E_2; in (b), we have first removed the state a, then the state ba, by trimming. □

Exercises

4.5 Compute the derived term $\mathbb{N}$-automaton of the expression $5\big((2\,a\,b)+((3\,b)\cdot(4\,(a\,b)^*))\big)^*$.

4.6 Let E be a rational $\mathbb{K}$-expression over the alphabet A. Show that the derived term automaton of E, i.e. $\mathcal{A}_{\mathsf{E}}$, is a quotient of the standard $\mathbb{K}$-automaton associated with E, i.e. $\mathcal{S}_{\mathsf{E}}$. *cf. Exer. II.3.8, p. 264 and Prop. 2.18, p. 415*

4.3 Series on a field

We now suppose that $\mathbb{K}$ is a *field*, not necessarily commutative, hence a *skew field*, or *division ring*. The preceding considerations about quotients of series will take on, we might say, a new dimension, since the ring of series $\mathbb{K}\langle\langle A^*\rangle\rangle$ is not only a $\mathbb{K}$-algebra, but a left and right $\mathbb{K}$-*vector space*, and we shall be able to use the notions of *dimension* of subspaces, of *duality*, and of *rank* of a matrix. *cf. Sec. 7, p. 488*

In particular, Theorem 4.1 (s is recognisable if and only if it belongs to a *stable* $\mathbb{K}$-module of finite type) becomes: *cf. p. 440*

Theorem 4.5 *Let $\mathbb{K}$ be a skew field. A series s of $\mathbb{K}\langle\langle A^*\rangle\rangle$ is recognisable if and only if the set of its (left) quotients* generates *a subspace of finite dimension of $\mathbb{K}\langle\langle A^*\rangle\rangle$, a left vector space over $\mathbb{K}$.*

Proof. The condition is sufficient since s is a quotient of itself by 1_{A^*}. It is necessary since, if the set of quotients *belongs* to a submodule of finite type, hence to a subspace of finite dimension E, then the set of quotients *generates* a subspace included in E, hence of finite dimension. ■

In fact the whole subsection rests on this transition from the *membership* of a submodule of finite type to the *generation* of a subspace of finite dimension, and hence to the *existence* of a finite number of quotients which generate the subspace. The consequence is that this subspace of finite dimension becomes an object intrinsically linked to the series: it plays for this series the role of the *minimal automaton* for a language by means of the notion of a *reduced representation* or representation of minimal dimension, and this subsection is in a sense devoted to the proof of the following result.

Theorem 4.6 *Let $\mathbb{K}$ be a skew field. Every $\mathbb{K}$-recognisable series s over A^* has representations of minimal dimension, effectively computable from an arbitrary representation which recognises s, and two representations of minimal dimension are similar.*

4.3.1 Rank of a series

The statement of Theorem 4.5 is clearly directed: *left* quotients, *left* vector space. The definition that follows will allow us to reestablish equilibrium.

Definition 4.3 The *Hankel matrix* of a series s of $\mathbb{K}\langle\langle A^*\rangle\rangle$ is the matrix $H^{(s)}$ of dimension $A^* \times A^*$ with entries in $\mathbb{K}$ defined by

$$\forall f, g \in A^* \qquad H^{(s)}_{f,g} = \langle s, f\,g\rangle \ . \qquad \square$$

cf. Exam. 2.5, p. 406

Example 4.5 Below is the upper left corner of the Hankel matrix of the series $c_1 = |\mathcal{C}_1|$. The rows and columns are in radix order; the first row and column repeat the indices.

	1_{A^*}	a	b	aa	ab	ba	bb	aaa	$\cdots$
1_{A^*}	0	0	1	0	1	2	3	0	$\cdots$
a	0	0	1	0	1	2	3	0	$\cdots$
b	1	2	3	4	5	6	7	8	$\cdots$
aa	0	0	1	0	1	2	3	0	$\cdots$
ab	1	2	3	4	5	6	7	8	$\cdots$
ba	2	4	5	8	9	10	11	16	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

$\square$

We observe that, by definition, the fth row of $H^{(s)}$ is (the sequence of coefficients of) the *left* quotient of s by f. Likewise, the gth column of $H^{(s)}$ is the *right* quotient of s by g. It is now timely to recall the following fundamental result:[52]

Theorem 4.7 *The row rank of a matrix over a skew field, if it is finite, is equal to its column rank.* ■

Two cases are possible: either the row rank of $H^{(s)}$ is finite and equal to its column rank, or it is countably infinite, and so is its column rank. We reformulate the preceding result by stating that the *dimension of the (left) subspace generated by the (left) quotients of a series is equal to the dimension of the (right) subspace generated by the (right) quotients of the same series*, which is not at all obvious if one does not think to introduce the matrix $H^{(s)}$. Theorems 4.5 and 4.7 can then be reformulated in the following manner.

Theorem 4.8 *Let s be a series of $\mathbb{K}\langle\langle A^*\rangle\rangle$ and $H^{(s)}$ its Hankel matrix. The rank of $H^{(s)}$ is called the* rank of *s; s is recognisable if and only if its rank is finite.* ■

It follows from Property 4.3 that the rank of a $\mathbb{K}$-recognisable series is less than or equal to the dimension of any of its $\mathbb{K}$-representations, and from Property 4.4 that this minimum can be reached.

Example 4.5 (continued) The rank of $c_1 = |\mathcal{C}_1|$ is 2: it is at most 2 since $\mathcal{C}_1$ is a 2-state automaton, and hence c_1 is recognised by a representation of dimension 2; it is at least 2 since rows a and b of the Hankel matrix of c_1 are independent.

Similarly, the rank of the series $c_2 = |\mathcal{C}_2| = \sum_{f \in A^*} \overline{f}^{\,2} f$ is 3. It is at most 3 since c_2 is recognised by $\mathcal{V}_2$ which has 3 states. It is at least 3 since we can extract from the Hankel matrix of c_2 the submatrix of rows and columns indexed by 1_{A^*}, b and $b\,b$, whose rows are independent. $\square$

0	1	9
1	9	49
9	49	225

[52] Its proof in a way caps the little non-commutative linear algebra primer given in Section 7.

4.3.2 Reduced representation

We will say that a $\mathbb{K}$-representation of a $\mathbb{K}$-recognisable series s is *reduced* if its dimension is minimal, equal to the rank of s. Reduced representations are characterised by the following result.

Proposition 4.8 *A $\mathbb{K}$-representation of dimension n is reduced if and only if $\lambda \cdot [A^*\mu]$ generates $\mathbb{K}^{1\times n}$, a left vector space,* and *$[A^*\mu] \cdot \nu$ generates $\mathbb{K}^{n\times 1}$, a right vector space.*

The proof of this proposition and, more generally, the reasoning which will follow about the dimension of the generated subspaces relies on the following lemma.[53]

Lemma 4.9 *Let V be a left $\mathbb{K}$-vector space, $G = \{v_1, v_2, \ldots, v_n\}$ a finite set of vectors of V and X a subspace of dimension m of $\mathbb{K}^n$, the left vector space over $\mathbb{K}$.*

(i) *If G is free, $X \cdot G$ is of dimension m.*

(ii) *If $X \cdot G$ is of dimension n, $m = n$ and G is free.*

Proof. If X is of dimension m ($m \leqslant n$), there exists a basis of X made up of m row vectors $\{l_1, l_2, \ldots, l_m\}$. Let L be the $m\times n$ matrix whose ith row is the vector l_i. We have $X = \mathbb{K}^m \cdot L$ and $X \cdot G = \mathbb{K}^m \cdot (L \cdot G)$ is a subspace of dimension at most m.

(i) Suppose G is free. If $L \cdot G$ were not free we would have an x in $\mathbb{K}^m$ such that $x \cdot (L \cdot G) = (x \cdot L) \cdot G = 0$, so that, since G is free, $x \cdot L = 0$: contradiction with the fact that the rows of L are independent.

(ii) If G is not free, there exists a basis B of the space generated by G which contains r vectors, with $r < n$. We can write $G = U \cdot B$, where U is the $n\times r$ matrix whose ith row is the sequence of coordinates of v_i in the basis B.

We have $X \cdot G = (X \cdot U) \cdot B$, $X \cdot U$ is a subspace of $\mathbb{K}^r$ and $X \cdot G$ is at most of dimension r. The conclusion follows by reusing (i). ■

Proof of Proposition 4.8. Let s be the series recognised by (λ, μ, ν) and, as in Property 4.3, the family of series $T = \{t_i\}_{i\in[n]}$: *cf. p. 440*

$$\forall i \in [n] \qquad t_i = \sum_{g\in A^*} (g\mu \cdot \nu)_i \, g \qquad \text{or} \qquad T = \sum_{g\in A^*} (g\mu \cdot \nu)\, g \ .$$

We saw that for all f in A^*, we have *cf. Equat. (4.8), p. 441*

$$f^{-1}T = f\mu \cdot T \ , \qquad \text{and hence} \qquad f^{-1}s = \lambda \cdot f\mu \cdot T \ .$$

The subspace generated by the left quotients of s is equal to $X \cdot T$ where X is the subspace generated by $\lambda \cdot [A^*\mu]$. The representation (λ, μ, ν) is thus reduced if and only if $n = \mathsf{dim}(X \cdot T)$.

[53] This lemma, associated with Theorem 4.7, gives, in the case of a field with non-commutative coefficients, an alternative to the use of the *determinant* in the standard case. There exist theories of determinants in the non-commutative case (see the notes section), but it is not necessary to cover them here.

By Lemma 4.9 we then deduce that in this case we necessarily have $\dim X = n$: that is, that $\lambda \cdot [A^*\mu]$ generates $\mathbb{K}^{1\times n}$ *and* that $\dim \langle T \rangle = n$.

It follows from Theorem 4.7 that the dimension of the left subspace of $\mathbb{K}\langle\langle A^*\rangle\rangle$ generated by T is equal to the dimension of the right subspace of $\mathbb{K}^n$ generated by the columns $g\mu \cdot \nu$ when g ranges over A^*: it is the rank of the $n \times A^*$ matrix whose entry (i, g) is $(g\mu \cdot \nu)_i = \langle t_i, g\rangle$. Thus $[A^*\mu] \cdot \nu$ generates a subspace of dimension n of $\mathbb{K}^{n\times 1}$, hence all of $\mathbb{K}^{n\times 1}$.

Conversely, it again follows from Lemma 4.9 that if X and T are of dimension n, then $X \cdot T$ and hence the subspace generated by all the quotients of s are both of dimension n. ■

The subspace generated by the (left) quotients of a (recognisable) series plays the role of the minimal automaton of a (recognisable) language; it is the choice of a basis in the subspace that fixes a reduced representation for a given recognisable series. Two reduced representations are therefore not necessarily equal; they are 'similar', in a very strong and mathematically defined sense. Two $\mathbb{K}$-representations (λ, μ, ν) and (λ', μ', ν') of the same dimension Q are *similar* if there exists an *invertible* $Q\times Q$ matrix P such that

$$\lambda = \lambda' \cdot P \ , \quad a\mu = P^{-1} \cdot a\mu' \cdot P \ , \quad \forall a \in A \qquad \text{and} \qquad \nu = P^{-1} \cdot \nu' \ .$$

Proposition 4.10 *Two reduced representations of the same recognisable series on a skew field are similar.*

Proof. Let (λ, μ, ν) and (λ', μ', ν') be two representations of s, both reduced, hence of the same dimension Q. Let $T = \{t_p\}_{p\in Q}$ and $T' = \{t'_p\}_{p\in Q}$ be the two sets of series defined as in Property 4.3:

$$\forall p \in Q \qquad t_p = \sum_{f\in A^*} (f\mu \cdot \nu)_p \, f \qquad \text{and} \qquad t'_p = \sum_{f\in A^*} \left(f\mu' \cdot \nu'\right)_p \, f \ .$$

Since (λ, μ, ν) and (λ', μ', ν') are reduced, T and T' are two *bases* of the *same* subspace of dimension Q of $\mathbb{K}\langle\langle A^*\rangle\rangle$, the subspace generated by the quotients of s. Let P be the matrix of coordinates of T' in T and P' that of T in T':

$$T' = P \cdot T \qquad \text{and} \qquad T = P' \cdot T' \ ,$$

so that $T = P \cdot P' \cdot T$ or $P \cdot P' = I$ and P is invertible.

We have $s = \lambda \cdot T = \lambda' \cdot T' = \lambda \cdot P \cdot T$ whence $\lambda = \lambda' \cdot P$ since T is free. We then have, for all a in A, $a^{-1}T = a\mu \cdot T$ and $a^{-1}T' = a\mu' \cdot T'$, or

$$a^{-1}(P \cdot T) = P \cdot (a^{-1}T) = P \cdot a\mu \cdot T = a\mu' \cdot T' = a\mu' \cdot P \cdot T \ ,$$

from which, as above, $P \cdot a\mu = a\mu' \cdot P$. We can finally write $\nu = \langle T, 1_{A^*}\rangle$ and $\nu' = \langle T', 1_{A^*}\rangle$, from which $\nu' = P \cdot \langle T, 1_{A^*}\rangle = P \cdot \nu$ and the proof is complete. ■

Example 4.5 (continued) The series $c_2 = |\mathcal{C}_2| = \sum_{f\in A^*} \overline{f}^2 f$ is recognised by $\mathcal{V}_2$ and by $\mathcal{V}'_2$ and hence by the two representations of dimension 3 (which we know are reduced): *cf. Exam. 2.8, p. 417*

$$\eta_2 = \begin{pmatrix}1 & 0 & 0\end{pmatrix} \ , \quad a\kappa_2 = \begin{pmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 4\end{pmatrix} \ , \quad b\kappa_2 = \begin{pmatrix}1 & 2 & 1\\ 0 & 2 & 2\\ 0 & 0 & 4\end{pmatrix} \ , \quad \xi_2 = \begin{pmatrix}0\\ 0\\ 1\end{pmatrix} \ ,$$

$$\eta'_2 = \begin{pmatrix}1 & 0 & 0\end{pmatrix} \ , \quad a\kappa'_2 = \begin{pmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 4\end{pmatrix} \ , \quad b\kappa'_2 = \begin{pmatrix}1 & 1 & 1\\ 0 & 2 & 4\\ 0 & 0 & 4\end{pmatrix} \ , \quad \xi'_2 = \begin{pmatrix}0\\ 0\\ 1\end{pmatrix} \ .$$

Let T_2 and T'_2 be the series vectors $\sum_{f\in A^*}(f\kappa_2 \cdot \xi_2)\, f$ and $\sum_{f\in A^*}(f\kappa'_2 \cdot \xi'_2)\, f$ respectively; if we retain from T_2 and T'_2 only the columns with index 1_{A^*}, b and bb, then T_2 and T'_2 are represented by the matrices

$$T_2 = \begin{pmatrix}0 & 1 & 9\\ 0 & 2 & 12\\ 1 & 4 & 16\end{pmatrix} \quad \text{and} \quad T'_2 = \begin{pmatrix}0 & 1 & 9\\ 0 & 4 & 24\\ 1 & 4 & 16\end{pmatrix} \ .$$

It is then immediate that the matrix of the coordinates of the rows of T'_2 in those of T_2 is

$$P_2 = \begin{pmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 1\end{pmatrix} \ ,$$

and we verify easily (the other equations are immediate) that

$$b\kappa_2 = \begin{pmatrix}1 & 2 & 1\\ 0 & 2 & 2\\ 0 & 0 & 4\end{pmatrix} = \begin{pmatrix}1 & 0 & 0\\ 0 & \frac{1}{2} & 0\\ 0 & 0 & 1\end{pmatrix} \cdot \begin{pmatrix}1 & 1 & 1\\ 0 & 2 & 4\\ 0 & 0 & 4\end{pmatrix} \cdot \begin{pmatrix}1 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 1\end{pmatrix} = P_2^{-1} \cdot b\kappa'_2 \cdot P_2 \ . \qquad \square$$

4.3.3 Linear recurrence

We shall now characterise the rational series by a procedure which generalises linear recurrences on the coefficients of classic rational series. We begin with a result which extends Proposition 4.8.

Proposition 4.11 *Let (λ, μ, ν) be a $\mathbb{K}$-representation of dimension Q of a series s of $\mathbb{K}\langle\langle A^*\rangle\rangle$, with the assumption that λ is not the null vector. Let $U \subseteq \mathbb{K}^{1\times Q}$ be the subspace generated by $\lambda \cdot [A^*\mu]$ and P a* maximal *prefix-closed subset such that*

$$\textit{the set of vectors} \qquad L = \{\lambda \cdot p\mu \mid p \in P\} = \lambda \cdot P\mu \qquad \textit{is free.} \qquad (*)$$

Then:

(i) *L generates U, hence $\|P\| = \dim U$;*

(ii) *there exists a $\mathbb{K}$-representation (λ', μ', ν') of s, of dimension P, such that $\lambda' \cdot [A^*\mu']$ generates $\mathbb{K}^{1\times P}$.*

Proof. The family of prefix-closed subsets which satisfy $(*)$ is not empty since, assuming λ is non-zero, it contains at least the subset reduced to the empty word. Every subset (prefix-closed or not) which satisfies $(*)$ contains at most $k \leqslant \|Q\|$ elements, where k is the dimension of U. Therefore there exist *maximal* prefix-closed subsets which satisfy $(*)$.

For each a in A, let τ_a be the morphism from $\mathbb{K}^{1\times Q}$ to itself whose matrix in the canonical basis is $a\mu$. The subspace U is the smallest subspace of $\mathbb{K}^{1\times Q}$ that is stable for all the τ_a and which contains λ.

Let p be in P (supposed to be maximal such that $(*)$ holds) and a in A; if $\lambda\cdot(p\,a)\mu$ is not in the subspace generated by L (and hence, in particular, $p\,a$ is not in P), then $L\cup\{\lambda\cdot(p\,a)\mu\}$ is free and P is not maximal. Hence $(\lambda\cdot p\mu)\cdot a\mu$ belongs to the subspace generated by L for all p and this subspace is stable for all the τ_a; it contains λ since P contains 1_{A^*}, and it is obviously contained in U: it is equal to U and P contains exactly k elements.

For each a in A, let $a\mu'$ be the matrix of coordinates, in the basis L of U, of the image of the row vectors of L under $a\mu$:

$$\forall a \in A \qquad L\cdot a\mu = a\mu'\cdot L \ . \tag{4.26}$$

We suppose that λ is the first vector of L and let $\lambda' = (1,0,0,\ldots,0)$: we thus have $\lambda = \lambda'\cdot L$, then deduce from (4.26) by induction on the length of f:

$$\forall f \in A^* \qquad \lambda\cdot f\mu = \lambda'\cdot f\mu'\cdot L \ .$$

From Lemma 4.9, we deduce that $\lambda'\cdot[A^*\mu'] = \mathbb{K}^{1\times P}$ and if we set $\nu' = L\cdot\nu$, it follows that

$$\forall f \in A^* \qquad \lambda\cdot f\mu\cdot\nu = \lambda'\cdot f\mu'\cdot\nu' \ . \qquad \blacksquare$$

cf. Exer. 3.12 **Example 4.6** The series $c_2 = |\mathcal{C}_2|$ is recognised by (λ_2,μ_2,ν_2) , the $\mathbb{N}$-representation defined by:

$$\lambda_2 = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix} \ , \quad a\mu_2 = \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 2 & 0 & 0\\ 0 & 0 & 2 & 0\\ 0 & 0 & 0 & 4\end{pmatrix} \ , \quad b\mu_2 = \begin{pmatrix}1 & 1 & 1 & 1\\ 0 & 2 & 0 & 2\\ 0 & 0 & 2 & 2\\ 0 & 0 & 0 & 4\end{pmatrix} \ , \quad \nu_2 = \begin{pmatrix}0\\0\\0\\1\end{pmatrix} \ .$$

We verify that the three vectors

$$\lambda_2\cdot(1_{A^*})\mu_2 = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix} \ , \quad \lambda_2\cdot b\mu_2 = \begin{pmatrix}1 & 1 & 1 & 1\end{pmatrix} \ , \quad \lambda_2\cdot(b\,a)\mu_2 = \begin{pmatrix}1 & 2 & 2 & 4\end{pmatrix} \ ,$$

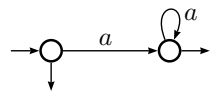
are independent and that they generate a stable subspace under (right) multiplication by $a\mu_2$ and $b\mu_2$:

$$\begin{aligned}
(\lambda_2\cdot(1_{A^*})\mu_2)\cdot a\mu_2 &= \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix} &&= \lambda_2\\
(\lambda_2\cdot b\mu_2)\cdot b\mu_2 &= \begin{pmatrix}1 & 3 & 3 & 9\end{pmatrix} &&= \lambda_2 - 3\,\lambda_2\cdot b\mu_2 + 3\,\lambda_2\cdot(b\,a)\mu_2\\
(\lambda_2\cdot(b\,a)\mu_2)\cdot a\mu_2 &= \begin{pmatrix}1 & 4 & 4 & 16\end{pmatrix} &&= 3\,\lambda_2 - 8\,\lambda_2\cdot b\mu_2 + 6\,\lambda_2\cdot(b\,a)\mu_2\\
(\lambda_2\cdot(b\,a)\mu_2)\cdot b\mu_2 &= \begin{pmatrix}1 & 5 & 5 & 25\end{pmatrix} &&= 6\,\lambda_2 - 15\,\lambda_2\cdot b\mu_2 + 10\,\lambda_2\cdot(b\,a)\mu_2 \ .
\end{aligned}$$

Hence the representation $(\lambda_2', \mu_2', \nu_2')$:

$$\lambda_2' = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \ , \quad a\mu_2' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 3 & -8 & 6 \end{pmatrix} \ , \quad b\mu_2' = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -3 & 3 \\ 6 & -15 & 10 \end{pmatrix} \ , \quad \nu_2' = \begin{pmatrix} 0 \\ 1 \\ 4 \end{pmatrix} \ . \quad \square$$

Example 4.7 The series $\underline{a^*}$ is recognised by the adjacent $\mathbb{N}$-automaton, and hence by the $\mathbb{N}$-representation:



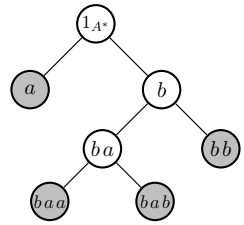
$$\lambda_1 = \begin{pmatrix} 1 & 0 \end{pmatrix} \ , \quad a\pi_1 = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \ , \quad \psi_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \ .$$

In this case, the two vectors $l_1 = \lambda_1 \cdot (1_{a^*})\pi_1 = (1 \ \ 0)$ and $l_2 = \lambda_1 \cdot a\pi_1 = (0 \ \ 1)$ are independent.[54] The prefix-closed subset is $P = \{1_{a^*}, a\}$ and we have $l_1 \cdot a\pi_1 = l_2 \cdot a\pi_1 = (0 \ \ 1)$ from which $\lambda_1' = \lambda_1$, $\pi_1' = \pi_1$ and $\psi_1' = \psi_1$. $\square$

Definition 4.4 If P is a non-empty prefix-closed subset of A^*, the *border of* P is the set:

$$C = P\,A \setminus P \ . \qquad \square$$

For example, the prefix-closed subset $\{1_{A^*}, b, b\,a\}$ and its border $\{a, b\,b, b\,a\,a, b\,a\,b\}$ are shown in the adjacent figure.



Lemma 4.12 *Let P be a non-empty prefix-closed subset and $C = P\,A \setminus P$ its border. Every word f of $A^* \setminus P$ can be written uniquely as $f = c\,g$ with c in C and g in A^*.*

Proof. Let $h = p\,a$ with p in P and a in A (this is the unique expression of h in this form). There are two, mutually exclusive, possible cases: h is in C or h is in P. Conversely, every word of $P \cup C$ can be written in this way, except 1_{A^*}. Hence we deduce the equality *between characteristic series*:

$$\underline{C} + \underline{P} = \underline{P}\,\underline{A} + \underline{1_{A^*}} \ ,$$

which we first rewrite as $\underline{1_{A^*}} - \underline{C} = \underline{P}(\underline{1_{A^*}} - \underline{A})$ then, by right multiplication by *cf. Rem. 2.2, p. 394*
$\underline{A^*} = (\underline{1_{A^*}} - \underline{A})^{-1}$, as $\underline{A^*} - \underline{C}\,\underline{A^*} = \underline{P}$, which proves the lemma. ■

Theorem 4.9 *A series s of $\mathbb{K}\langle\langle A^* \rangle\rangle$ is recognisable if and only if there exists a finite prefix-closed subset P such that, if we write its border $C = P\,A \setminus P$, there exists, for each pair (c, p) in $C \times P$, a coefficient $k_{c,p}$ in $\mathbb{K}$ such that*

$$\forall g \in A^* \ , \ \forall c \in C \qquad \langle s, c\,g \rangle = \sum_{p \in P} k_{c,p} \langle s, p\,g \rangle \ . \tag{4.27}$$

[54] This example, which is a bit stupid, will be useful later.

Proof. Let (λ, μ, ν) be a $\mathbb{K}$-representation that recognises s and let P be a prefix-closed subset and (λ', μ', ν') the equivalent $\mathbb{K}$-representation which we deduce from it as in Proposition 4.11. For each $c = pa$ in C, $\lambda \cdot c\mu$ belongs to the subspace (freely) generated by $\lambda \cdot P\mu$ and, for all q in P, we set $k_{c,q} = (a\mu')_{p,q}$. It follows that

$$\forall g \in A^* \qquad \langle s, cg\rangle = \lambda \cdot p\mu \cdot a\mu \cdot g\mu \cdot \nu = \sum_{q\in P} a\mu'_{p,q}\, \lambda \cdot q\mu \cdot g\mu \cdot \nu = \sum_{q\in P} k_{c,q}\langle s, qg\rangle\,.$$

Conversely, (4.27) implies that every quotient $f^{-1}s$ belongs to the subspace T generated by $p^{-1}s$ for p in P. This last property is trivially verified if f is in P and (4.27) can be rewritten as

$$\forall c \in C \qquad c^{-1}s = \sum_{p\in P} k_{c,p}\, p^{-1}s\,;$$

that is, the property is verified for f in C. *A contrario*, suppose that $f^{-1}s$ does not belong to T; we have $f = cg$ and choose f such that g is of minimal length. By (4.27), we have

$$\forall h \in A^*\,,\ \langle s, cgh\rangle = \sum_{p\in P} k_{c,p}\langle s, pgh\rangle : \qquad \text{that is,} \qquad f^{-1}s = \sum_{p\in P} k_{c,p}(pg)^{-1}s\ .$$

For each p in P, either pg is in P, or $pg = c'g'$ with c' in C; then $|c'| > |p|$ hence $|g'| < |g|$ and $(pg)^{-1}s$ is in T by the assumption of minimality of g. Hence, $f^{-1}s$ belongs to T, which is a contradiction. Also, s is recognisable by Theorem 4.5. ■

Remark 4.5 If $A = \{a\}$, every prefix-closed subset of A^* has the form

$$P = \{1_{a^*}, a, \ldots, a^{r-1}\}$$

for some integer r, and C is a singleton: $C = \{a^r\}$. Equation (4.27) becomes

$$\forall n \in \mathbb{N} \qquad \langle s, a^{n+r}\rangle = k_{r-1}\langle s, a^{n+r-1}\rangle + k_{r-2}\langle s, a^{n+r-2}\rangle + \cdots + k_0\langle s, a^n\rangle\ ;$$

that is, a linear recurrence in its standard form. □

It should be obvious that Proposition 4.11 (like Theorem 4.9) has a dual formulation:

Proposition 4.11 *Let (λ, μ, ν) be a $\mathbb{K}$-representation of dimension Q of a non-zero series s of $\mathbb{K}\langle\langle A^*\rangle\rangle$ and $V \subseteq \mathbb{K}^{Q\times 1}$ the* right *subspace generated by $[A^*\mu]\cdot\nu$. Let S be a* maximal *suffix-closed subset such that $K = S\mu\cdot\nu$ is free.*

(i) *K generates V, hence $\|S\| = \mathsf{dim}\, V$.*

(ii) *There exists a $\mathbb{K}$-representation (λ', μ', ν') of s, of dimension S, such that $[A^*\mu']\cdot\nu'$ generates $\mathbb{K}^{S\times 1}$.* ■

Another way to exploit Proposition 4.11, and hence Proposition 4.8, is by 'computing' the coefficients of a *reduced representation* of a recognisable series as a function of the coefficients of the series itself. Going from the series back to the representation

does not so much correspond to an effective procedure, like those described in Proposition 4.11 and Theorem 4.9, as express a fundamental property of recognisable series on a field, which will be used in the next chapter. see Prop. IV.3.2, p. 559

Proposition 4.13 *Let $\mathbb{K}$ be a skew field, s a $\mathbb{K}$-recognisable series of rank n, and (λ, μ, ν) a reduced representation of s. There exist two sets of n words: $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$ (which we can choose to be respectively prefix-closed and suffix-closed) and two $n \times n$ matrices α_P and β_Q such that*

$$\forall f \in A^* \qquad f\mu = \alpha_P \cdot (\langle s, p_i\, f\, q_j \rangle) \cdot \beta_Q \ , \tag{4.28}$$

where $(\langle s, p_i\, f\, q_j \rangle)$ denotes the $n \times n$ matrix whose entry (i, j) is $\langle s, p_i\, f\, q_j \rangle$.

Proof. As in Proposition 4.11, let there be a prefix-closed subset P such that $\lambda \cdot P\mu$ is a basis of $\mathbb{K}^{1\times n}$. There then exists an $n \times n$ matrix α_P such that $\alpha_P \cdot (\lambda \cdot P\mu)$ is the identity matrix (which describes the canonical basis of $\mathbb{K}^{1\times n}$). Symmetrically, let Q be a suffix-closed subset such that $Q\mu \cdot \nu$ is a basis of $\mathbb{K}^{n\times 1}$. There therefore exists an $n \times n$ matrix β_Q such that $(Q\mu \cdot \nu) \cdot \beta_Q$ is the identity matrix. It then follows that

$$f\mu = \alpha_P \cdot \lambda \cdot P\mu \cdot f\mu \cdot Q\mu \cdot \nu \cdot \beta_Q \ , \tag{4.29}$$

which is another formulation of (4.28). ■

4.3.4 Effective computations

The foregoing proofs all correspond to effective computations, assuming of course that the operations in $\mathbb{K}$ (addition, multiplication, inverse) are effective. All the complexities that follow are calculated assuming that each operation in $\mathbb{K}$ has a fixed constant cost, independent of its operands. Computations in $\mathbb{K}^n$ are based on the *Gaussian elimination* procedure.

Definition 4.5 A sequence of k vectors $(v^1, v^2, \ldots, v^k)$ of $\mathbb{K}^n$ is an *echelon system* if, for all i in $[k]$:

(i) $v^i{}_i = 1_{\mathbb{K}}$; (ii) $\forall j < i \qquad v^i{}_j = 0_{\mathbb{K}}$. □

An echelon system is free and hence $k \leqslant n$. The following proposition is classic (for commutative fields) and its proof is left as an exercise.

Proposition 4.14 [**Gaussian elimination**] *Let $\mathbb{K}$ be a skew field and let us view $\mathbb{K}^n$ as a left vector space over $\mathbb{K}$. Let $S = (v^1, v^2, \ldots, v^k)$ be an echelon system and let w be a vector in $\mathbb{K}^n$.*

(i) *We can decide whether w is in $\langle S \rangle$, the subspace generated by S, and, in this case, compute effectively the coordinates of w in S.*

(ii) *If w is not in $\langle S \rangle$, we can compute effectively w' such that $S' = S \cup \{w'\}$ is echelon and generates the same subspace as $S \cup \{w\}$.*

The complexity of these operations (deciding whether w is in $\langle S \rangle$ and computing the coordinates of either w or w') is $O(k\,n)$. ■

From this proposition we deduce the effective nature of the assertions, constructions, and specifications used in the proofs of this section. More precisely:

Corollary 4.15 *Let S be a finite set of vectors of $\mathbb{K}^n$ and let w be in $\mathbb{K}^n$.*

(i) *We can decide whether w belongs to $\langle S \rangle$.*

(ii) *We can extract effectively from S a basis T of $\langle S \rangle$.*

(iii) *We can compute effectively the coordinates in T of an (explicitly given) vector of $\langle S \rangle$.* ■

The following proposition and its proof exhibit the computation underlying Proposition 4.11.

Proposition 4.16 *Let (λ, μ, ν) be a $\mathbb{K}$-representation on A^*. We can compute effectively a maximal prefix-closed subset P of A^* such that the vectors $\lambda \cdot P\mu = \{\lambda \cdot p\mu \mid p \in P\}$ are independent. The complexity of this computation is $O(dn^3)$, where d is the cardinal of A.*

Proof. We set $P_0 = \{1_{A^*}\}$ and $C_0 = \emptyset$. The algorithm to compute P can be written in the following manner.

If $E_k = (P_k A \setminus P_k) \setminus C_k$ is non-empty, choose an arbitrary f in E_k and decide whether $\lambda \cdot f\mu$ belongs to $\langle \lambda \cdot P_k \mu \rangle$.

(i) If not, then $P_{k+1} = P_k \cup \{f\}$ and $C_{k+1} = C_k$.

(ii) If so, then $P_{k+1} = P_k$ and $C_{k+1} = C_k \cup \{f\}$.

Set $k = k + 1$ and start again.

The algorithm terminates when E_k is empty and at that moment $C_k = P_k A \setminus P_k$ is the border of P_k. The algorithm must terminate since P_k has at most n elements, so $P_k \cup C_k$ has at most $\|A\| n + 1$ elements and this set grows by 1 at each step of the algorithm.

By construction, P_k is prefix-closed, and each element f of C_k is such that $\lambda \cdot f\mu$ belongs to $\langle \lambda \cdot P_k \mu \rangle$: when C_k is the border of P_k, P_k is maximal. ■

Corollary 4.17 *Given a $\mathbb{K}$-representation (λ, μ, ν), we can compute effectively an equivalent reduced $\mathbb{K}$-representation (λ', μ', ν'), with complexity $O(dn^3)$.*

Proof. We start by applying the dual version of Proposition 4.11 and compute a suffix-closed subset S of A^* which determines a representation of s, $(\lambda'', \mu'', \nu'')$, such that $A^*\mu'' \cdot \nu''$ generates $\mathbb{K}^{S \times 1}$, a *right* vector space. Then, starting from this representation $(\lambda'', \mu'', \nu'')$, we construct by the forward direction of Proposition 4.11 a prefix-closed subset P which determines a representation (λ', μ', ν') such that $\lambda' \cdot A^*\mu'$ generates $\mathbb{K}^{1 \times P}$, a left vector space.

Theorem 4.7 (twice) and Lemma 4.9 then ensure that $A^*\mu' \cdot \nu'$ generates $\mathbb{K}^{P \times 1}$ and Proposition 4.8 that (λ', μ', ν') is reduced. ■

Example 4.7 shows that a reduction based only on the computation of a prefix-closed subset is insufficient for our needs. From there, it is not hard to construct an example for which the *two* successive computations are necessary.

Another consequence of the preceding results is the *decidability of equality* of $\mathbb{K}$-recognisable series (the decidability of equivalence of $\mathbb{K}$-automata or $\mathbb{K}$-representations). It is important to note that the propositions apply then not only when $\mathbb{K}$ is a skew field but also when $\mathbb{K}$ is only *embeddable* into a skew field $\mathbb{F}$ (on condition that we can compute effectively in $\mathbb{F}$).

Theorem 4.10 *Let A be a finite alphabet of cardinal d and $\mathbb{K}$ a sub-semiring of a skew field $\mathbb{F}$. The equality of two $\mathbb{K}$-recognisable series on A^*, s_1 and s_2, is decidable with complexity $O(d\,(n_1+n_2)^3)$, where n_1 and n_2 are the dimensions of two representations which define s_1 and s_2 respectively.*

Proof. The series s_1 and s_2 are equal if and only if $s_1 - s_2$ is the zero series. This last is recognised by (η, κ, ζ) with

$$\eta = \begin{pmatrix} \lambda_1 & -\lambda_2 \end{pmatrix}, \quad \kappa = \begin{pmatrix} \mu_1 & 0 \\ 0 & \mu_2 \end{pmatrix} \quad \textit{and} \quad \zeta = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \quad ,$$

where $(\lambda_1, \mu_1, \nu_1)$ and $(\lambda_2, \mu_2, \nu_2)$ are two representations which recognise s_1 and s_2 respectively. Corollary 4.17 and the fact that a series is zero if and only if its reduced representation is zero gives us the conclusion. ∎

Corollary 4.18 *If A is a finite alphabet and $\mathbb{K}$ a sub-semiring of a skew field $\mathbb{F}$, then the equivalence of finite $\mathbb{K}$-automata over A^* is decidable.* ∎

We can also deduce from the preceding results some statements that allow us to prove more simply the decidability of the equality of recognisable series, though the proofs do not correspond to procedures of optimal complexity.

Corollary 4.19 *Let $\mathbb{K}$ be a sub-semiring of a skew field $\mathbb{F}$. A series s on A^* recognised by a $\mathbb{K}$-representation (λ, μ, ν) of dimension n is uniformly zero if and only if it is zero on every word of length less than n; that is,*

$$\forall f \in A^* \quad \lambda \cdot f\mu \cdot \nu = 0 \quad \Longleftrightarrow \quad \forall f \in A^*\,,\ |f| < n \quad \lambda \cdot f\mu \cdot \nu = 0 \ .$$

Proof. Let P be the prefix-closed subset of A^*, C its border, and $\{k_{c,p}\}$ the family of elements of $\mathbb{F}$ whose existence is ensured by Theorem 4.9.

Proposition 4.11 ensures that P contains at most n words and, since it is a prefix-closed subset, these words are at most of length $n-1$. Hence, by assumption, $<s,p> = 0_{\mathbb{K}}$ for all p in P so, by (4.27), $<s,c> = 0_{\mathbb{K}}$ for all c in C. The same procedure of induction on the length of the proof (of the sufficient condition) of Theorem 4.9 proves that $<s,f> = 0_{\mathbb{K}}$ for all f in A^*. ∎

Hence we deduce, by using the same representation as in the proof of Theorem 4.10,

Theorem 4.11 *Let $\mathbb{K}$ be a sub-semiring of a skew field $\mathbb{F}$. Two $\mathbb{K}$-recognisable series on A^* are equal if and only if they coincide on all the words of length less than the sum of the dimensions of two representations which recognise them.* ∎

By virtue of the Kleene–Schützenberger Theorem, we deduce that *two* $\mathbb{K}$*-automata over* A^* *whose labels are linear combinations of letters in* A *are equivalent if and only if their behaviours coincide on all the words of length less than the sum of the number of states of these two* $\mathbb{K}$*-automata* (see Exercise 4.11 for the fact that the condition is necessary).

Exercises & Problems

• 4.7 Verify that, if $\mathbb{K}$ is not a field, it is false that the set of quotients of a recognisable series generates a sub-$\mathbb{K}$-module of $\mathbb{K}\langle\langle A^*\rangle\rangle$ of finite type.

cf. Exam. 4.5 and Exer. 3.12

• 4.8 Show that the rank of the series $|\mathcal{C}_n| = \sum_{f\in A^*} \overline{f}^n f$ is $n+1$.

4.9 Prove Proposition 4.14 and Corollary 4.15.

• 4.10 Give a direct proof of Corollary 4.19 (that is, a proof based on linear algebra but without using Theorem 4.9).

4.11 Verify with the example of the two automata below that the bound on the length given in the statement of Corollary 4.19 is optimal.

• 4.12 Show with a simple example that if $\mathbb{K}$ is not a sub-semiring of a skew field, we can construct a $\mathbb{K}$-recognisable series not uniformly zero, but zero on all the words of length less than or equal to n and recognised by a $\mathbb{K}$-representation of dimension d strictly less than n.

cf. Exer. I.3.10, p. 111

• 4.13 **Discriminating length II.** Recall that we call the *discriminating length* between two non-equivalent automata $\mathcal{A}$ and $\mathcal{B}$ the length of a shortest word which is accepted by one and not the other. We write $L_{\mathsf{d}}(n,m)$ (resp. $L_{\mathsf{nd}}(n,m)$) for the maximum of the discriminating lengths when $\mathcal{A}$ and $\mathcal{B}$ have respectively n and m states and are deterministic (resp. and are non-deterministic).

(a) Compute $L_{\mathsf{d}}(n,m)$; (b) Give an upper bound for $L_{\mathsf{nd}}(n,m)$.

4.14 **Syntactic algebras and ideals.**[55] We first suppose that $\mathbb{K}$ is a *commutative* ring. Thus, $\mathbb{K}\langle A^*\rangle$ is a $\mathbb{K}$-module freely generated by A^*.

(a) Verify that $\mathbb{K}\langle\langle A^*\rangle\rangle$ can be identified with the dual of $\mathbb{K}\langle A^*\rangle$ by associating with each series s the linear form $p \mapsto \langle s,p\rangle$. For each s of $\mathbb{K}\langle\langle A^*\rangle\rangle$, we write $\mathsf{Ker}\, s$ for the *kernel* of s viewed as a linear form:

$$\mathsf{Ker}\, s = \{p \in \mathbb{K}\langle A^*\rangle \mid \langle s,p\rangle = 0_{\mathbb{K}}\} \ .$$

(b) The *syntactic ideal* of s is the greatest bilateral ideal of $\mathbb{K}\langle A^*\rangle$ contained in $\mathsf{Ker}\, s$ and the *syntactic algebra* of s is the quotient of $\mathbb{K}\langle A^*\rangle$ by the syntactic ideal of s. Prove:

Theorem 4.12 *A series of* $\mathbb{K}\langle\langle A^*\rangle\rangle$ *is recognisable if and only if its syntactic algebra is of finite type.*

[55] Definitions, statements and proofs can be found in Berstel and Reutenauer, *op. cit.*, Chapter II.

(c) Now suppose that $\mathbb{K}$ is a (commutative) *field*. The *right syntactic ideal* of s is the greatest right ideal of $\mathbb{K}\langle A^*\rangle$ contained in $\mathsf{Ker}\, s$. Prove:

Theorem 4.13 *The rank of a series is equal to the co-dimension of its right syntactic ideal.*

4.4 Rational series and their supports

We have seen that the support of a $\mathbb{K}$-rational series on A^* is a rational subset of A^* if the semiring $\mathbb{K}$ is *positive*. This last assumption is necessary: *cf. Cor. 2.19, p. 415*

Example 4.8 The series $z_1 = s_1 - t_1 = \sum_{f\in A^*} (|f|_a - |f|_b)\, f$ is a $\mathbb{Z}$-rational series. The complement of $\mathsf{supp}\, z_1 = \{f \in A^* \mid |f|_a \neq |f|_b\}$ is the language $Z_1 = \{f \in A^* \mid |f|_a = |f|_b\}$, which we know is not rational. □ *cf. Exam. I.1.6, p. 64*

We shall now study certain conditions which ensure the rationality of support of a series, then present a remarkable result which links an $\mathbb{N}$-rational series to the characteristic series of its support. We end with several undecidable properties for $\mathbb{Z}$-rational series, somewhat surprising in this context where properties seem to be all decidable and effective.

4.4.1 Rationality of supports

Recall that a series s of $\mathbb{K}\langle\langle A^*\rangle\rangle$ is fundamentally a map from A^* to $\mathbb{K}$. It is therefore natural to write $\mathsf{Im}\, s$ for the set of coefficients of s, and, for every subset U of $\mathbb{K}$, $U\, s^{-1}$ for the set of words of A^* whose coefficient in s belongs to U:

$$U\, s^{-1} = \{f \in A^* \mid \langle s, f\rangle \in U\} \ .$$

Proposition 4.20 *Let $\mathbb{K}$ be a* finite *semiring and let s be a $\mathbb{K}$-rational series. For all subsets U of $\mathbb{K}$, $U\, s^{-1}$ is rational.*

Proof. Since s is also recognisable, s is recognised by a $\mathbb{K}$-representation (λ, μ, ν), of finite dimension Q. The language $U\, s^{-1}$ is recognised by the morphism $\mu\colon A^* \to \mathbb{K}^{Q\times Q}$, a morphism in a finite monoid, exactly as in the proof of Proposition II.2.4. ■ *cf. p. 238*

We then deduce: *cf. Prop. 2.8, p. 400*

Corollary 4.21 *Let s be a $\mathbb{Z}$-rational series on A^*. For all r and p in $\mathbb{Z}$, $p \neq 0$, the language $(r + p\mathbb{Z})\, s^{-1}$ is rational.* ■

Corollary 4.22 *Let s be an $\mathbb{N}$-rational series on A^*. For every integer k, the languages*

$$\{f \in A^* \mid \langle s, f\rangle \geqslant k\} \ , \quad \{f \in A^* \mid \langle s, f\rangle = k\} \quad \textit{and} \quad \{f \in A^* \mid \langle s, f\rangle \leqslant k\}$$

are rational. ■

Proposition 4.20 generalises in a remarkable way if $\mathbb{K}$ is a field.

Theorem 4.14 *Let $\mathbb{K}$ be a skew field. If s is a $\mathbb{K}$-rational series with a finite image, then $k\, s^{-1}$ is rational for all k in $\mathbb{K}$.*

cf. Th. 4.9, p. 457 *Proof.* There exists a finite prefix-closed subset P such that if we write its border $C = PA \setminus P$, then for each pair (c,p) in $C \times P$, there exists a coefficient $k_{c,p}$ in $\mathbb{K}$ such that

$$\forall f \in A^*\,,\ \forall c \in C \qquad \langle s, c\,f\rangle = \sum_{p\in P} k_{c,p}\langle s, p\,f\rangle \ . \tag{4.30}$$

Let us write $\overline{s}$ for the map from A^* to $\mathbb{K}^P$ defined by

$$\forall f \in A^* \qquad f\overline{s} = (\langle s, p_1 f\rangle, \langle s, p_2 f\rangle, \ldots, \langle s, p_n f\rangle)$$

and, for all a in A, write τ_a for *left* multiplication by a: $\tau_a(f) = a\,f$. The crux of the proof is the fact that the translation of τ_a by $\overline{s}$ is a map from the image $X \subseteq \mathbb{K}^P$ of $\overline{s}$ to itself, which we write ψ_a.

In fact, if $\langle s, p\,f\rangle = \langle s, p\,g\rangle$ for all p in P, it follows that if pa is in P, then $\langle s, pa\,f\rangle = \langle s, pa\,g\rangle$ by assumption, and if pa is not in P, it is in C and $\langle s, pa\,f\rangle = \langle s, pa\,g\rangle$ by (4.30). This proves that if $f\,\overline{s} = g\,\overline{s}$ then $(a\,f)\,\overline{s} = (a\,g)\,\overline{s}$ and that the function $\psi_a\colon X \to X$ is well defined by $\psi_a(f\overline{s}) = [\tau_a(f)]\,\overline{s}$.

It then follows that the *left* multiplication, which is a *left action* of A^* on A^*, is translated by $\overline{s}$ into a *left action* ψ of A^* on X, which is a *finite* set by assumption.

For each k in $\mathbb{K}$, $k\,s^{-1}$ is the set of words in A^* which send by ψ the vector $(1_{A^*})\overline{s} = (\langle s, p_1\rangle, \langle s, p_2\rangle, \ldots, \langle s, p_n\rangle)$ to a vector whose first coordinate is k. ∎

$$\begin{array}{ccc} A^* & \xrightarrow{\overline{s}} & X \subseteq \mathbb{K}^P \\ {\scriptstyle \tau_a}\downarrow & & \downarrow{\scriptstyle \psi_a} \\ A^* & \xrightarrow{\overline{s}} & X \subseteq \mathbb{K}^P \end{array}$$

Proposition 4.23 *Let $\mathbb{K}$ be a sub-semiring of $\mathbb{R}$. The set of supports of $\mathbb{K}$-rational series on A^* contains* $\mathrm{Rat}\,A^*$ *and is closed under union, product, star and intersection.*

cf. p. 415 *Proof.* The first assertion is a restatement of Corollary 2.20 (associated with the fact that $\mathbb{N}$ is contained in every sub-semiring of $\mathbb{R}$). Since $\mathbb{K}\mathrm{Rat}\,A^*$ is closed under the Hadamard product, we deduce first the closure by intersection, then, because s and $s \odot s$ have the same support, it follows that every support of a $\mathbb{K}$-rational series is the support of a $\mathbb{K}$-rational series with *non-negative coefficients*. Then, for such series, we clearly have

$$\mathsf{supp}(s + s') = \mathsf{supp}\,s \cup \mathsf{supp}\,s' \ , \quad \mathsf{supp}(s\,s') = \mathsf{supp}\,s\ \mathsf{supp}\,s'$$
$$\text{and} \qquad \mathsf{supp}(s^*) = (\mathsf{supp}\,s)^* \ .$$ ∎

cf. Exer. IV.3.4, p. 561 The properties of closure under morphism and inverse morphism (between free monoids) will be treated in the next chapter. The set $\mathrm{Rat}\,A^*$ is also closed under complement, but if $\mathbb{K}$ is not positive, the set of supports of $\mathbb{K}$-rational series can strictly contain $\mathrm{Rat}\,A^*$. The following result says that in this case the closure under complement is precisely characteristic of membership of $\mathrm{Rat}\,A^*$.

Theorem 4.15 [**Restivo and Reutenauer**] *Let $\mathbb{K}$ be a (sub-semiring of a) skew field. If a language and its complement are each the support of a $\mathbb{K}$-rational series over A^*, then this language is rational.*

Proof. Let (λ, μ, ν) and (λ', μ', ν') be representations of two series s and s' whose respective supports L and L' are complementary, and let n be the maximal dimension of the representations. Let f be a word of A^* of length greater than n and let

$$f = u\, v_1\, v_2 \cdots v_n\, w$$

be an arbitrary factorisation (where none of the v_i is empty).

Suppose that f is in $L = \mathsf{supp}\, s$. There exists a j, $j \leqslant n$, such that $\lambda \cdot (u\, v_1\, v_2 \cdots v_j)\mu$ belongs to the subspace generated by the vectors $\lambda \cdot (u\, v_1\, v_2 \cdots v_i)\mu$, $0 \leqslant i < j$. Hence there exist some k_i in $\mathbb{K}$ such that

$$\begin{aligned} \langle s, j \rangle &= \sum_{i=0}^{i=j-1} k_i\, \lambda \cdot (u\, v_1\, v_2 \cdots v_i)\mu \cdot (v_{j+1}\, v_{j+2} \cdots v_n\, w)\mu \cdot \nu \\ &= \sum_{i=0}^{i=j-1} k_i \langle s, u\, v_1\, v_2 \cdots v_i\, v_{j+1}\, v_{j+2} \cdots v_n\, w \rangle \ . \end{aligned}$$

None of the terms of the last sum can be zero and there exists (at least) one i, $0 \leqslant i < j$, such that $u\, v_1\, v_2 \cdots v_i\, v_{j+1}\, v_{j+2} \cdots v_n\, w$ belongs to $L = \mathsf{supp}\, s$.

In the same way, if f belongs to $L' = \mathsf{supp}\, s' = A^* \setminus L$, there exists j', $j' \leqslant n$, such that there exists i', $0 \leqslant i' < j'$, such that $u\, v_1\, v_2 \cdots v_{i'}\, v_{j'+1}\, v_{j'+2} \cdots v_n\, w$ belongs to $A^* \setminus L$.

We have thus just proved that L satisfies the condition that we termed $\mathfrak{E}_n$ and is thus rational by Theorem I.3.3 (the Ehrenfeucht–Parikh–Rozenberg Theorem). ■ *cf. Sec. I.3.4, p. 119*

We deduce for instance from this theorem (and from Example 4.8) that no $\mathbb{Z}$-rational series exists whose support is the language Z_1.

4.4.2 The Rational Skimming Theorem I

We now consider the case where $\mathbb{K} = \mathbb{N}$. We saw in the previous chapter that for every automaton $\mathcal{A}$ over A^* we can construct a particular covering, called the Schützenberger covering of $\mathcal{A}$, and from this 'extract' sub-automata which are both unambiguous and equivalent to $\mathcal{A}$.

The same construction applied to an $\mathbb{N}$-automaton $\mathcal{A}$ will give unambiguous $\mathbb{N}$-automata $\mathcal{T}$ whose behaviours have the same support as that of $\mathcal{A}$ (that is, $\mathbb{N}$-automata $\mathcal{T}$ such that $|\mathcal{T}| = \underline{\mathsf{supp}\, |\mathcal{A}|}$), which in itself is unsurprising. What is remarkable is that from the same construction we will obtain for each such $\mathcal{T}$, using the notion of $\mathbb{N}$-covering, an $\mathbb{N}$-automaton $\mathcal{D}$ which is the *complement* of $\mathcal{T}$ with respect to $\mathcal{A}$; that is, $|\mathcal{A}| = |\mathcal{T}| + |\mathcal{D}|$, which is stated by: *cf. the commentary that follows Cor. II.3.14.*

Theorem 4.16 [**Schützenberger**] *If s is an $\mathbb{N}$-rational series on A^*, the series*

$$s - \underline{\mathsf{supp}\, s}$$

is also $\mathbb{N}$-rational.

In other words, if we subtract 1 from each non-zero coefficient of an $\mathbb{N}$-rational series on A^* we again obtain an $\mathbb{N}$-rational series, a proposition which we can visualise as in Figure 4.2: the series s is shown as a sequence of values of coefficients, shown descending, and we skim the thinnest possible layer, thus of height 1, off the top;[56] what is left is the picture of another $\mathbb{N}$-rational series.

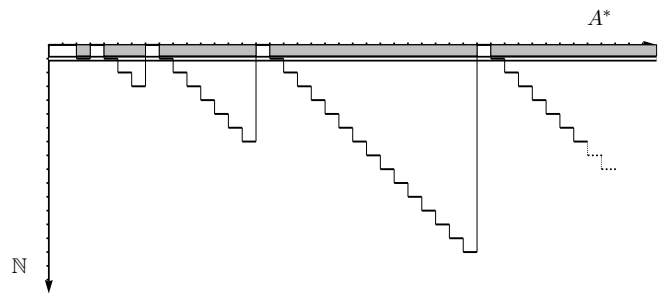


Figure 4.2: A milky view of $\mathbb{N}$-series

Proof of Theorem 4.16. Let $\mathcal{A} = \langle I, E, T \rangle$ be an $\mathbb{N}$-automaton on A^* (assumed to be proper) whose behaviour is equal to s, let $\mathcal{S} = \langle J, F, U \rangle$ be its S-covering, an $\mathbb{N}$-automaton of dimension R, and let $\mathcal{T} = \langle J, F_1, U_1 \rangle$ be an S-immersion in $\mathcal{A}$, also of dimension R. By definition, $\mathcal{T}$ is a sub-automaton of $\mathcal{S}$ and there exists a matrix F_2 with entries in $\mathbb{N}\langle\langle A^* \rangle\rangle$ and a vector U_2 with entries in $\mathbb{N}$ such that $F = F_1 + F_2$ and $U = U_1 + U_2$.

We then observe that the following automaton $\mathcal{S}'$ of dimension $R \times \{1, 2, 3\}$ is equivalent to $\mathcal{S}$, hence to $\mathcal{A}$:

$$\mathcal{S}' = \left\langle \begin{pmatrix} J & J & 0 \end{pmatrix}, \begin{pmatrix} F_1 & 0 & F_2 \\ 0 & F_1 & 0 \\ 0 & 0 & F \end{pmatrix}, \begin{pmatrix} U_1 \\ U_2 \\ U \end{pmatrix} \right\rangle = \left\langle \begin{pmatrix} J & J & 0 \end{pmatrix}, F', \begin{pmatrix} U_1 \\ U_2 \\ U \end{pmatrix} \right\rangle .$$

In fact, by summing in $\mathcal{S}'$ the *rows* of indices $(r, 1)$ and $(r, 2)$ for r in R, we obtain the matrices:

$$\begin{pmatrix} J & J & 0 \end{pmatrix} , \quad \begin{pmatrix} F_1 & F_1 & F_2 \\ 0 & 0 & F \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} U \\ U \end{pmatrix} ,$$

whose *columns* of indices $(r, 1)$ and $(r, 2)$, for r in R, are equal: $\mathcal{S}'$ is an $\mathbb{N}$-co-covering of:

$$\mathcal{S}'' = \left\langle \begin{pmatrix} J & 0 \end{pmatrix}, \begin{pmatrix} F_1 & F_2 \\ 0 & F \end{pmatrix}, \begin{pmatrix} U \\ U \end{pmatrix} \right\rangle .$$

The automaton $\mathcal{S}''$, of dimension $R \times \{1, 2\}$, is itself an $\mathbb{N}$-covering of $\mathcal{S}$ since if we sum the *columns* of indices $(r, 1)$ and $(r, 2)$, for r in R, we obtain the matrices:

$$J , \quad \begin{pmatrix} F \\ F \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} U \\ U \end{pmatrix} ,$$

[56] As the milk-maid skims the cream off the top of the milk.

whose *rows* of indices $(r,1)$ and $(r,2)$, for r in R, are equal to those of $\mathcal{S}$. It should be obvious that

$$|\mathcal{T}| = |\langle J, F_1, U_1 \rangle| = |\left\langle \begin{pmatrix} J & J & 0 \end{pmatrix}, F', \begin{pmatrix} U_1 \\ 0 \\ 0 \end{pmatrix} \right\rangle| = \underline{\mathsf{supp}\,|\mathcal{A}|}$$

and that

$$|\mathcal{A}| = |\mathcal{S}| = |\mathcal{S}'| = |\left\langle \begin{pmatrix} J & J & 0 \end{pmatrix}, F', \begin{pmatrix} U_1 \\ 0 \\ 0 \end{pmatrix} \right\rangle| + |\left\langle \begin{pmatrix} J & J & 0 \end{pmatrix}, F', \begin{pmatrix} 0 \\ U_2 \\ U \end{pmatrix} \right\rangle| \ .$$

The behaviour of the automaton

$$\mathcal{D} = \left\langle \begin{pmatrix} J & J & 0 \end{pmatrix}, F', \begin{pmatrix} 0 \\ U_2 \\ U \end{pmatrix} \right\rangle$$

is equal to $s - \underline{\mathsf{supp}\, s}$. ■

Example 2.5 (continued) We apply the preceding construction to the S-covering of the automaton

$$\mathcal{C}_1 = \left\langle \begin{pmatrix} 1 & 0 \end{pmatrix}, \begin{pmatrix} a+b & b \\ 0 & 2a+2b \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\rangle \ ,$$

simplified by the fact that in this case (and using the notation of the proof) we have $U_1 = U$ and hence, directly, $\mathcal{S}' = \mathcal{S}''$ is an $\mathbb{N}$-covering of the S-covering of $\mathcal{C}_1$. The corresponding automaton $\mathcal{D}_1$ is shown in Figure 4.3. □

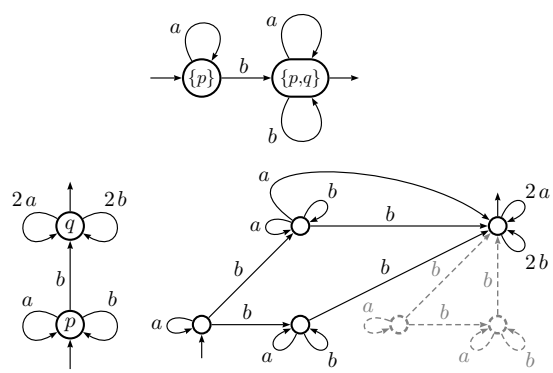


Figure 4.3: The automaton $\mathcal{D}_1$ whose behaviour is equal to $|\mathcal{C}_1| - \underline{\mathsf{supp}\,|\mathcal{C}_1|}$

An $\mathbb{N}$-series on A^* is said to be *bounded (by k)* if its image, i.e. the set of its coefficients, is a subset bounded (by k) of $\mathbb{N}$. We deduce from Theorem 4.16 a sequence of corollaries which specify the structure of $\mathbb{N}$-rational series.

Corollary 4.24 *An $\mathbb{N}$-rational series on A^* bounded by k is the sum of k unambiguous $\mathbb{N}$-rational series.* ■

This corollary may as well be seen as a consequence of Theorem 4.14.

Let s and t be two $\mathbb{N}$-series on A^*. We will write $s \leqslant t$ if $\langle s, f\rangle \leqslant \langle t, f\rangle$ for all f in A^* – that is, if there exists an $\mathbb{N}$-series u on A^* such that $s + u = t$ – when we will write $u = t - s$. More generally, we define $t \dot{-} s$ by

$$\langle t \dot{-} s, f\rangle = \mathsf{sup}\{0, (\langle t, f\rangle - \langle s, f\rangle)\} \ ,$$

for all f in A^*.

Corollary 4.25 *Let s and t be two $\mathbb{N}$-rational series on A^* with the property that s is bounded. Then $t \dot{-} s$ is an $\mathbb{N}$-rational series.* ■

Corollary 4.25 would be false without the assumption that s is bounded, as the following example shows.

Example 4.8 (continued) We return to the two $\mathbb{N}$-rational series s_1 and t_1 defined by $\langle s_1, f\rangle = |f|_a$ and $\langle t_1, f\rangle = |f|_b$ for all f in A^*. The series

$$u_1 = s_1 \odot s_1 + t_1 \odot t_1 \qquad \text{and} \qquad v_1 = 2(s_1 \odot t_1)$$

cf. Th. 3.1, p. 428 are $\mathbb{N}$-rational. We have, for all f in A^*,

$$\langle u_1, f\rangle = |f|_a^2 + |f|_b^2 \qquad \text{and} \qquad \langle v_1, f\rangle = 2\,|f|_a\,|f|_b$$

and hence $u_1 - v_1 = z_1 \odot z_1$ with $z_1 = s_1 - t_1$. It then becomes obvious first that $u_1 \geqslant v_1$, and hence $u_1 \dot{-} v_1 = u_1 - v_1$, and also that $u_1 - v_1$ is *not* an $\mathbb{N}$-rational series since $\mathsf{supp}(u_1 - v_1) = \mathsf{supp}\, z_1$ is not rational. □

4.4.3 Undecidable questions

We construct, with this simple model of finite weighted automata, some series for which we cannot answer some elementary questions, as soon as the semiring of coefficients contains $\mathbb{Z}$.

Theorem 4.17 *It is undecidable if the support of a $\mathbb{Z}$-rational series on A^* is all of A^*.*

cf. Exer. 0.6.10, p. 32 *Proof.* Let $B = \{x, y\}$; the morphism $\alpha\colon B^* \to \mathbb{N}^{2\times 2}$ defined by

$$x\alpha = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \ , \qquad \text{and} \qquad y\alpha = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

is *injective*.

Then let $\theta\colon A^* \to B^*$ and $\theta\colon A^* \to B^*$ be two morphisms. For i and j equal to 1 or 2, the series $s_{i,j}$ defined by

$$\forall f \in A^* \qquad \langle s_{i,j}, f\rangle = ((f\theta)\alpha)_{i,j} - ((f\mu)\alpha)_{i,j}$$

are $\mathbb{Z}$-rational, hence so are the series $t_{i,j} = s_{i,j} \odot s_{i,j}$, and the series

$$t = \sum_{i,j} t_{i,j} \quad .$$

The support of t is not all of A^* if and only if there exists a word f such that $\langle t, f \rangle = 0$; that is, since α is *injective*, if and only if $f\theta = f\mu$, which we know to be undecidable. ■ *cf. Th. 0.8.2, p. 37*

With the same construction, we easily obtain:

Corollary 4.26 *Let s be a $\mathbb{Z}$-rational series on A^*. It is undecidable whether:*

(i) *s has infinitely many coefficients equal to zero;*

(ii) *s has at least one positive coefficient;*

(iii) *s has infinitely many positive coefficients.* ■

Corollary 4.27 *It is undecidable whether the supports of two $\mathbb{Z}$-rational series on A^* are equal.* ■

Exercises

4.15 (a) Prove Corollaries 4.21 and 4.22 using Proposition 4.20.

(b) Prove Corollary 4.22 using Theorem 4.16.

4.16 Prove Corollaries 4.24 and 4.25.

• 4.17 Let $t \in \mathbb{N}\mathrm{Rat}\, A^*$ and $f \in \mathsf{supp}\, t$. Give an elementary proof (that is, without using Corollary 4.25, of which this result is only the simplest instance) that $u = t - \underline{f}$ belongs to $\mathbb{N}\mathrm{Rat}\, A^*$.

4.18 Prove the assertions of Corollary 4.26.

4.19 Prove Corollary 4.27.

4.20 Show that it is decidable whether the support of a $\mathbb{Z}$-rational series on A^* is finite.

4.21 Let s and t be in $\mathbb{N}\mathrm{Rat}\, A^*$. Show that it is undecidable whether $s \leqslant t$.

DEEPER & BROADER

5 Series on an arbitrary monoid

When we defined series in Section 1, we noted two difficulties. The first is to make $\mathbb{K}\langle\langle M\rangle\rangle$ a semiring (that is, to ensure that the Cauchy product is defined); the second is to equip $\mathbb{K}\langle\langle M\rangle\rangle$ with a topology suitable for defining and computing with the star operation. We chose not to make assumptions about $\mathbb{K}$, and we developed the entire theory supposing M to be graded, both because that was the most direct generalisation of the near-classic theory of series on the free monoid and because that is the case that will be useful in the next two chapters on relations and functions of words realised by finite automata.

Now we must consider defining $\mathbb{K}\langle\langle M\rangle\rangle$ without restrictive assumptions on M, in order to be able at least to integrate into the general framework of formal series the results we have proved for rational subsets of arbitrary monoids. It is also a necessary basis for the study[57] of algebraic or *context-free* languages and series, in the course of laying down which we shall consider rational series *on the free group* (and on other monoids of the same sort) that are not graded.

In the rest of this section, M is an arbitrary monoid.

5.1 Complete semirings, continuous semirings

A semiring $\mathbb{K}$ is called *complete* if *addition* in $\mathbb{K}$ can be extended to a *possibly infinite* number of operands (and not just for a finite number) and in such a way that the properties of associativity and distributivity are also satisfied.

In fact, we shall not extend a binary operation, because that is not really possible; we shall start with a more general operation which, restricted to two operands, will coincide with normal addition. Instead of binary addition we shall define on $\mathbb{K}$, as well as multiplication, a *sum* operation, written $\sum$, which acts on indexed sets of elements of $\mathbb{K}$ and which satisfies the following axioms:

- SAC0 $\mathbb{K}$ has a zero element, written $0_{\mathbb{K}}$, such that

$$\sum_{i\in I} k_i = 0_{\mathbb{K}} \quad \text{if} \quad \forall i \in I\,,\ k_i = 0_{\mathbb{K}} \qquad \text{and} \qquad \sum_{i\in\emptyset} k_i = 0_{\mathbb{K}}\ .$$

- SAC1 $\sum$ is associative: let $\{J_i \mid i \in I\}$ be a family of pairwise disjoint sets

[57] Which we shall not even begin here.

and $L = \bigcup_{i\in I} J_i$; then

$$\forall k_j \in \mathbb{K} \qquad \sum_{i\in I}\left(\sum_{j\in J_i} k_j\right) = \sum_{j\in L} k_j \ .$$

• SAC2 $\mathbb{K}$ is a monoid for multiplication, whose neutral element is written $1_{\mathbb{K}}$ (or 1).

• SAC3 Multiplication is distributive over $\sum$: for every set I and J of indices we have

$$\forall k_i, l_j \in \mathbb{K} \qquad \left(\sum_{i\in I} k_i\right)\cdot\left(\sum_{j\in J} l_j\right) = \sum_{(i,j)\in I\times J} k_i \cdot l_j \ .$$

• SAC4 $0_{\mathbb{K}}$ is a zero for multiplication.

We can then define, from $\sum$, 'normal' addition

$$k_1 + k_2 = \sum_{i\in\{1,2\}} k_i \ ,$$

which indeed makes $\mathbb{K}$ a semiring.

Examples & non-examples

The semirings $\mathbb{B}$, $\mathcal{N}$, $\mathbb{N}_{r,1}$[58] for all integers r, $\mathfrak{P}(M)$ for all monoids M, are complete, as are the square matrices (of a given size) over a complete semiring.

The semirings $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Z}/p\mathbb{Z}$ and $\mathbb{N}_{r,p}$ for all p greater than 1 are not complete; neither is $\mathrm{Rat}\, M$ for every infinite monoid M.

A sub-semiring of a complete semiring is not necessarily complete: $\mathbb{N}$ and $\mathrm{Rat}\, M$ are two examples. □

Proposition 5.1 *Let M be an arbitrary monoid and $\mathbb{K}$ a complete semiring. Then $\mathbb{K}\langle\langle M\rangle\rangle$ equipped with pointwise addition and the Cauchy product is a* complete *semiring.* ■

We suppose from now on that $\mathbb{K}$ is complete and now make a minimal assumption to have a notion of convergence in $\mathbb{K}$ which carries over to $\mathbb{K}\langle\langle M\rangle\rangle$.

Definition 5.1 A *continuous semiring* is a complete semiring in which the two following axioms hold:

• SAC5 The relation $\leqslant$ defined on $\mathbb{K}$ by

$$\forall k, l \in \mathbb{K} \qquad k \leqslant l \quad \Longleftrightarrow \quad \exists h \in \mathbb{K} \quad k + h = l$$

is an *order relation.*

• SAC6 For every family $\{k_i \mid i \in I\}$ of elements of $\mathbb{K}$, $\sum_{i\in I} k_i$ is the least upper bound of the $\sum_{i\in J} k_i$ for all finite J contained in I. □

[58] $\mathbb{N}_{r,p}$ is the quotient of $\mathbb{N}$ by the congruence generated by $r = r + p$, for all integers r and all positive integers p (*cf.* Exer. 0.4.2, p. 26).

Axiom SAC5 makes $\mathbb{K}$ an ordered semiring (the order is compatible with addition and multiplication):

Property 5.1 *If $\mathbb{K}$ is a continuous semiring then for all h, k and l in $\mathbb{K}$:*
(i) $k \leqslant l \implies k + h \leqslant l + h$;
(ii) $k \leqslant l \implies h\,k \leqslant h\,l$ *and* $k\,h \leqslant l\,h$. ■

What is more, SAC5 implies that $0_{\mathbb{K}}$ is the *smallest element* of $\mathbb{K}$ (which is not the case in all ordered semirings) from which we deduce:

Property 5.2 *A continuous semiring is positive.* ■

We easily verify:

Property 5.3
If $\mathbb{K}$ is a continuous semiring, then $\mathbb{K}\langle\langle M\rangle\rangle$ is a continuous semiring. ■

Property 5.4
If $\mathbb{K}$ is a continuous semiring, then $\mathbb{K}^{Q\times Q}$ is a continuous semiring. ■

5.2 Star of a series

In what follows we assume[59] that $\mathbb{K}$ is a *continuous semiring*. As in all complete semirings, the *star* of an element is defined by

$$k^* = \sum_{n\in\mathbb{N}} k^n$$

and as usual we write $k^+ = \sum_{n\in\mathbb{N}_*} k^n$. The axioms of associativity and distributivity of complete semirings ensure that the identities $\mathbf{U_l}$ and $\mathbf{U_r}$ hold; that is,

$$\forall k \in \mathbb{K} \qquad k^* = 1_{\mathbb{K}} + k\,k^* = 1_{\mathbb{K}} + k^*\,k \ . \tag{5.1}$$

cf. Lem. 2.4, p. 395 The fact that $\mathbb{K}$ is continuous leads to a variant of Arden's Lemma:

Lemma 5.2 [**Arden II**] *Let $\mathbb{K}$ be a continuous semiring and k and l be two arbitrary elements of $\mathbb{K}$. Then $k^*\,l$ (resp. $l\,k^*$) is* the smallest solution *of the equation*

$$X = l + k\,X \qquad\qquad (\textit{resp. } X = l + X\,k) \ . \tag{5.2}$$

Proof. That $k^*\,l$ is a solution is a consequence of (5.1). Conversely, if s is a solution, we have

$$s = l + k\,s \ ,$$

and hence, substituting this equality into itself n times,

$$s = l + k\,l + k^2\,l + \cdots + k^n\,l + k^{n+1}\,s \ .$$

Thus, for all n, $s \geqslant k^{\leqslant n}\,l$ whence since $\mathbb{K}$ is continuous $s \geqslant k^*\,l$. ■

[59] We shall nonetheless restate the assumption in our main propositions, for skimming readers.

Remark 5.1 The *same* proof shows that k^*l is the smallest solution of the *inequation*

$$X \geqslant l + k\,X \quad . \qquad \square$$

In fact, the solution of (5.2) is no longer unique, in contrast with the situation that prevailed under the assumptions of Lemma 2.4. We can nevertheless prove the standard identities. *cf. Prop. 2.5, p. 396*

Proposition 5.3 *Let $\mathbb{K}$ be a continuous semiring and k and l two arbitrary elements of $\mathbb{K}$; we have the identities*

$$(k+l)^* = k^*\,(l\,k^*)^* = (k^*\,l)^*\,k^* \quad , \tag{\mathbf{S}}$$

$$(k\,l)^* = 1 + k\,(l\,k)^*\,l \quad , \tag{\mathbf{P}}$$

$$\forall n \in \mathbb{N} \qquad k^* = k^{<n}\,(k^n)^* \quad . \tag{\mathbf{Z}_n}$$

Proof. We begin by verifying that $k^*\,(l\,k^*)^*$ is a solution of

$$X = (k+l)\,X + 1 \quad , \tag{5.3}$$

with the same sequence of calculations as for the proof of Proposition 2.5: *cf. p. 396*

$$\begin{aligned}(k+l)\,(k^*(l\,k^*)^*) + 1 &= k^+(l\,k^*)^* + \underbrace{l\,k^*(l\,k^*)^* + 1}_{\mathbf{U_1}} = k^+(l\,k^*)^* + (l\,k^*)^* \\ &= (k^+ + 1)(l\,k^*)^* = k^*\,(l\,k^*)^* \quad ,\end{aligned}$$

from which we deduce $k^*\,(l\,k^*)^* \geqslant (k+l)^*$.

Now let us turn again to the decomposition used in the proof of Proposition 2.6: *cf. p. 396*
for every pair of integers q and r we set

$$P_{q,r} = \sum_{\substack{p_0,p_1,\dots,p_q \in \mathbb{N} \\ p_0+p_1+\cdots+p_q = r}} k^{p_0}\,l\,k^{p_1}\,l \cdots k^{p_{q-1}}\,l\,k^{p_q} \tag{5.4}$$

with, by convention, $P_{0,r} = k^r$ and $P_{q,0} = l^q$. We verify by inspection that for every integer n

$$(k+l)^n = \sum_{q+r=n} P_{q,r}$$

and hence

$$(k+l)^{\leqslant n} = \sum_{m=0}^{n} \left(\sum_{q+r=m} P_{q,r} \right) \qquad \text{or} \qquad (k+l)^{\leqslant n} = \sum_{q=0}^{n} \left(\sum_{r=0}^{n-q} P_{q,r} \right) \quad .$$

Then for every integer q we have, by simple observation of (5.4)

$$\sum_{r=0}^{s} P_{q,r} \leqslant k^*\,(l\,k^*)^q \qquad \text{for every integer } s,$$

and hence

$$(k+l)^{\leqslant n} \leqslant \sum_{q=0}^{n} k^*\,(l\,k^*)^q = k^*\,(l\,k^*)^{\leqslant q}$$

$$(k+l)^{\leqslant n} \leqslant k^*\,(l\,k^*)^* \qquad \text{whence} \qquad (k+l)^* \leqslant k^*\,(l\,k^*)^* \ ,$$

which is the inequality we want. The two other identities are left as exercises. ■

5.3 $\mathbb{K}$-rational series

cf. p. 398 We proceed as in Section 2.2, except that the problems we had defining the star operation have disappeared. The *rational operations* on $\mathbb{K}\langle\langle M\rangle\rangle$ are:

(i) the *operations of a* $\mathbb{K}$*-semialgebra*; that is,
- the two *exterior multiplications* by the elements of $\mathbb{K}$;
- *addition*;
- *product*;

(ii) the *star* operation.

A subset of $\mathbb{K}\langle\langle M\rangle\rangle$ is *rationally closed* if it is closed under rational operations; that is, if it is a sub-semialgebra of $\mathbb{K}\langle\langle M\rangle\rangle$ which is closed under the star operation. The intersection of every family of rationally closed subsets is rationally closed and thus the *rational closure* of a set $\mathcal{E}$ is the smallest rationally closed subset which contains $\mathcal{E}$; it is written $\mathbb{K}\mathrm{Rat}\,\mathcal{E}$. As star is always defined we can also talk about rational closure even inside $\mathbb{K}$, by identifying $\mathbb{K}$ and $\mathbb{K}\langle\langle 1\!\!1\rangle\rangle$. We also reproduce Definition 2.2.

Definition 5.2 A series of $\mathbb{K}\langle\langle M\rangle\rangle$ is *rational* if it belongs to the rational closure of $\mathbb{K}\langle M\rangle$. The set ($\mathbb{K}$-semialgebra in fact) of rational series (over M with coefficients in $\mathbb{K}$) is written $\mathbb{K}\mathrm{Rat}\,M$. □

In order to establish a theory of rational series we still need to know how to compute the star of a matrix. The result we need is a copy of that which we proved in Section 2; its proof is slightly different, because the version of Arden's Lemma that we use for continuous semirings is slightly different.

Proposition 5.4 *Let $\mathbb{K}$ be a continuous semiring. The entries of the star of a square matrix m with entries in $\mathbb{K}$ belong to the rational closure of the entries of m.*

We know that this proposition is a consequence of the following proposition.

Proposition 5.5 *Let $\mathbb{K}$ be a continuous semiring and $m = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ a square matrix of dimension n over $\mathbb{K}$, decomposed into blocks A, B, C and D, where A and D are square blocks. We then have*

$$m^* = \begin{pmatrix} (A + B\,D^*\,C)^* & A^*\,B\,(D + C\,A^*\,B)^* \\ D^*\,C(A + B\,D^*\,C)^* & (D + C\,A^*\,B)^* \end{pmatrix} \ .$$

Proof [Conway]. Since $\mathbb{K}$ is continuous, so is $\mathbb{K}^{Q\times Q}$, and $m^* = \sum_{n\in\mathbb{N}} m^n$ exists and is, by Lemma 5.2, the smallest solution of the equation $X = 1 + m\,X$.

Set $m^* = \begin{pmatrix} E & F \\ G & H \end{pmatrix}$ where the blocks E, F, G and H are of the same size as A, B, C and D respectively. We can then write

$$m^* = \begin{pmatrix} E & F \\ G & H \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} \quad ,$$

a matrix equation which, as in the proof of Proposition 2.10, can be decomposed into a system of four equations:

$$E = 1 + A\,E + B\,G \tag{5.5}$$
$$F = A\,F + B\,H \tag{5.6}$$
$$G = C\,E + D\,G \tag{5.7}$$
$$H = 1 + C\,F + D\,H \tag{5.8}$$

Lemma 5.2, applied to (5.6) and (5.7), gives

$$F \geqslant A^*\,B\,H \quad \text{and} \quad G \geqslant D^*\,C\,E \quad .$$

Substituting into (5.5) and (5.8), we obtain

$$E \geqslant 1 + A\,E + B\,D^*\,C\,E = 1 + (A + B\,D^*\,C)\,E$$
$$\text{and} \qquad H \geqslant 1 + C\,A^*\,B\,H + D\,H = 1 + (C\,A^*\,B + D)\,H \quad ,$$

from which we obtain (*cf.* Remark 5.1)

$$E \geqslant (A + B\,D^*\,C)^* \quad \text{and} \quad H \geqslant (D + C\,A^*\,B)^* \quad ,$$

or, by substitution,

$$G \geqslant D^*\,C\,(A + B\,D^*\,C)^* \quad \text{and} \quad F \geqslant A^*\,B\,(D + C\,A^*\,B)^* \quad .$$

As the four right hand sides satisfy the system (5.5)–(5.8), they form a matrix n which satisfies both $m^* \geqslant n$ and $n = 1 + m\,n$. Hence $m^* = n$. ■

We can then repeat Sections 2.3 and 2.4 verbatim, to obtain:

Theorem 5.1 *A series of $\mathbb{K}\langle\langle M\rangle\rangle$ is rational if and only if it is the behaviour of a finite $\mathbb{K}$-automaton over M.* ■

Exercise

5.1 Prove the properties in this section that were left without proof.

6 Rational subsets in free products

Automata, again automata, always automata.

Danton

By its object, namely the characterisation of rational subsets in a certain family of monoids, this section falls under the scope of the last chapter where it would naturally have found its place: the results shown here extend those of Section II.6 and echo those of Section II.7. However, the techniques used are all based on the manipulations of automata in their matrix form which was introduced in Section 2.2 of this chapter, hence the need for the reader to wait until now.

We begin by defining the free product of two monoids; we do it in an elementary fashion, operationally, without bending to the imperatives of category theory. We then define *bipartite automata* and show, in the course of the three subsequent subsubsections, how far they can play the part of classic automata over a free monoid.

6.1 Free product of monoids

In what follows, M and N are two *disjoint* monoids. We write $M_\bullet$ for the complement of 1_M in M:

$$M_\bullet = M \setminus 1_M \quad .$$

We say that M has 'no divisors of the identity' if there does not exist p and q in $M_\bullet$ such that $pq = 1_M$; that is, if $M_\bullet$ is a *sub-semigroup* of M; that is, if 1_M is an adjoint identity to $M_\bullet$, or in other words if $\alpha\colon A^* \to M$ is a morphism then $(1_M)\alpha^{-1} = 1_{A^*}$.

The *free product* of M and N, written $M*N$, can be identified with the monoid whose elements other than the identity are finite sequences $(u_1, u_2, \ldots, u_k)$ of elements of $M_\bullet \cup N_\bullet$ taken alternately from $M_\bullet$ and $N_\bullet$; that is, $u_i \in M_\bullet \Leftrightarrow u_{i+1} \in N_\bullet$.

The product $(u_1, u_2, \ldots, u_k)(v_1, v_2, \ldots, v_l)$ of two such sequences is equal to:

(i) $(u_1, u_2, \ldots, u_k, v_1, v_2, \ldots, v_l)$ if u_k and v_1 are not in the same monoid;

(ii) $(u_1, u_2, \ldots, u_k v_1, v_2, \ldots, v_l)$ if u_k and v_1 are in the same monoid and if $u_k v_1$ is not its neutral element;

(iii) $(u_1, u_2, \ldots, u_{k-1})(v_2, ..., v_l)$ otherwise (that is, if $u_k v_1 = 1_M$ or 1_N).

The identity of $M*N$ is the *empty sequence*; that is, (1_M) and (1_N) are not elements of $M*N$. Each of M and N is (isomorphic to) a submonoid of $M*N$, identified with sequences of length 0 and 1. This allows us to write $u = u_1 u_2 \cdots u_k$ instead of $u = (u_1, u_2, \ldots, u_k)$; such a factorisation of an element u of $M*N$ is unique and is called its *canonical factorisation*. We also note that modulo this identification, M and N are not exactly disjoint but that their intersection is reduced to their identity. That is what we shall understand from now on when we say that two monoids are disjoint.

The *initial monoid* of u, written $\mathrm{MI}(u)$, is the monoid, M or N, to which u_1 belongs. The *monoid* of u, written $\mathrm{MF}(u)$, is the monoid to which u_k belongs. The product $u\,v$ of two elements u and v of $M{*}N$ is called *immiscible* if $\mathrm{MF}(u) \neq \mathrm{MI}(v)$; that is, if the canonical factorisation of $u\,v$ is:

$$u_1\,u_2 \cdots u_k\,v_1\,v_2 \cdots v_l \quad .$$

Given three monoids M, M' and M'', we verify that $(M * M') * M''$ is isomorphic to $M * (M' * M'')$ and we write simply $M * M' * M''$. In what follows we shall only study the free product of *two* monoids but all the results will hold for the free product of any *finite number* of monoids.

Examples 6.1 (i) If A and B are two disjoint alphabets then

$$(A \cup B)^* = (A^*) * (B^*) \quad .$$

In particular, if A has k elements then A^* is isomorphic to the free product of k copies of $\mathbb{N}$.

(ii) If A and B are two disjoint alphabets, then $F(A \cup B) = F(A) * F(B)$. In particular, if A has k elements then $F(A)$ is isomorphic to the free product of k copies of $\mathbb{Z}$.

(iii) Let $A = \{a, b, c\}$. The quotient of A^* by the congruence generated by the relation $a\,b = b\,a$ is isomorphic to the free product $\mathbb{N}^2 * \mathbb{N}$.

(iv) Let $M = \langle\, A; R \,\rangle$ and $N = \langle\, B; S \,\rangle$ be two monoids defined by a *presentation*. If A and B are disjoint, then $M * N = \langle\, A \cup B; R \cup S \,\rangle$. This example subsumes the three preceding ones. □

In what follows, we shall use the following notations: A and B are two alphabets assumed to be disjoint and M and N are two monoids, also disjoint. Two surjective morphisms $\varphi\colon A^* \to M$ and $\psi\colon B^* \to N$ determine a surjective morphism

$$\varphi{*}\psi\colon (A \cup B)^* \to M{*}N$$

by: $$\forall c \in A \cup B \qquad c\,[\varphi{*}\psi] = \begin{cases} c\varphi & \text{if} \quad c \in A \\ c\psi & \text{if} \quad c \in B \end{cases} .$$

We say the '*inverse image* of an element u of $M{*}N$' (resp. of M, of N), to mean the inverse image of u under $\varphi{*}\psi$ (resp. under φ, under ψ). With these notations, there follows:

Lemma 6.1 *If M and N have no divisors of the identity, the inverse image of an element u of $M{*}N$ is the product of the inverse images of the factors of its canonical factorisation; thus for example, if $\mathrm{MI}(u) = M$ and $\mathrm{MF}(u) = N$, we have*

$$u\,[\varphi{*}\psi]^{-1} = u_1\varphi^{-1}\,u_2\psi^{-1}\,u_3\varphi^{-1} \cdots u_{n-1}\varphi^{-1}\,u_n\psi^{-1} \quad . \qquad \blacksquare$$

Exercises

- 6.1 Prove Lemma 6.1.

6.2 Show that for every monoid S and every pair of morphisms $\alpha\colon M \to S$ and $\beta\colon N \to S$ there exists a unique morphism $\theta\colon M{*}N \to S$ which makes the adjacent diagram commute.

$$\begin{tikzcd} M \arrow[d] \arrow[dr, "\alpha"] & \\ M{*}N \arrow[r, "\theta"] & S \\ N \arrow[u] \arrow[ur, "\beta"'] & \end{tikzcd}$$

Show that conversely every monoid which has this property with respect to M and N is isomorphic to $M{*}N$.

• 6.3 Show that $(\mathbb{Z}/2\mathbb{Z}) * (\mathbb{Z}/2\mathbb{Z})$ contains a subgroup of index 2 isomorphic to $\mathbb{Z}$.

6.4 Observe that a rational subset of a free product is not necessary recognisable.

6.2 Bipartite automaton over a free product

We shall define a particular class of automata over the free products of monoids which are in many ways similar to automata over the free monoids. Recall first that a *graph* is called *bipartite* if there exists a partition of the set of its vertices into two classes such that no two vertices in the same class are adjacent.

We say that an *automaton* $\mathcal{A}$ over $M{*}N$ is *bipartite* if the following conditions are met:

(i) $\mathcal{A}$ is a bipartite graph; let $S = P + Q$ be the partition of the set of its vertices;
(ii) the transitions from P to Q are labelled by subsets of M;
(iii) the transitions from Q to P are labelled by subsets of N.

A bipartite automaton $\mathcal{A} = \langle I, E, T \rangle$ can then be written

$$\mathcal{A} = \left\langle \begin{pmatrix} I' & I'' \end{pmatrix}, \begin{pmatrix} 0 & V \\ W & 0 \end{pmatrix}, \begin{pmatrix} T' \\ T'' \end{pmatrix} \right\rangle \tag{6.1}$$

where V and W are two matrices of dimension $P{\times}Q$ and $Q{\times}P$ with values in $\mathfrak{P}(M)$ and $\mathfrak{P}(N)$ respectively; these notations will be used implicitly in the rest of the section.

The automaton $\mathcal{A}$ is *proper* if the neutral element (of M or N) belongs to no label; that is,

$$\forall p \in P\,,\ \forall q \in Q \qquad V_{p,q} \subseteq M_\bullet \qquad \text{and} \qquad W_{q,p} \subseteq N_\bullet\ .$$

A *proper* bipartite automaton $\mathcal{A}$ can (and should?) be seen as an automaton over a free monoid, in general not finitely generated: each element of $M_\bullet$ and $N_\bullet$ is considered as a letter. If $\mathcal{A}$ is bipartite and proper, two consecutive transitions in any computation are labelled by elements taken from two different monoids: their product is immiscible and the canonical factorisation of every element of $|\mathcal{A}|$ is 'read' in the computation which recognises it, exactly as a word recognised by a classic automaton over an alphabet A is 'spelt out' by the transitions of the computation which recognises it.

This identification forces us to take the whole set $M_\bullet \cup N_\bullet$ as the alphabet, which is infinite even if M and N are themselves finitely generated. It allows us to extend most of the usual definitions and constructions on automata over a free monoid to bipartite automata.

We say for example that a bipartite automaton is *deterministic* if:

(i) the transitions leaving from each state have *disjoint* labels; that is,

$$\begin{array}{ll} \forall p \in P\,,\ \forall q', q'' \in Q & V_{p,q'} \cap V_{p,q''} = \emptyset \\ \forall q \in Q\,,\ \forall p', p'' \in P & W_{q,p'} \cap W_{q,p''} = \emptyset \ , \end{array}$$

(ii) there is at most one initial state in each subset[60] P and Q:

$$\|I'\| \leqslant 1 \qquad \text{and} \qquad \|I''\| \leqslant 1 \ .$$

A proper bipartite automaton is *complete*[61] if:

(i) the union of transition labels leaving from each state is $M_\bullet$ or $N_\bullet$; that is,

$$\forall p \in P \qquad \bigcup_{q \in Q} V_{p,q} = M_\bullet \qquad \text{and} \qquad \forall q \in Q \qquad \bigcup_{p \in P} W_{q,p} = N_\bullet \ ,$$

(ii) neither I' nor I'' is empty.

We say that an automaton over a monoid is *rational* if all its coefficients are rational subsets.

Proposition 6.2 *Every rational of $M * N$ is the behaviour of a rational bipartite automaton.*

Proof. Every rational of $M * N$ is the behaviour of a proper finite automaton $\mathcal{A} = \langle I, E, T \rangle$; that is, one whose coefficients are finite subsets of $M_\bullet \cup N_\bullet$ (and where I and T are boolean vectors), since $M_\bullet \cup N_\bullet$ generates $M * N$. We write $E = F + G$ and expand E^* by the classical identities

cf. also Exer. I.4.3, p. 133

$$\begin{aligned} (F+G)^* &= (F^* G)^* F^* = (G + F^+ G)^* F^* = G^* (F^+ G^+)^* F^* \\ &= (1 + G^+)(F^+ G^+)^* (1 + F^+); \end{aligned}$$

that is, by expanding and rearranging,

$$\begin{aligned} (F+G)^* &= (F^+ G^+)^* + G^+ (F^+ G^+)^* + (F^+ G^+)^* F^+ + G^+ (F^+ G^+)^* F^+ \\ &= (F^+ G^+)^* + G^+ (F^+ G^+)^* + F^+ (G^+ F^+)^* + (G^+ F^+)^+ \ . \end{aligned}$$

which we can write, since $1 + 1 = 1$,

$$(F+G)^* = (F^+ G^+)^* + G^+ (F^+ G^+)^* + F^+ (G^+ F^+)^* + (G^+ F^+)^* \ .$$

We then see that $|\mathcal{A}| = I \cdot E^* \cdot T$ is the behaviour of a bipartite automaton

cf. also Exer. 2.1, p. 397

$$\mathcal{A}' = \left\langle \begin{pmatrix} I & I \end{pmatrix}, \begin{pmatrix} 0 & F^+ \\ G^+ & 0 \end{pmatrix}, \begin{pmatrix} T \\ T \end{pmatrix} \right\rangle$$

whose coefficients are rationals of M and N. ■

[60] Throughout this section we steer between the subset interpretation and the 'boolean vector' interpretation of the symbols that denote the initial and final states.

[61] This does not mean that $\mathcal{A}$ is a *complete bipartite graph*.

It follows from the proof of Proposition 6.2 that if M and N have *no divisors of the identity* the bipartite automaton $\mathcal{A}'$ is proper. If we remember that it is the *proper* bipartite automata that play the role of automata over a free monoid, we shall see that this conclusion, which is important, holds under a much weaker assumption on M and N, but at the cost of a harder proof.

We say that a monoid M has the *detachable identity* property (DI) if the rationals of M are stable under subtraction of the neutral element:

$$\forall R \in \operatorname{Rat} M \qquad R \setminus 1_M = R \cap M_\bullet \in \operatorname{Rat} M \ .$$

cf. Exer. II.1.6, p. 225 Note that the property (DI) is obviously satisfied if $\operatorname{Rat} M$ is a Boolean algebra, and also if M has no divisors of the identity.

Theorem 6.1 *Let M and N be two monoids with the (DI) property. Every rational of $M*N$ is the behaviour of a proper rational bipartite automaton.*

Proof. We start with two matrix equalities whose proofs are left as exercises.

Lemma 6.3 *Let E be a square matrix which can be decomposed into the following blocks:*

$$E = \begin{pmatrix} 0 & V \\ X+Z & 0 \end{pmatrix} \ ,$$

where V, X and Z are square matrices (hence of the same size) and X is a boolean matrix. Then

(i) $$E^* = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \begin{pmatrix} 0 & (V\,X)^*V \\ Z & 0 \end{pmatrix}^* \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \ ;$$

(ii) $$E^* = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \begin{pmatrix} 0 & (V\,X)^*V \\ X+Z & 0 \end{pmatrix}^* \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \ . \qquad \blacksquare$$

Let $\langle I, E, T \rangle$ be a bipartite automaton constructed as in Proposition 6.2; that is,

$$E = \begin{pmatrix} 0 & V_0 \\ W_0 & 0 \end{pmatrix} \ ,$$

where V_0 and W_0 are two square matrices of the same size. For every integer n, let V_n, W_n, X_n, Z_n, U_n and Y_n be defined recursively by the following equations:

$$\begin{aligned} W_n &= X_{n+1} + Z_{n+1} && \text{where } X_{n+1} \text{ is boolean and } Z_{n+1} \text{ is proper;} \\ V_{n+1} &= (V_n\,X_{n+1})^*\,V_n = U_{n+1} + Y_{n+1} && \text{where } U_{n+1} \text{ is boolean and } Y_{n+1} \text{ proper;} \\ W_{n+1} &= (W_n\,U_{n+1})^*\,W_n \ . \end{aligned}$$

The (DI) property implies that if the entries of E are rational, then so are those of Y_n and Z_n, for all n. To simplify the notation, set

$$G_0 = D_0 \qquad \text{i.e. the identity matrix of the same size as } E,$$

$$L_n = \begin{pmatrix} 1 & 0 \\ X_n & 1 \end{pmatrix} \qquad \text{and} \qquad R_n = \begin{pmatrix} 1 & U_n \\ 0 & 1 \end{pmatrix} \ ,$$

$$G_n = G_{n-1}\,L_n\,R_n \qquad \text{and} \qquad D_n = R_n\,L_n\,D_{n-1} \ .$$

By Lemma 6.3, we then deduce

$$\forall n \in \mathbb{N} \qquad E^* = G_n \begin{pmatrix} 0 & V_n \\ W_n & 0 \end{pmatrix}^* D_n \ . \tag{6.2}$$

We note that for all n, $X_n \leqslant X_{n+1}$ and $U_n \leqslant U_{n+1}$ (as boolean matrices) and that thus the L_n, R_n, G_n and D_n are monotonic increasing sequences of boolean matrices: they are stationary from a certain rank k onwards. Set $X = X_k$, $U = U_k$, $L = L_k$, $R = R_k$, $G = G_k$ and $D = D_k$; we have

$$G\,L = G \ , \qquad G\,R = G \ , \qquad L\,D = D \ , \qquad R\,D = D \ .$$

From $V_k = (V_{k-1}\,X)^*\,V_{k-1} = U + Y_k$, we conclude

$$V_{k+1} = (V_k\,X)^*\,V_k = ((V_{k-1}\,X)^*\,V_{k-1}\,X)^*\,V_{k-1} = V_k = U + Y_{k+1} \ ,$$

and hence that the sequence of Y_n is stationary from $Y_k = Y$ (and $V_k = V = U+Y$); symmetrically, the sequence of the Z_n is stationary from $Z_k = Z$ (and $W_k = W = X + Z$). Set

$$F = \begin{pmatrix} 0 & Y \\ Z & 0 \end{pmatrix} \qquad \text{and} \qquad H = \begin{pmatrix} 0 & U \\ X & 0 \end{pmatrix} \ ,$$

and (6.2) becomes

$$E^* = G \cdot (F + H)^* \cdot D \ .$$

We verify $H \leqslant L\,R$ and $H \leqslant R\,L$, from which we obtain $G\,H \leqslant G$ and $H\,D \leqslant D$. From

$$[(U + Y)\,X]^*\,(U + Y) = U + Y \qquad \text{and} \qquad [(X + Z)\,U]^*\,(X + Z) = X + Z \ ,$$

we deduce

$$Y\,X\,Y \leqslant Y \ , \qquad Y\,X\,U \leqslant Y \ , \qquad Z\,U\,Z \leqslant Z \ , \qquad Z\,U\,X \leqslant Z \ ,$$

then

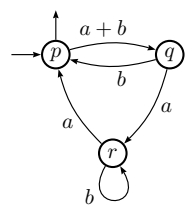
$$F\,H\,F \leqslant F \qquad \text{and} \qquad F\,H^{2n} \leqslant F \ ,$$

and finally

$$G\,F^*\,D = G\,(H + F)^*\,D \ .$$

The automaton $\langle\, I \cdot G, F, D \cdot T \,\rangle$ is (bipartite and) equivalent to $\langle\, I, E, T \,\rangle$ and is *proper*. ■

Example 6.2 Let F be the free group over two generators. We have $F = \mathbb{Z}_a * \mathbb{Z}_b$ where $\mathbb{Z}_a$ and $\mathbb{Z}_b$ are two distinct copies of $\mathbb{Z}$. Let $\mathcal{S}_1 = \langle\, I_1, E_1, T_1 \,\rangle$ be the automaton shown in the adjacent figure (with the convention that a transition (p, a, q) also represents the transition $(q, \overline{a}, p)$):



$$I_1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \ , \qquad E_1 = \begin{pmatrix} 0 & a + b + \overline{b} & \overline{a} \\ \overline{a} + b + \overline{b} & 0 & a \\ a & \overline{a} & b + \overline{b} \end{pmatrix} \ , \qquad T_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \ .$$

Set

$$X = \{a^n \mid n \in \mathbb{Z} \ , \quad n \equiv 0 \bmod 3 \quad \text{and} \quad n \neq 0\} \ ,$$

$$Y = \{a^n \mid n \in \mathbb{Z} \ , \quad n \equiv 1 \bmod 3\} \qquad \text{and} \qquad Z = \{a^n \mid n \in \mathbb{Z} \ , \quad n \equiv 2 \bmod 3\} \ ,$$

$$R = \{b^n \mid n \in \mathbb{Z} \ , \quad n \equiv 0 \bmod 2 \quad \text{and} \quad n \neq 0\}$$

$$\text{and} \qquad S = \{b^n \mid n \in \mathbb{Z} \ , \quad n \equiv 1 \bmod 2\} \ ,$$

with the convention that $a^n = \overline{a}^{-n}$ for negative n. It follows that

$$\mathbb{Z}_a = 1_{\mathbb{Z}_a} + X + Y + Z \qquad \text{and} \qquad \mathbb{Z}_b = 1_{\mathbb{Z}_b} + R + S \ .$$

When all the calculations have been performed, we obtain the proper bipartite automaton (which we note is deterministic and complete):

$$\left\langle \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} , \begin{pmatrix} & 0 & & X & Y & Z \\ & & & Z & X & Y \\ & & & Y & Z & X \\ R & S & 0 & & & \\ S & R & 0 & & 0 & \\ 0 & 0 & R+S & & & \end{pmatrix} , \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \right\rangle$$

shown in Figure 6.1. □

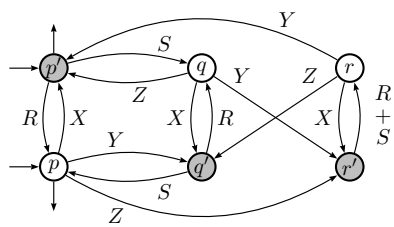


Figure 6.1: Proper bipartite automaton of Example 6.2

Exercises

• 6.5 Prove Lemma 6.3.

6.6 Perform the calculations that yield Example 6.2.

6.3 Bipartite deterministic automaton

We strengthen the parallel between (finite) automata over a free monoid and bipartite automata over a free product of monoids by proving that the latter can, like the former, be *determinised*, and with the same *subset construction*.

Theorem 6.2 *Every proper bipartite automaton $\mathcal{A}$ is equivalent to a deterministic and complete proper bipartite automaton, whose coefficients belong to the Boolean algebra generated by the coefficients of $\mathcal{A}$, and is effectively computable if this algebra is effective.*

Proof. We use the same notation as in (6.1):

$$\mathcal{A} = \left\langle \begin{pmatrix} I' & I'' \end{pmatrix}, \begin{pmatrix} 0 & V \\ W & 0 \end{pmatrix}, \begin{pmatrix} T' \\ T'' \end{pmatrix} \right\rangle .$$

Then $\mathcal{A}$ determines the two functions $\delta' \colon P \times M_\bullet \to \mathfrak{P}(Q)$ and $\delta'' \colon Q \times N_\bullet \to \mathfrak{P}(P)$ by

$$\forall p \in P\,,\ \forall m \in M_\bullet \qquad (p,m)\delta' = \{q \in Q \mid m \in V_{p,q}\}$$
$$\text{and} \qquad \forall q \in Q\,,\ \forall n \in N_\bullet \qquad (q,n)\delta'' = \{p \in P \mid n \in W_{q,p}\} \ .$$

These functions are extended by additivity to $\mathfrak{P}(P) \times M_\bullet$ and $\mathfrak{P}(Q) \times N_\bullet$ respectively by

$$\forall K \subseteq P \quad (K,m)\delta' = \bigcup_{p \in K} (p,m)\delta' \qquad \text{and} \qquad \forall L \subseteq Q \quad (L,n)\delta'' = \bigcup_{q \in L} (q,n)\delta'' ,$$

which in return enables us to define a bipartite automaton $\widehat{\mathcal{A}}$ of size $\mathfrak{P}(P) + \mathfrak{P}(Q)$. We set, for all K in $\mathfrak{P}(P)$ and all L in $\mathfrak{P}(Q)$,

$$\widehat{V}_{K,L} = \{m \in M_\bullet \mid (K,m)\delta' = L\} \qquad \text{and} \qquad \widehat{W}_{L,K} = \{n \in N_\bullet \mid (L,n)\delta'' = K\}$$

$$\widehat{I'}_K = 1 \Leftrightarrow K = \{p \in P \mid I'_p = 1\} \qquad \text{and} \qquad \widehat{I''}_L = 1 \Leftrightarrow L = \{q \in Q \mid I''_q = 1\}$$

$$\widehat{T'}_K = 1 \Leftrightarrow \exists p \in K \quad T'_p = 1 \qquad \text{and} \qquad \widehat{T''}_L = 1 \Leftrightarrow \exists q \in L \quad T'_q = 1$$

and finally

$$\widehat{\mathcal{A}} = \left\langle \begin{pmatrix} \widehat{I'} & \widehat{I''} \end{pmatrix}, \begin{pmatrix} 0 & \widehat{V} \\ \widehat{W} & 0 \end{pmatrix}, \begin{pmatrix} \widehat{T'} \\ \widehat{T''} \end{pmatrix} \right\rangle .$$

We first verify that the coefficients of $\widehat{\mathcal{A}}$ belong to the Boolean algebra generated by the coefficients of $\mathcal{A}$; more precisely:

Property 6.1

$$\forall K \in \mathfrak{P}(P)\,,\ \forall L \in \mathfrak{P}(Q) \qquad \widehat{V}_{K,L} = \left[\bigcap_{q \in L} \left(\bigcup_{p \in K} V_{p,q}\right)\right] \setminus \left[\bigcup_{q \notin L} \left(\bigcup_{p \in K} V_{p,q}\right)\right]$$

$$\textit{and} \qquad \widehat{W}_{L,K} = \left[\bigcap_{p \in K} \left(\bigcup_{q \in L} W_{q,p}\right)\right] \setminus \left[\bigcup_{p \notin K} \left(\bigcup_{q \in L} W_{q,p}\right)\right] \qquad \blacksquare$$

We then verify that $\widehat{\mathcal{A}}$ is equivalent to $\mathcal{A}$.

Property 6.2 $\quad |\widehat{\mathcal{A}}| = |\mathcal{A}|$. $\blacksquare$

Clearly, the automaton $\widehat{\mathcal{A}}$ is deterministic and complete and this completes the proof of the theorem. $\blacksquare$

As in the case of classic automata, if $\mathcal{A} = \langle I, E, T \rangle$ is a complete deterministic bipartite automaton, then the behaviour of the automaton $\overline{\mathcal{A}} = \langle I, E, \overline{T} \rangle$, where $\overline{T}$ is the 'complement' of T (that is, the 0's are replaced by 1's and *vice versa*) is the complement of the behaviour of $\mathcal{A}$. We deduce immediately:

Corollary 6.4 *If* $\mathrm{Rat}\, M$ *and* $\mathrm{Rat}\, N$ *are two (effective) Boolean algebras, then* $\mathrm{Rat}\, M{*}N$ *is also an (effective) Boolean algebra.* ■

If X is a subset of M and Y is a subset of N, the products XY and YX are *unambiguous*; if furthermore X *and* Y are proper, then XY and YX each generate a free submonoid of $M{*}N$. In particular, $M{*}N$ belongs to the unambiguous rational closure of $M_\bullet$ and $N_\bullet$:

$$M{*}N = 1_{M*N} + (M_\bullet N_\bullet)^+(1_M + M_\bullet) + (N_\bullet M_\bullet)^+(1_N + N_\bullet) \tag{6.3}$$

This unambiguous decomposition of $M{*}N$ is the exact counterpart of the unambiguous generation of a free monoid A^* by its basis A and we easily derive from it the generalisation of the fact that (classic) deterministic automata are unambiguous.

Proposition 6.5 *The behaviour of a proper bipartite deterministic automaton belongs to the unambiguous rational closure of its coefficients.* ■

By Corollary 6.4 and Proposition 6.5, we then obtain:

Corollary 6.6 *If* $\mathrm{Rat}\, M$ *and* $\mathrm{Rat}\, N$ *are two Boolean algebras and* $\mathrm{Rat}\, M = \mathrm{URat}\, M$ *and* $\mathrm{Rat}\, N = \mathrm{URat}\, N$ *, then* $\mathrm{Rat}\,(M{*}N) = \mathrm{URat}\,(M{*}N)$ *.* ■

cf. Th. II.7.5 and Th. II.7.6, p. 342

Note in particular that Corollaries 6.4 and 6.6 apply in particular to *free products of free commutative monoids.*

Exercise

6.7 **Subset construction.** Prove Property 6.1.

6.4 Minimal deterministic bipartite automaton

We pursue the parallel between classic automata and bipartite automata; we now show that we can relate a (finite) minimal bipartite automaton to each (rational) subset of a free product.

If R is a subset (*a priori* arbitrary) of $M{*}N$, and, as in Chapter II, we write ρ_R for the coarsest right regular equivalence which saturates R; that is,

$$u \cong v \bmod \rho_R \quad \Longleftrightarrow \quad u^{-1}R = v^{-1}R \ ,$$

and σ_R for the coarsest congruence on $M{*}N$ which saturates R:

$$u \equiv v \bmod \sigma_R \quad \Longleftrightarrow \quad C_R(u) = C_R(v) \ ,$$

where $$C_R(u) = \{(x,y) \in (M{*}N)\times(M{*}N) \mid x\,u\,y \in R\} \ .$$

Let τ be the equivalence defined on $M{*}N$ by the equality of final monoids:

$$u \approx v \bmod \tau \quad \Longleftrightarrow \quad \mathrm{MF}(u) = \mathrm{MF}(v) \ .$$

This three-class equivalence corresponds to the following partition of $M*N$, which is a rewriting of (6.3):

$$M*N = 1_{M*N} + \{(1_N + N_\bullet)(M_\bullet N_\bullet)^+\} + \{(1_M + M_\bullet)(N_\bullet M_\bullet)^+\}$$

We note that if M and N have no divisors of the identity, then τ is a congruence. We then define the equivalences γ_R on $M*N$ by the following conditions:

$$u \cong v \bmod \gamma_R \quad \Longleftrightarrow \quad \begin{cases} \text{(i)} & u \approx v \bmod \tau \\ \text{(ii)} & u \in R \Leftrightarrow v \in R \\ \text{(iii)} & \forall w\,,\ \mathrm{MI}(w) \neq \mathrm{MF}(u) \quad u\,w \in R \Leftrightarrow v\,w \in R \end{cases}$$

In other words, $u \cong v \bmod \gamma_R$ if and only if u and v have the *same immiscible right contexts* for R. By (i), 1_{M*N} is unique in its class modulo γ_R. The following properties are immediate.

Property 6.3 (i) *γ_R is an equivalence which saturates R;*
(ii) *$\rho_R \wedge \tau$ is finer than γ_R;*
(iii) *γ_R is not necessarily right regular but we have*

$$u \cong v \bmod \gamma_R \quad \Longleftrightarrow \quad \forall w\,,\ \mathrm{MI}(w) \neq \mathrm{MF}(u) \qquad u\,w \cong v\,w \bmod \gamma_R\ . \qquad \blacksquare$$

Furthermore, we shall adapt the notion of an *action* to the case of complete deterministic bipartite automata. Let $\mathcal{A} = \langle I, E, T \rangle$ be such an automaton of size $P+Q$ over $M*N$ and let p be in P, q in Q, and u in $M*N$. If $\mathrm{MI}(u) = M$, there exists a unique computation of $\mathcal{A}$, with source p and label u; we write $p \cdot u = t$ for its destination. Symmetrically, if $\mathrm{MI}(u) = N$, there exists a unique computation of $\mathcal{A}$, with source q and label u; we write $q \cdot u = t$ for its destination. Thus $\mathcal{A}$ determines a map, not totally defined, from $(P+Q)\times M*N$ to $P+Q$; in the cases not covered above, $s \cdot u$ is undefined. Since it is partial, this map is not an action, but the following property justifies the notation.

Property 6.4 $\forall s \in P+Q\,,\ \forall u, w \in M*N$
(i) $s \cdot u$ *defined* $\Longrightarrow$ $[(s \cdot u) \cdot w$ *defined* $\Leftrightarrow$ $\mathrm{MF}(u) \neq \mathrm{MI}(w)]$;
(ii) $(s \cdot u) \cdot w$ *defined* $\Longrightarrow$ $[(s \cdot u) \cdot w = s \cdot (u\,w)]$. $\blacksquare$

We write $I \cdot u$ to denote the destination of the unique computation of $\mathcal{A}$ labelled u whose source is the unique initial state of I' or of I'' according to whether $\mathrm{MI}(u) = M$ or N and we have:

$$|\mathcal{A}| = \{u \in M*N \mid I \cdot u \in T\}\ .$$

We finally say that $\mathcal{A}$ is *normalised* if its initial states are not the destination of any transition.

Lemma 6.7 *Let $\mathcal{A}$ be a normalised complete deterministic bipartite automaton and $R = |\mathcal{A}|$ its behaviour. It then follows that*

$$\forall u, v \in M*N \qquad I \cdot u = I \cdot v \quad \Longrightarrow \quad u \cong v \bmod \gamma_R\ .$$

Proof. If $u \neq 1_{M*N}$, $I \cdot u$ is disjoint from I, since $\mathcal{A}$ is normalised, and hence $u = 1_{M*N} \Leftrightarrow v = 1_{M*N}$. Suppose $u \neq 1_{M*N}$, then:

(i) $I \cdot u = I \cdot v \quad \Longrightarrow \quad \mathrm{MF}(u) = \mathrm{MF}(v)$ and hence $u \approx v \bmod \tau$;

(ii) $I \cdot u = I \cdot v \quad \Longrightarrow \quad [u \in R \Leftrightarrow v \in R]$;

(iii) $\forall w \in M*N \qquad \mathrm{MI}(w) \neq \mathrm{MF}(u) \quad \Longrightarrow$

$I \cdot (u\,w) = (I \cdot u) \cdot w = (I \cdot v) \cdot w = I \cdot (v\,w)$ and hence $u\,w \in R \Leftrightarrow v\,w \in R$. ■

Thus, two 'words' u and v which take $\mathcal{A}$ to the same state are equivalent modulo γ_R. Conversely, this equivalence γ_R allows us to define a normalised complete deterministic bipartite automaton. We write C and D for the two classes, other than 1_{M*N}, of τ,

$$C = \big\{(1_N + N_\bullet)(M_\bullet N_\bullet)^+\big\} \qquad \text{and} \qquad D = \big\{(1_M + M_\bullet)(N_\bullet M_\bullet)^+\big\} \ .$$

Let R be an arbitrary subset of $M*N$ and γ_R the equivalence associated with it. We rewrite Property 6.3 (iii) in the form

$$\begin{array}{lll} \forall u, v \in \{1_{M*N}\} + C\,,\ \forall m \in M & u \cong v \bmod \gamma_R & \Rightarrow \quad u\,m \cong v\,m \bmod \gamma_R \\ \forall u, v \in \{1_{M*N}\} + D\,,\ \forall n \in N & u \cong v \bmod \gamma_R & \Rightarrow \quad u\,n \cong v\,n \bmod \gamma_R \ . \end{array}$$

Let P be the set of classes modulo γ_R in $\{1_{M*N}\} + C$ and Q the set of classes modulo γ_R in $\{1_{M*N}\} + D$. Let u be in $\{1_{M*N}\} + C$ and $[u]$ its class modulo γ_R. For each m in M, the equality

$$[u] \cdot m = [u\,m]$$

defines a function from $P \times M$ to Q. Symmetrically, for each n in N, the equality

$$[u] \cdot n = [u\,n]$$

defines a function from $Q \times N$ to P. We then define a bipartite automaton $\mathcal{A}_R$ of dimension $P + Q$: for each pair (p, q) of $P \times Q$, the transition from p to q is labelled

$$V_{p,q} = \{m \in M_\bullet \mid p \cdot m = q\} \ ,$$

and the transition from q to p is labelled

$$W_{q,p} = \{n \in N_\bullet \mid q \cdot n = p\} \ .$$

For each p in P, $\{V_{p,q} \mid q \in Q\}$ is a partition of $M_\bullet$ and for each q in Q, $\{W_{q,p} \mid p \in P\}$ is a partition of $N_\bullet$: $\mathcal{A}_R$ is proper, deterministic and complete. Its initial states are the classes of 1_{M*N} in P and Q and $\mathcal{A}_R$ is normalised. Its construction itself implies that the behaviour of $\mathcal{A}_R$ is equal to R (but not that $\mathcal{A}_R$ is *finite*). We call $\mathcal{A}_R$ the *minimal bipartite automaton* of R.

Proposition 6.8 *Every normalised complete deterministic proper bipartite automaton, whose behaviour is R, is a covering of the minimal bipartite automaton of R.* ■

Hence we obtain:

Corollary 6.9 *The minimal bipartite automaton of a rational subset of $M*N$ is finite.* ■

Of course, a subset of $M*N$ whose minimal bipartite automaton is finite is not necessarily rational. In a particular case, the minimal bipartite automaton enables us to characterise the *recognisable subsets* of free products.

Theorem 6.3 *Let M and N be two monoids* without divisors of the identity. *A subset R of $M*N$ is recognisable if and only if it is the behaviour of a finite bipartite automaton all of whose coefficients are recognisable subsets of M or of N.*

Proof. The condition is sufficient. The behaviour R of a finite bipartite automaton $\mathcal{A}$ is a rational expression E in the coefficients of $\mathcal{A}$:

$$R = \mathsf{E}\left(\{V_{p,q}, W_{q,p} \mid p \in P\,,\ q \in Q\}\right) \quad .$$

Since the two monoids are without divisors of the identity and by Lemma 6.1 the inverse image of R under $(A \cup B)^*$ is equal to the same expression in the inverse images of coefficients of $\mathcal{A}$:

$$R\,[\varphi*\psi]^{-1} = \mathsf{E}\left(\{(V_{p,q})\varphi^{-1}, (W_{q,p})\psi^{-1} \mid p \in P\,,\ q \in Q\}\right) \quad .$$

Since the $V_{p,q}$ and the $W_{q,p}$ are recognisable, their inverse images are rational, so $R\,[\varphi*\psi]^{-1}$ is rational, hence R is recognisable. *cf. Cor. II.2.12, p. 250*

The following proposition implies that the condition is necessary, and gives a more precise statement.

Proposition 6.10 *Let M and N be two monoids without divisors of the identity. The minimal bipartite automaton of a recognisable subset R is a finite bipartite automaton all of whose coefficients are recognisable subsets of M or N.*

Proof. Let R be a recognisable subset of $M*N$ and $\mathcal{A}_R$ its minimal bipartite automaton. The equivalence ρ_R and the congruence σ_R are of finite index and hence so are $\rho_R \wedge \tau$ and $\sigma_R \wedge \tau$. By Property 6.3 (ii), γ_R is of finite index and $\mathcal{A}_R$ is a finite automaton of dimension $P + Q$.

Let us fix p in P and let θ_p be the coarsest right regular equivalence on M which saturates the partition $\{V_{p,q} \mid q \in Q\}$; that is,

$$\forall f, g \in M \qquad f \cong g \bmod \theta_p \iff [\forall h \in M\,,\ \forall q \in Q \quad f\,h \in V_{p,q} \Leftrightarrow g\,h \in V_{p,q}] \quad .$$

We shall show that $\sigma_R \wedge \tau$ is finer than θ_p. The identity 1_M is unique in its class modulo τ and also modulo θ_p since it is the only element of M which does not belong to any $V_{p,q}$. Let f and g be elements of $M_\bullet$ such that

$$f \cong g \bmod \sigma_R \wedge \tau \quad . \tag{6.4}$$

Since M and N are without divisors of the identity, τ is a congruence and (6.4) implies

$$\forall u, w \in M*N \qquad u\,f\,w \in R \Leftrightarrow u\,g\,w \in R$$

which can be rewritten as

$$\forall h \in M\,,\ \forall u, w \in M*N\,,\ \mathrm{MI}(w) = N \qquad u\,f\,h\,w \in R \Leftrightarrow u\,g\,h\,w \in R\ . \tag{6.5}$$

If we choose u such that $I \cdot u = p$, (6.5) becomes

$$\forall h \in M \qquad u\,f\,h \cong u\,g\,h \bmod \gamma_R$$

or again

$$\forall h \in M \qquad [u] \cdot f\,h = [u] \cdot g\,h\ ,$$

which means that $f\,h$ and $g\,h$ belongs to the same $V_{p,q}$ with $p = [u]$ and $q = [u\,f\,h]$. Thus $\sigma_R \wedge \tau$ is finer than θ_p which is thus of finite index: the $V_{p,q}$ are recognisable. By symmetric reasoning the $W_{q,p}$ are recognisable, which proves the proposition and hence Theorem 6.3. ■

Corollary 6.11 *Let M and N be two monoids without divisors of the identity. If* $\mathrm{Rat}\,M$ *and* $\mathrm{Rat}\,N$ *are two effective Boolean algebras in which the recognisable subsets are decidable, then so is* $\mathrm{Rat}\,M*N$.

Proof. Let R be a rational subset of $M*N$. By Proposition 6.2, R is the behaviour of a rational bipartite automaton, and by Theorem 6.1 that of a *proper* bipartite automaton, say $\mathcal{B}$ of dimension $P + Q$, on $\mathrm{Rat}\,M$ and $\mathrm{Rat}\,N$, whose coefficients are also rational and effectively computable, by assumption on $\mathrm{Rat}\,M$ and $\mathrm{Rat}\,N$.

By trying *all the partitions* of P and Q, and by testing the equality of the behaviour of the automata thus obtained with R, which is effective by Corollary 6.4, we can compute the minimal automaton $\mathcal{A}_R$. It only remains to test whether the coefficients of $\mathcal{A}_R$ are recognisable, which is also decidable by assumption on $\mathrm{Rat}\,M$ and $\mathrm{Rat}\,N$. ■

Remark 6.1 Example 6.2 shows that Proposition 6.10 no longer holds without the assumption that M and N have no divisors of the identity, but this example does not invalidate the conclusions of Corollary 6.11. □

Exercise

- 6.8 Characterise the subsets of $M*N$ whose minimal bipartite automaton is finite.

7 A non-commutative linear algebra primer

The following pages gather together, for readers unfamiliar with the subject, the definitions and rudiments of linear algebra used in this chapter, essentially what one might call a simple theory of dimension and rank. One of the aims is to prove Theorem 4.7, which was at the heart of the reduction procedure of rational series (over a free monoid) with coefficients in a field:

Theorem 4.7 *The row rank of a matrix over a skew field, if it is finite, is equal to its column rank.*

The only originality with respect to the manuals of algebra in which these questions are classically treated is that I shall work solely with skew fields: that is, without assuming that the coefficients commute. This does not complicate matters but requires solely that we distinguish *left* vector spaces from *right* vector spaces; for example, between the left vector space generated by the *rows* of a matrix and the right vector space generated by its *columns*.

To start with, $\mathbb{K}$ is a non-commutative semiring.

Modules

A *(left) module*[62] over $\mathbb{K}$, or $\mathbb{K}$-*module*, is a *commutative* monoid U, in general written additively, equipped with an exterior law $\mathbb{K} \times U \to U$, written by simple juxtaposition:[63] $(k, u) \mapsto k\,u$, which satisfies the following axioms.

• Mod1 The exterior multiplication is consistent with multiplication in $\mathbb{K}$:

cf. the definition of an action *Section II.2.1*

$$\forall k, l \in \mathbb{K}\,,\ \forall u \in U \qquad (kl)\,u = k\,(l\,u) \qquad \text{and} \qquad 1_{\mathbb{K}}\,u = u \ .$$

• Mod2 The exterior multiplication distributes over the addition of $\mathbb{K}$ and that of U:

$$\forall k, l \in \mathbb{K}\,,\ \forall u, v \in U \qquad k\,(u + v) = (k\,u + k\,v) \qquad \text{and} \qquad (k + l)\,u = k\,u + l\,u \ .$$

• Mod3 The zero of $\mathbb{K}$ sends every element of U to its zero:

$$\forall u \in U \qquad 0_{\mathbb{K}}\,u = 0_U \ .$$

We define *right modules* over $\mathbb{K}$ symmetrically. If the semiring $\mathbb{K}$ is *commutative* the notions of left and right modules coincide. In general, we shall consider left modules, which we shall call simply 'modules'.

Every semiring $\mathbb{K}$ is a left module over itself; it is also a right $\mathbb{K}$-module, but the two structures are not isomorphic if $\mathbb{K}$ is not commutative. The same is true for $\mathbb{K}^n$, for every integer n. Every commutative monoid is an $\mathbb{N}$-module.

A *submodule* of U is a submonoid of U that is stable for the exterior multiplication. Let S be a (possibly infinite) subset of a $\mathbb{K}$-module U. A *linear combination* of elements of S is a sum:

$$t = \sum_{s \in S} k_s\,s \ ,$$

where the k_s, elements of $\mathbb{K}$, are *all zero except for a finite number*, which ensures that the sum t is well defined. The linear combination is *non-trivial* if at least one of the k_s is non-zero. The set of linear combinations of elements of S, written $\langle S \rangle$, is a submodule of U, contained in all the submodules of U that contain S: it is the

[62] The usual term is *semi-module* [218, 142]. I follow Berstel and Reutenauer [31] who say more simply, and, in my opinion, more elegantly, 'module'. It is the (semiring or ring) structure of $\mathbb{K}$ which determines the structure of U.

[63] I have added a little space between the two terms in the hope of improving legibility.

submodule generated by S. A module U is called *of finite type* if it is generated by a finite set. The module $\mathbb{K}^n$ is of finite type: the set B of vectors of dimension n all of whose entries are zero except for one, which has the value $1_{\mathbb{K}}$, generates $\mathbb{K}^n$.

A subset S of U is *free*, or *linearly independent*, if two linear combinations are equal only if their coefficients are term-wise equal; that is, if

$$\sum_{s\in S} k_s\, s = \sum_{s\in S} l_s\, s \qquad \Longrightarrow \qquad \forall s \in S \qquad k_s = l_s \ .$$

This implies that *if* S is free, *then* any non-trivial linear combination of elements of S is non-zero; that is,

$$S \quad \text{free} \qquad \Longrightarrow \qquad \left\{ \sum_{s\in S} k_s\, s = 0_U \quad \Rightarrow \quad \forall s \in S \quad k_s = 0_{\mathbb{K}} \right\} \ . \qquad (*)$$

A subset S is a *basis* of U if it is free and generates U. A module is called *free* if it has a basis. The module $\mathbb{K}^n$ is free, with basis B.

A *morphism* of left $\mathbb{K}$-modules $\alpha\colon U \to V$ is a morphism[64] of monoids which satisfies

$$\forall k \in \mathbb{K}\,,\ \forall u \in U \qquad (k\,u)\alpha = k\,[u\alpha] \ .$$

The image $\mathsf{Im}\,\alpha$ of α is a submodule of V.

Modules over a ring

This is the standard notion. We showed above that the usual definitions can be applied without difficulty to the more general setting of semirings. However, the assumption that $\mathbb{K}$ is a ring implies some strong properties for a (left) $\mathbb{K}$-module U which are worth mentioning. First, thanks to subtraction, a subset S of U such that any non-trivial combination of elements of S is non-zero, is free (or linearly independent); that is, the converse implication of $(*)$ is true. Next, U is necessarily a commutative *group*, and because of this, the congruences and submodules of U coincide.

More precisely, every submodule V of U defines a congruence α by the relation

$$x \equiv y \quad [\alpha] \quad \Leftrightarrow \quad x - y \in V \ , \qquad (**)$$

so V is the set of elements of U that are congruent to 0_U modulo α. Conversely, every congruence α of U can be defined by $(**)$ by taking for V the set of elements of U congruent to 0_U modulo α. The quotient of U by the congruence defined by a submodule V is written U/V.

If $\alpha\colon U \to W$ is a morphism of $\mathbb{K}$-modules, then the *kernel* of α, written $\mathsf{Ker}\,\alpha$, is the submodule $(0_W)\alpha^{-1}$. The map equivalence of α is a congruence, that defined by $\mathsf{Ker}\,\alpha$, and we easily verify that $\mathsf{Im}\,\alpha$ is isomorphic to $U/\mathsf{Ker}\,\alpha$.

[64] In this case we appreciate postfix notation for function symbols enormously. So much so, that we write morphisms of right modules *on the left*.

Vector spaces

Usage dictates that we call a module over a *field* a *vector space*. A vector space has a basis (it is a free module) and two distinct bases have the same number of elements. The classic proof does not require commutativity of 'coefficients' and the same properties hold for modules over *skew fields* which we also call vector spaces.[65]

Theorem 7.1 *Let V be a vector space over a skew field $\mathbb{K}$. All the bases of V have the same cardinal. This common cardinal is called the* dimension *of V over $\mathbb{K}$ and is written* $\dim V$ *(sometimes* $\dim_{\mathbb{K}} V$*).*

Proof. (a) Let B and G be two subsets of V such that $B \subseteq \langle G \rangle$. Each element of B is a linear combination of elements of G and, as every linear combination of linear combinations of elements of G is a linear combination of elements of G, it follows that $\langle B \rangle \subseteq \langle G \rangle$.

(b) Let S be a free subset of V and x an element of V which does not belong to $\langle S \rangle$, the subspace generated by S. Then $S \cup \{x\}$ is a free subset. Indeed, there otherwise exist scalars $\{k_s\}_{s \in S}$ and l in $\mathbb{K}$ such that

$$\sum_{s \in S} k_s\, s + l\, x = 0_U \quad . \tag{7.1}$$

If $l \neq 0_{\mathbb{K}}$, then $x = \sum_{s \in S} -(l^{-1} k_s)\, s$ belongs to $\langle S \rangle$, which is a contradiction. If $l = 0_{\mathbb{K}}$, (7.1) contradicts the assumption that S is free.

(c) Let S be a free subset of V and G a generating set of V which contains S. Then there exists a basis of V that is contained in G and contains S.

Indeed, let $\mathcal{T}$ be the set of subsets of G which contain S and which are linearly independent; $\mathcal{T}$ is non-empty since it contains S. If $\{T_i\}_{i \in I}$ is a subset of $\mathcal{T}$ (that is totally ordered by inclusion) then $\bigcup_{i \in I} T_i$ is also a set of linearly independent vectors of G that contain S and hence belong to $\mathcal{T}$.

Let B be a maximal element[66] of $\mathcal{T}$. If B is not a basis of V, i.e. if $\langle B \rangle \neq V = \langle G \rangle$, we find an x in G which is not in $\langle B \rangle$ by (a); $B \cup \{x\}$ is free by (b); we have a contradiction with the maximality of B.

(d) Let $S = \{s_1, s_2, \ldots, s_n\}$ and $T = \{t_1, t_2, \ldots, t_m\}$ be two finite free subsets of V. If $T \subseteq \langle S \rangle$, then $m \leqslant n$. In fact, if $T \subseteq \langle S \rangle$ there exist some k_i, $i \in [n]$, such that

$$t_1 = k_1\, s_1 + k_2\, s_2 + \cdots + k_n\, s_n \quad , \tag{7.2}$$

and we can assume, up to renumbering the s_i, that $k_1 \neq 0_{\mathbb{K}}$. Thus, s_1 is in the subspace generated by $U_1 = \{t_1, s_2, \ldots, s_n\}$, hence equal to $\langle S \rangle$. If t_1 belonged to the subspace generated by $\{s_2, \ldots, s_n\}$, we would deduce by (7.2), since $k_1 \neq 0_{\mathbb{K}}$, a linear dependence relation between the s_i, which cannot exist since S is free. Hence, by (b), U_1 is free.

---

[65] In this, as well as in the proof of the next two theorems, we follow S. Lang, *Algebra*, [145].

[66] If G is infinite, the axiom of choice (or Zorn's Lemma) is necessary to ensure the existence of such a maximal element.

Hence suppose that for all r, $1 \leqslant r < n$, and up to a renumbering of the s_i, we have shown that $\{t_1, t_2, \ldots, t_r, s_{r+1}, \ldots, s_n\}$ is a basis of $\langle S \rangle$. We write that t_{r+1} belongs to $\langle S \rangle$:

$$t_{r+1} = a_1 t_1 + a_2 t_2 + \cdots + a_r t_r + b_{r+1} s_{r+1} + \cdots + b_n s_n \quad .$$

Not all the b_j can be zero, because the t_j are linearly independent. We can suppose that $b_{r+1} \neq 0_{\mathbb{K}}$ and by an argument identical to the preceding one, we prove that $\{t_1, t_2, \ldots, t_r, t_{r+1}, \ldots, s_n\}$ is a basis of $\langle S \rangle$. We can repeat the procedure until $r = \mathsf{min}(m, n)$. If $n < m$, it follows that t_{n+1} belongs to the subspace generated by $\{t_1, t_2, \ldots, t_n\}$; hence $m \leqslant n$.

From this we deduce that if S and T are two bases of V of finite cardinal n and m respectively, we must have both $m \leqslant n$ and $n \leqslant m$. The case of infinite bases is left as an exercise to the reader. ■

Corollary 7.1 *Let V be a vector space of finite dimension over a skew field $\mathbb{K}$ and W a subspace of V. If $\mathsf{dim}\, W = \mathsf{dim}\, V$, then $W = V$.*

Proof. Let T be a basis of W. If $W = \langle T \rangle$ is strictly contained in v, there exists x in $V \setminus W$, and $T \cup \{x\}$ is free. This is impossible if a basis of V has the same cardinal as T; hence $W = V$. ■

Theorem 7.2 *Let V be a vector space of finite dimension over a skew field $\mathbb{K}$ and $\alpha \colon V \to U$ a morphism (of vector spaces). Then*

$$\mathsf{dim}\, V = \mathsf{dim}\, \mathsf{Ker}\, \alpha + \mathsf{dim}\, \mathsf{Im}\, \alpha \quad . \tag{7.3}$$

Proof. (a) Two subspaces X and Y of V are called *supplementary (in V)* if every v in V decomposes *uniquely* into a sum $v = x + y$ with x in X and y in Y. (We also say in this case that Y is a *supplement* (space) of X.) We shall first show that every subspace X of V has (at least) one supplement and that if V is of finite dimension, and if X and Y are supplementary, then $\mathsf{dim}\, V = \mathsf{dim}\, X + \mathsf{dim}\, Y$.

Let $B = \{b_i\}_{i \in I}$ be a basis of X. As $B \cup (V \setminus X)$ generates V, we can find a basis S of V which contains B, say $S = B \cup C$. Then $Y = \langle C \rangle$ is a supplement subspace of X. Indeed, every v in V can be written

$$v = \sum_{i \in I} k_i b_i + \sum_{j \in J} h_j c_j = x + y$$

(with the k_i and h_j almost all zero) since S generates V. If $v = x' + y'$ with $x' \in X$ and $y' \in Y$, then $x' = \sum_{i \in I} k'_i b_i$ and $y' = \sum_{j \in J} h'_i c_j$ from which we obtain

$$\sum_{i \in I} (k_i - k'_i)\, b_i + \sum_{j \in J} (h_j - h'_j)\, c_j = 0_V \quad ,$$

a linear combination which cannot be non-trivial since S is a basis; hence $x = x'$ and $y = y'$.

Conversely, if X and Y are supplementary spaces, the union of a basis B of X and a basis C of Y gives a basis S of V. In fact, S generates V and if we had a linear combination such that

$$\sum_{i\in I} k_i\, b_i + \sum_{j\in J} h_j\, c_j = 0_V \quad ,$$

we would have $\sum_{i\in I} k_i\, b_i = 0_V$ and $\sum_{j\in J} h_i\, c_j = 0_V$, a *unique* decomposition of 0_V in terms of X and Y and the conclusion follows since B and C are free.

Hence $\mathsf{dim}\, V = \mathsf{dim}\, X + \mathsf{dim}\, Y$ if V is of finite dimension.

(b) Let $\{u_i\}_{i\in I}$ be a basis of $\mathsf{Im}\,\alpha$ and $\{w_j\}_{j\in J}$ a basis of $\mathsf{Ker}\,\alpha$. Let $B = \{v_i\}_{i\in I}$ be a family of elements of V such that $v_i\alpha = u_i$ for each i in I. Note first that B is free since the image under α of a linear combination of the v_i is a linear combination of the u_i, which must be trivial to be zero; that is, if $x \in X = \langle B \rangle$ and $x\alpha = 0_U$ then $x = 0_V$.

Let v be in V; there exists some k_i (almost all zero) such that

$$v\alpha = \sum_{i\in I} k_i\, u_i \quad , \qquad \text{whence} \qquad \left(v - \sum_{i\in I} k_i\, v_i\right)\alpha = v\alpha - \sum_{i\in I} k_i\, v_i\alpha = 0_V \quad .$$

Thus, $v - \sum_{i\in I} k_i\, v_i$ is in $Y = \mathsf{Ker}\,\alpha$ and $X \cup Y$ generates V. The two subspaces X and Y are supplementary since if we had a double decomposition $v = x + y = x' + y'$ with x and x' in X and y and y' in Y, it would follow that $(x - x')\alpha = 0_V$ hence $x = x'$ by the preceding remark, and as a result $y = y'$. The theorem follows from (a). ■

Linear functionals

If V is a left $\mathbb{K}$-vector space, a (linear) $\mathbb{K}$-*functional on* V is a morphism from V to $\mathbb{K}$. Since $\mathbb{K}$, whether commutative or not, is as much a right $\mathbb{K}$-vector space as a left $\mathbb{K}$-vector space, the set of $\mathbb{K}$-functionals on V is naturally a *right* $\mathbb{K}$-*vector space*, called the *dual* of V and written[67] $V^\star$:

$$\forall \sigma, \tau \in V^\star\,,\ \forall v \in V\,,\ \forall k \in \mathbb{K} \qquad v\,[\sigma + \tau] = v\sigma + v\tau \quad \text{and} \quad v\,[\sigma\, k] = (v\sigma)\, k\ . \tag{7.4}$$

For v in V and τ in $V^\star$, we write $\langle v, \tau \rangle$ instead of $(v)\tau$, which gives V and $V^\star$ a symmetric role. This symmetry is even more marked in finite dimensions because of the following result, for the statement of which we use the *Kronecker delta* $\delta_{i,j}$: $\delta_{i,j} = 1_{\mathbb{K}}$ if $i = j$, $\delta_{i,j} = 0_{\mathbb{K}}$ otherwise.

Proposition 7.2 *Let V be a $\mathbb{K}$-vector space of finite dimension n and let $B = \{b_1, b_2, \ldots, b_n\}$ be a basis of V. The set $B^\star = \{\beta_1, \beta_2, \ldots, \beta_n\}$ of $\mathbb{K}$-functionals defined by*

$$\forall i, j \in [n] \qquad \langle b_i, \beta_j \rangle = \delta_{i,j}$$

is a basis of $V^\star$, and hence $\mathsf{dim}\, V = \mathsf{dim}\, V^\star$.

[67] The notation is too standard to be able to change it. Context prevents us from confusing it with the free monoid generated by V and in any case this notion is not central to our discussion and will not be much used. Finally, the observant reader will have noticed that in any case we are using a different star.

Proof. Let σ be in $V^\star$ and, for each i in $I = [n]$, $k_i = \langle b_i, \sigma\rangle$. For all x in V, $x = \sum_{i\in I} x_i\, b_i$, we have

$$\begin{aligned}\langle x, \sigma\rangle &= \langle \sum_{i\in I} x_i\, b_i, \sigma\rangle = \sum_{i\in I} x_i \langle b_i, \sigma\rangle = \sum_{i\in I} x_i\, k_i \\ &= \sum_{i\in I}\sum_{j\in I} x_i \langle b_i, \beta_j\rangle k_j = \sum_{j\in I}\sum_{i\in I} x_i \langle b_i, \beta_j\rangle k_j = \sum_{j\in I} \langle x, \beta_j\rangle k_j\end{aligned}$$

which is the same as saying that $\sigma = \sum_{j\in I} \beta_j\, k_j$, and hence that $B^\star$ generates $V^\star$ (this last assertion is only possible as I is finite; otherwise, there would be no reason for the k_j to be almost all zero).

Furthermore, $B^\star$ is free since an arbitrary linear combination $\sum_{j\in I} \beta_j\, h_j$ is a functional which takes the value h_i on b_i and can thus only be the zero functional if it is trivial. ■

If W is a subspace of V, the set of $\mathbb{K}$-functionals on V which take the value 0 at every element of W is a subspace of $V^\star$, called the *orthogonal* of W and written $W^\perp$:

$$W^\perp = \{\tau \in V^\star \mid \forall w \in W \quad \langle w, \tau\rangle = 0_{\mathbb{K}}\} \ .$$

Theorem 7.3 *Let V be a $\mathbb{K}$-vector space of finite dimension and W a subspace of V. Then*

$$\mathsf{dim}\, V = \mathsf{dim}\, W + \mathsf{dim}\, W^\perp \ . \tag{7.5}$$

Proof. (a) Let $\alpha\colon V \to U$ be a morphism of *left vector spaces*. For all σ in $U^\star$, we write $\alpha^\star(\sigma)$ for the map from V to $\mathbb{K}$ defined by

$$\forall v \in V \qquad v\,[\alpha^\star(\sigma)] = \langle v\alpha, \sigma\rangle \ .$$

This map $\alpha^\star(\sigma)$ is a linear functional on V:

$$\begin{aligned}(v+w)\,[\alpha^\star(\sigma)] &= \langle (v+w)\alpha, \sigma\rangle = \langle v\alpha, \sigma\rangle + \langle w\alpha, \sigma\rangle \\ &= v\,[\alpha^\star(\sigma)] + w\,[\alpha^\star(\sigma)] \\ (k\,v)\,[\alpha^\star(\sigma)] &= \langle (k\,v)\alpha, \sigma\rangle = \langle k\,(v\alpha), \sigma\rangle = k\langle v\alpha, \sigma\rangle = k\ (v\,[\alpha^\star(\sigma)]) \ .\end{aligned}$$

We shall therefore write from now on $\langle v, \alpha^\star\sigma\rangle$ for $v\,[\alpha^\star(\sigma)]$. We then easily verify that $\alpha^\star$ is a morphism[68] of (right) vector spaces, from $U^\star$ to $V^\star$.

(b) We apply the above calculation to the canonical injection $\iota\colon W \to V$ of a subspace W of V into V.

Let T be a supplementary space of W: for each v in V there exists a unique w in W and a unique t in T such that $v = w + t$. For each σ in $W^\star$, we define a map τ from V to $\mathbb{K}$ by

$$v\tau = \langle w, \sigma\rangle \ .$$

The uniqueness of the decomposition ensures that τ is a functional: that is, belongs to $V^\star$. For all w in W, we have

$$\langle w, \sigma\rangle = \langle w, \tau\rangle = \langle w\iota, \tau\rangle = \langle w, \iota^\star\tau\rangle \ ,$$

[68] Morphisms of *right* vector spaces are written *on the left* of their argument (*cf.* Note 64).

which shows that $\iota^\star\tau = \sigma$ and hence that $\iota^\star\colon V^\star \to W^\star$ is *surjective*. Furthermore, we calculate

$$\begin{aligned}\mathsf{Ker}\,\iota^\star &= \{\tau \in V^\star \mid \iota^\star\tau = 0\} = \{\tau \in V^\star \mid \forall w \in W \quad \langle w, \iota^\star\tau\rangle = 0\} \\ &= \{\tau \in V^\star \mid \forall w \in W \quad \langle w\iota, \tau\rangle = \langle w, \tau\rangle = 0\} = W^\perp \ .\end{aligned}$$

(c) Theorem 7.2, applied to $\iota^\star$ gives $\mathsf{dim}\, V^\star = \mathsf{dim}\,\mathsf{Im}\,\iota^\star + \mathsf{dim}\,\mathsf{Ker}\,\iota^\star$, which transforms directly into (7.5) by applying the above equality and Proposition 7.2. ■

Morphisms of modules and matrices

For every finite P, we identify the *left* $\mathbb{K}$-module $\mathbb{K}^P$ with the row vectors (over $\mathbb{K}$) of dimension P. A matrix m of dimension $P\times Q$ over $\mathbb{K}$ then realises a morphism (on left $\mathbb{K}$-modules) from $\mathbb{K}^P$ to $\mathbb{K}^Q$ by multiplication *on the right* of the elements of $\mathbb{K}^P$.

This statement has a converse *when* $\mathbb{K}$ *is a (skew) field*. In this case, let V and U be two left vector $\mathbb{K}$-spaces of finite dimensions r and s respectively, and let $\mathcal{B} = \{b^1, b^2, \ldots, b^r\}$ be a basis of V and $\mathcal{C} = \{c^1, c^2, \ldots, c^s\}$ a basis of U. The choice of the basis $\mathcal{B}$ (resp. $\mathcal{C}$) determines a canonical isomorphism between V and $\mathbb{K}^r$ (resp. between U and $\mathbb{K}^s$): every v in V can be written uniquely in the form $v = v_1 b^1 + v_2 b^2 + \cdots + v_r b^r$ and we associate with it the vector $\underline{v} = (v_1, v_2, \ldots, v_r)$.

We also have a natural isomorphism between the set of morphisms from V to U and $\mathbb{K}^{r\times s}$ (these last being viewed as *right* vector spaces). For every morphism $\alpha\colon V \to U$, we write $\underline{\alpha} = \left(\underline{\alpha}^i_j\right)$ for the $r\times s$ matrix whose ith row $\underline{\alpha}^i$ is the (row) vector of the coordinates of $(b^i)\alpha$ in the basis $\mathcal{C}$; that is, the row vector $\underline{(b^i)\alpha}$. It is then immediate to verify

$$\forall v \in V \qquad \underline{(v)\alpha} = \underline{v}\cdot\underline{\alpha} \ . \tag{7.6}$$

If $\tau\colon V \to \mathbb{K}$ is a functional – that is, $\tau \in V^\star$ – we have $\langle b^i, \tau\rangle = \underline{\tau}^i$ and $\underline{\tau}$ is the (column) vector whose entries are the coordinates of τ in the basis $\mathcal{B}^\star = \{\beta_1, \beta_2, \ldots, \beta_r\}$, the dual of $\mathcal{B}$. Equation (7.6) becomes $\langle v, \tau\rangle = \underline{v}\cdot\underline{\tau}$ and we shall naturally say that $\underline{v}$ and $\underline{\tau}$ are *orthogonal* if $\underline{v}\cdot\underline{\tau} = 0$.

Every element of the subspace of $\mathbb{K}^s$ (left vector space) generated by the *rows* of $\underline{\alpha}$ is written

$$v_1\,\underline{\alpha}^1 + v_2\,\underline{\alpha}^2 + \cdots + v_r\,\underline{\alpha}^r = \underline{v}\cdot\underline{\alpha} \ ,$$

and hence corresponds to an element of $\mathsf{Im}\,\alpha$, and conversely. The row rank of $\underline{\alpha}$ is by definition the dimension of $\mathsf{Im}\,\alpha$. We can also characterise $\mathsf{Ker}\,\alpha$ with the help of the matrix $\underline{\alpha}$:

$$\mathsf{Ker}\,\alpha = \{v \in V \mid v\alpha = 0_U\} = \{\underline{v} \in \mathbb{K}^r \mid \underline{v}\cdot\underline{\alpha} = \underline{0_U}\} \ .$$

Thus, $\mathsf{Ker}\,\alpha$ is the set of vectors v such that $\underline{v}$ is orthogonal to each of the columns of $\underline{\alpha}$ and hence $\mathsf{Ker}\,\alpha$ is the *orthogonal* of the (right) subspace F of $\mathbb{K}^s$ generated by the *columns* of $\underline{\alpha}$. By definition, the column rank of $\underline{\alpha}$ is the dimension of F.

By Theorems 7.2 and 7.3 and by Proposition 7.2, we conclude

$$\mathsf{dim}\, V = \mathsf{dim}\,\mathsf{Ker}\,\alpha + \mathsf{dim}\,\mathsf{Im}\,\alpha = \mathsf{dim}\, V^\star = \mathsf{dim}\, F^\perp + \mathsf{dim}\, F \ .$$

As $\mathsf{Ker}\,\alpha = F^{\perp}$, we have $\mathsf{dim}\,\mathsf{Im}\,\alpha = \mathsf{dim}\,F$, so we have proved Theorem 4.7 in the case of matrices of finite dimension.

This proof[69] distinguishes itself from those given in most treatises on algebra which are based on the use of *determinants*, as it does not assume that $\mathbb{K}$ is commutative. The proof in the case of infinite matrices operates directly on the object matrix without requiring the identification between matrix and morphism, which may therefore seem simpler.

Proof of Theorem 4.7. Let α be a matrix of dimension $P \times Q$ (with entries in $\mathbb{K}$), where P and Q are arbitrary sets. For all p in P, rows α^p are elements of $\mathbb{K}^Q$; for all q in Q, columns α_q are elements of $\mathbb{K}^P$.

The row rank of α is finite and hence there exists a finite subset R of P such that, for all p in P, there exists, for each r in R, a coefficient k^p_r in $\mathbb{K}$ such that

$$\alpha^p = \sum_{r \in R} k^p_r \, \alpha^r \quad ,$$

and hence in particular

$$\forall q \in Q \qquad \alpha^p_q = \sum_{r \in R} k^p_r \, \alpha^r_q \quad . \tag{7.7}$$

For every column α_q of α, we write $\overline{\alpha}_q$ for the restriction of α_q to indices that belong to R: α_q is a column vector of dimension R. Let S be a subset of Q such that the set $\{\overline{\alpha}_s\}_{s \in S}$ is not free; that is, there exist coefficients $\{h^s\}_{s \in S}$ such that

$$\sum_{s \in S} \overline{\alpha}_s \, h^s = 0 \quad ,$$

and hence in particular

$$\forall r \in R \qquad \sum_{s \in S} \overline{\alpha}^r_s \, h^s = 0_{\mathbb{K}} \quad ,$$

from which we deduce by (7.7) that

$$\sum_{s \in S} \alpha_s \, h^s = 0 \quad .$$

Thus, the column rank of α is equal to the dimension of the space generated by the vectors $\{\overline{\alpha}_q\}_{q \in Q}$, necessarily less or equal to than the cardinal of R, the row rank of α. By symmetry, the two ranks are equal. ■

[69] Which in substance is that given for example in N. Jacobson, *Lecture in Abstract Algebra*, vol. II [128].

Exercises

7.1 **Modules over a ring.** In this exercise, $\mathbb{K}$ is a ring.

(a) Verify that a $\mathbb{K}$-module is necessarily a (commutative) *group*.

(b) Verify that the equation

$$x \equiv y \quad [\alpha] \quad \Leftrightarrow \quad x - y \in \mathsf{Ker}\,\alpha \tag{$**$}$$

induces a bijection between the congruences of a $\mathbb{K}$-module U and its sub-$\mathbb{K}$-modules.

(c) Verify that if $\alpha\colon U \to W$ is a morphism of $\mathbb{K}$-modules, then $\mathsf{Im}\,\alpha$ is isomorphic to $U/\mathsf{Ker}\,\alpha$.

• 7.2 **Dual vector space.** In this exercise, $\mathbb{K}$ is a skew field and V a left $\mathbb{K}$-vector space.

(a) Verify that the set of $\mathbb{K}$-functionals on V is a right $\mathbb{K}$-vector space.

(b) Verify that there exists a canonical *injection* from V to $V^{\star\star}$, the dual of the dual of V.

• 7.3 **Morphisms and matrices.** In this exercise, $\mathbb{K}$ is a skew field, V and U are two left $\mathbb{K}$-vector spaces and $\alpha\colon V \to U$ is a morphism.

(a) Verify that, for all v in V, $\underline{(v)\alpha} = \underline{v} \cdot \underline{\alpha}$ (Equation (7.6)).

(b) Let $\alpha^\star\colon U^\star \to V^\star$ be the dual morphism of α. Verify that the matrix of $\alpha^\star$ with bases $\mathcal{C}^\star$ and $\mathcal{B}^\star$ is equal to $\underline{\alpha}$, the matrix of α with bases $\mathcal{B}$ and $\mathcal{C}$.

SOLUTIONS TO THE EXERCISES

Skimming through the geometry textbooks,
he wondered if it was really worth knowing how to read.
Luis Sepulveda

1. Formal power series on a graded monoid

1.1 *Verify that the Cauchy product indeed makes $\mathbb{K}\langle M\rangle$ a semiring, and $\mathbb{K}\langle\langle M\rangle\rangle$ too, if $\mathbb{K}$ is complete.*

Solution: The Cauchy product is defined in both cases, but we have not verified that it is indeed distributive over addition. Let m be in M and s', s'' and t in $\mathbb{K}\langle M\rangle$ (or $\mathbb{K}\langle\langle M\rangle\rangle$). Distributivity can be written on the one hand as

$$\begin{aligned}\langle (s'+s'')t, m\rangle &= \sum_{\substack{u,v\in M\\ uv=m}} \langle (s'+s''), u\rangle\langle t, v\rangle = \sum_{\substack{u,v\in M\\ uv=m}} (\langle s', u\rangle + \langle s'', u\rangle)\langle t, v\rangle \\ &= \sum_{\substack{u,v\in M\\ uv=m}} [\langle s', u\rangle\langle t, v\rangle + \langle s'', u\rangle\langle t, v\rangle] \ , \qquad (*)\end{aligned}$$

and on the other,

$$\begin{aligned}\langle s't + s''t, m\rangle &= \langle s't, m\rangle + \langle s''t, m\rangle \\ &= \sum_{\substack{u,v\in M\\ uv=m}} \langle s', u\rangle\langle t, v\rangle + \sum_{\substack{u,v\in M\\ uv=m}} \langle s'', u\rangle\langle t, v\rangle \qquad (**)\end{aligned}$$

The right hand sides of $(*)$ and $(**)$ are equal by associativity.

1.2 *Recognise the semirings $\mathbb{Z}\langle\mathbb{N}\rangle$, $\mathbb{Q}\langle\langle\mathbb{N}\rangle\rangle$, $\mathbb{R}\langle\langle\mathbb{N}^2\rangle\rangle$.*

Solution: Everything becomes clear if we identify $\mathbb{N}$ with the free monoid with one generator $\{x\}^*$, and hence n with x^n, and $\mathbb{N}^2$ with the commutative monoid with two generators, and hence (m, n) with $x^m y^n$:

$\mathbb{Z}\langle\mathbb{N}\rangle$ is the ring of (classical) polynomials in one variable with integer coefficients.

$\mathbb{Q}\langle\langle\mathbb{N}\rangle\rangle$ is the rings of (classical) series in one variable with rational coefficients.

$\mathbb{R}\langle\langle\mathbb{N}^2\rangle\rangle$ is the ring of (classical) series in two (commutative) variables with real coefficients.

1.3 (a) *Verify that the Cauchy product is not defined on $\mathbb{Q}\langle\langle\mathbb{Z}\rangle\rangle$.*
(b) *Recognise $\mathbb{Q}_{\mathsf{wo}}\langle\langle\mathbb{Z}\rangle\rangle$, the subset of $\mathbb{Q}\langle\langle\mathbb{Z}\rangle\rangle$ consisting of the power series whose support has a least element.*

Solution: (a) Let s be the characteristic series of $\mathbb{Z}$, an element of $\mathbb{N}\langle\langle\mathbb{Z}\rangle\rangle$, hence of $\mathbb{Q}\langle\langle\mathbb{Z}\rangle\rangle$. The square of s is not defined since, for each integer k, we have

$$\langle s^2, x^k\rangle = \sum_{z\in\mathbb{Z}} \langle s, x^{k-z}\rangle\langle s, x^z\rangle = \sum_{z\in\mathbb{Z}} 1$$

which is undefined.

(b) An element of $\mathbb{Q}_{\mathsf{wo}}\langle\langle\mathbb{Z}\rangle\rangle$ can be written in the form $s = \sum_{z\in\mathbb{Z}, z\geqslant k}^{+\infty} q_z\, x^z$ This is what is called a *Laurent series*. We see that $\mathbb{Q}_{\mathsf{wo}}\langle\langle\mathbb{Z}\rangle\rangle$ is a semiring since, given two series s and t of $\mathbb{Q}_{\mathsf{wo}}\langle\langle\mathbb{Z}\rangle\rangle$, for each k in $\mathbb{Z}$, the coefficient of x^k in $s\,t$:

$$\langle s\,t, x^k\rangle = \sum_{z\in\mathbb{Z}} \langle s, x^{k-z}\rangle\langle t, x^z\rangle \ ,$$

is a sum with only a finite number of non-zero terms, and thus well-defined.

1.6 (a) *Show that a graded monoid has no idempotents other than* 1_M.

(b) *Give an example of a graded monoid that is not finitely decomposable.*

Solution: (a) If e is idempotent, we should have $\varphi(e) = \varphi(e^2) = \varphi(e) + \varphi(e)$, which contradicts the assumption that $\varphi(e)$ is strictly positive if e is not the identity.

(b) By Corollary 1.2, such a monoid must have an infinite number of generators. Let $A = \{a_n \mid n \in \mathbb{N}\}$ be an infinite alphabet. The quotient of A^* by the relations $a_i\, a_j = (a_0)^2$ is a monoid equipped with a length function of which no element of length greater than 1 is finitely decomposable.

1.8 *Let A be a finite alphabet.*

(a) *Show that a monoid M presented by $M = \langle\, A\,;\, \{u_i = v_i \mid i \in I\}\,\rangle$ with $|u_i| = |v_i|$ for all i in I is a graded monoid.*

(b) *Show that a monoid M presented by $M = \langle\, A\,;\, u = 1_{A^*}\,\rangle$ is not a graded monoid, whatever the word u is.*

(c) *Give an example of a graded monoid M that is presented by $M = \langle\, A\,;\, u = v\,\rangle$ with $|u| > |v| > 0$.*

(d) *Give an example of a non-graded monoid M that is presented by $M = \langle\, A\,;\, u = v\,\rangle$ with $|u| > |v| > 0$.*

Solution: (a) M is composed of classes of words which are all of the same length (in A^*). We can take as the length (in M) of an element of M the common length of words in A^* that it represents.

(b) In full rigour, the proposition is false:[70] if u is a letter of A, M is isomorphic to $(A \setminus \{a\})^*$ (and hence equipped with a length).

If u is of length greater than or equal to 2, it can be written $u = a\,u'$ with a in A and where a is a divisor of the identity in M.

(c) We take $A = \{a, b\}$, $u = a\,a$, and $v = b$. Every element m in M has a unique representative in A^* which is a word without any factor equal to a^2. The length in M is defined by $\varphi(a) = 1$ and $\varphi(b) = 2$.

(d) We take $A = \{a, b\}$, $u = a\,b\,a$, and $v = a$. The representative of $a\,b$ is idempotent. Without even examining the structure of M, this is enough to conclude that either $a\,b$ is not the identity of M and M is not graded (*cf.* above); or $a\,b$ is the identity, and a and b are divisors of the identity.

1.15 *Explain why the definition of the distance $\mathbf{d}'$ in (1.5) cannot be replaced by*

$$\mathbf{f}(s,t) = \big(\mathsf{max}\{\mathbf{c}(\langle s,m\rangle, \langle t,m\rangle) \mid |m| = \mathbf{e}(s,t)\}\big)\, 2^{-\,\mathbf{e}(s,t)} \quad .$$

Solution: The function defined in this way is not a distance since it does not satisfy the triangular inequality in $\mathbb{Q}\langle\langle a^*\rangle\rangle$:

$$\mathbf{f}\left(0, \frac{1}{8}\right) = \mathbf{f}\left(0, \frac{1}{8} + a\right) = \frac{1}{8} \qquad \text{and} \qquad \mathbf{f}\left(\frac{1}{8}, \frac{1}{8} + a\right) = \frac{1}{2} > \frac{1}{8} + \frac{1}{8} \quad .$$

1.16 ***Simple convergence.*** *Verify that the distance defined on $\mathbb{K}\langle\langle M\rangle\rangle$ is indeed such that a sequence $\{s_n\}_{n\in\mathbb{N}}$ of series of $\mathbb{K}\langle\langle M\rangle\rangle$ converge to s if and only if for all m in M, $\langle s_n, m\rangle$ tends to $\langle s, m\rangle$ (Property 1.2).*

[70] After discovering this, I thought it better from a pedagogical point of view to leave such an error in the text: the reader should always be vigilant. *Note added in translation:* During the translation of the book from French into English, many mistakes were discovered and I now think it was not necessary to leave mistakes on purpose.

Solution: If $\{s_n\}_{n\in\mathbb{N}}$ converges to s – that is, if

$$\forall\varepsilon>0\,,\ \exists N\in\mathbb{N}\,,\ \forall n\geqslant N \qquad \mathbf{d}(s_n,s)<\varepsilon$$

with $\mathbf{d}(s_n,s)=\sum_{i\in\mathbb{N}}\left(\frac{1}{2^i}\max_{|m|=i}\mathbf{c}(\langle s_n,m\rangle,\langle s,m\rangle)\right)$, then – for every m in M,

$$\mathbf{c}(\langle s_n,m\rangle,\langle s,m\rangle)\leqslant 2^{|m|}\,\mathbf{d}(s_n,s)$$

and $\langle s_n,m\rangle$ tends to $\langle s,m\rangle$ for all m in M.
Conversely, if $\langle s_n,m\rangle$ tends to $\langle s,m\rangle$ for all m in M, we can write

$$\forall\varepsilon>0\,,\ \forall m\in M\,,\ \exists N=N(\varepsilon,m)\in\mathbb{N}\,,\ \forall n\geqslant N \qquad \mathbf{c}(\langle s_n,m\rangle,\langle s,m\rangle)<\varepsilon\ .$$

It is important to note that in this equation N depends on m and not just on ε. We therefore fix ε and let l be an integer such that $\frac{1}{2^l}\leqslant\varepsilon$. Let $T=T(\varepsilon)$ be the integer defined by $T=\max_{|m|\leqslant l}N(\varepsilon,m)$. This upper bound exists, and is *finite*, since there is only a finite number of elements m whose length is less than or equal to l. We then have, for all $n\geqslant T$,

$$\begin{aligned}\mathbf{d}(s_n,s)&=\sum_{i=0}^{i=l}\frac{1}{2^i}\left[\max_{|m|=i}\mathbf{c}(\langle s_n,m\rangle,\langle s,m\rangle)\right]+\sum_{i>l}\frac{1}{2^i}\left[\max_{|m|=i}\mathbf{c}(\langle s_n,m\rangle,\langle s,m\rangle)\right]\\&\leqslant\sum_{i=0}^{i=l}\left(\frac{1}{2^i}\varepsilon\right)+\sum_{i>l}\left(\frac{1}{2^i}1\right)\leqslant 3\,\varepsilon\end{aligned}$$

and $\{s_n\}_{n\in\mathbb{N}}$ converges to s.

1.20 ***Uniform convergence.***

(a) *Show that if the topology of $\mathbb{K}$ is discrete, the uniform limit of a sequence of polynomials is a polynomial.*

(b) *Characterise the series of $\mathbb{Q}\langle\langle M\rangle\rangle$ that are the uniform limit of polynomials of $\mathbb{Q}\langle M\rangle$.*

Solution: (a) Let $\{p_n\}_{n\in\mathbb{N}}$ be a sequence of polynomials of $\mathbb{K}\langle M\rangle$ that converge uniformly to s; that is,

$$\forall\varepsilon>0\,,\ \exists N=N(\varepsilon)\in\mathbb{N}\,,\ \forall m\in M\,,\ \forall n\geqslant N \qquad \mathbf{c}(\langle p_n,m\rangle,\langle s,m\rangle)<\varepsilon\ . \qquad (\mathsf{SE}.1.1)$$

Take ε strictly less than 1; we obtain

$$\exists N\in\mathbb{N}\,,\ \forall m\in M\,,\ \forall n\geqslant N \qquad \mathbf{c}(\langle p_n,m\rangle,\langle s,m\rangle)<1 \qquad \text{hence} \qquad \langle p_n,m\rangle=\langle s,m\rangle$$

and thus, there exists an integer N such that for all n greater than N, $p_n=s$ holds: the sequence $\{p_n\}$ is stationary from rank N on.

(b) By ($\mathsf{SE}.1.1$), we deduce that s is the uniform limit of the p_n in $\mathbb{Q}\langle M\rangle$ if and only if the coefficients $\langle s,m\rangle$ tend to 0 when the length of m tends to infinity:

$$\forall\varepsilon>0\,,\ \exists l\in\mathbb{N}\,,\ \forall m\in M \qquad |m|\geqslant l \implies |\langle s,m\rangle|<\varepsilon\ .$$

2. $\mathbb{K}$-automata and $\mathbb{K}$-rational power series

2.1 *Let s and t be two proper series. Compute the star of the matrix $m=\begin{pmatrix}0&s\\t&0\end{pmatrix}$.*

cf. Prop. 2.5, p. 396.

Solution: Since m is proper, we can apply $\mathbf{Z}_2$: $m^*=(1+m)(m^2)^*$ from which

$$\begin{pmatrix}0&s\\t&0\end{pmatrix}^*=\begin{pmatrix}1&s\\t&1\end{pmatrix}\cdot\begin{pmatrix}s\,t&0\\0&t\,s\end{pmatrix}^*=\begin{pmatrix}1&s\\t&1\end{pmatrix}\cdot\begin{pmatrix}(s\,t)^*&0\\0&(t\,s)^*\end{pmatrix}=\begin{pmatrix}(s\,t)^*&s\,(t\,s)^*\\t\,(s\,t)^*&(t\,s)^*\end{pmatrix}\ .$$

$\boxed{\mathbf{2.2}}$ *Compute* $(a^*)^2$ *and* $(a^*)^3$ *in* $\mathbb{N}\langle\langle a^*\rangle\rangle$.

Solution: By definition, $\forall n \in \mathbb{N} \quad \langle a^*, a^n\rangle = 1$ and hence

$$\forall n \in \mathbb{N} \qquad \langle (a^*)^2, a^n\rangle = \sum_{\substack{p,q\in\mathbb{N}\\ p+q=n}} \langle a^*, a^p\rangle\langle a^*, a^q\rangle = \sum_{\substack{p,q\in\mathbb{N}\\ p+q=n}} (1) = n+1 \ .$$

Then $(a^*)^2 = \sum_{n\in\mathbb{N}} (n+1)\, a^n$. Likewise

$$\begin{aligned}\forall n \in \mathbb{N} \qquad \langle (a^*)^3, a^n\rangle &= \sum_{\substack{p,q\in\mathbb{N}\\ p+q=n}} \langle (a^*)^2, a^p\rangle\langle a^*, a^q\rangle = \sum_{\substack{p,q\in\mathbb{N}\\ p+q=n}} (p+1)\\ &= \sum_{p=0}^{p=n} (p+1) = \frac{(n+1)(n+2)}{2} \ .\end{aligned}$$

Hence $(a^*)^3 = \sum_{n\in\mathbb{N}} \frac{(n+1)(n+2)}{2}\, a^n$.

$\boxed{\mathbf{2.3}}$ *Let* p *be an integer; we define* $a^{(p)}$ *by induction:* $a^{(0)} = a$ *and* $a^{(p+1)} = \left(a^{(p)}\right)^+$. *Show that* $a^{(p)} = (p\,a)^* a$.

Solution: The relation is true for $p = 0$. We then compute[71]

$$\left((p\,a)^* a\right)^+ = \underbrace{\left((p\,a)^* a\right)^* (p\,a)^*}_{\text{by } (\mathbf{S})} a = (p\,a + a)^* a = \left((p+1)\,a\right)^* a \ .$$

$\boxed{\mathbf{2.6}}$ *Suppose that* $\mathbb{K}$ *is a ring. Show that a series of* $\mathbb{K}\langle\langle M\rangle\rangle$ *is invertible if and only if its constant term is invertible.*

Solution: The condition is necessary. Let t be the inverse of s; that is, $s\,t = t\,s = 1_{\mathbb{K}}$. Since M is graded, $\langle s\,t, 1_M\rangle = \langle s, 1_M\rangle\langle t, 1_M\rangle = 1_{\mathbb{K}}$ and $\langle t\,s, 1_M\rangle = \langle t, 1_M\rangle\langle s, 1_M\rangle = 1_{\mathbb{K}}$, and hence $s_0 = \langle s, 1_M\rangle$ is invertible.

Conversely, let s_{p} be the proper part of s and t_0 the inverse of the constant term s_0 of s. We can write $s = s_0\,[1_M - (-t_0\, s_{\mathsf{p}})]$. The star of $(-t_0\, s_{\mathsf{p}})$, a proper series, is defined, and $t = [-t_0\, s_{\mathsf{p}}]^*\, t_0$ is the inverse of s.

$\boxed{\mathbf{2.9}}$ *Compute the star of the matrix* $m = \begin{pmatrix} a & b & c\\ d & e & f\\ g & h & j\end{pmatrix}$.

Solution: This is an exercise in pure calculation, which should convince the reader that the induction in the proof of Proposition 2.10 is unusable in practice. We decompose m into blocks,

$$m = \begin{pmatrix} A & B\\ C & D\end{pmatrix} \qquad \text{with} \qquad A = \begin{pmatrix} a & b\\ d & e\end{pmatrix}, \quad B = \begin{pmatrix} c\\ f\end{pmatrix}, \quad C = \begin{pmatrix} g & h\end{pmatrix}, \quad \text{and} \quad D = j \ .$$

By (2.13), $m^* = \begin{pmatrix} (A+BD^*C)^* & A^*B(D+CA^*B)^*\\ D^*C(A+BD^*C)^* & (D+CA^*B)^*\end{pmatrix}$. First we calculate $(A+BD^*C)^*$:

$$(A+BD^*C)^* = \left(\begin{pmatrix} a & b\\ d & e\end{pmatrix} + \begin{pmatrix} c\\ f\end{pmatrix} j^* \begin{pmatrix} g & h\end{pmatrix}\right)^* = \left(\begin{pmatrix} a + c\,j^*g & b + c\,j^*h\\ d + f\,j^*g & e + f\,j^*h\end{pmatrix}\right)^* \ .$$

[71] This is an example of the application of rational identities given by Eilenberg, *op. cit.* (p. 168).

The entries (we can only write one per line!) of $(A+BD^*C)^*$, listing the entries in row order, from left to right, are thus

$$\begin{gathered}\big(a+cj^*g+(b+cj^*h)(e+fj^*h)^*(d+fj^*g)\big)^* \ , \\ (a+cj^*g)^*(b+cj^*h)\big(e+fj^*h+(d+fj^*g)(a+cj^*g)^*(b+cj^*h)\big)^* \ , \\ (e+fj^*h)^*(d+fj^*g)\big(a+cj^*g+(b+cj^*h)(e+fj^*h)^*(d+fj^*g)\big)^* \ , \\ \big(e+fj^*h+(d+fj^*g)(a+cj^*g)^*(b+cj^*h)\big)^* \ .\end{gathered}$$

Equation (2.13) gives $A^* = \begin{pmatrix} (a+be^*d)^* & a^*b(e+da^*b)^* \\ e^*d(a+be^*d)^* & (e+da^*b)^* \end{pmatrix}$ whence

$$(D+CA^*B)^* = \big(j+g(a+be^*d)^*c+ga^*b(e+da^*b)^*f+he^*d(a+be^*d)^*c+h(e+da^*b)^*f\big)^*$$

and we leave to the obsessive reader the drudgery of finishing this tedious calculation.

2.10 *Compute the star of the matrix* $\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}$.

Solution: It's a trap! Obviously we should not use the general method of the previous exercise. We write

$$\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ 0 & d & 0 \\ 0 & 0 & f \end{pmatrix} + \begin{pmatrix} 0 & b & c \\ 0 & 0 & e \\ 0 & 0 & 0 \end{pmatrix} \ ,$$

and use the identity $(X+Y)^* = (X^*Y)^*X^*$. We obtain

$$\begin{aligned}\begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}^* &= \left[\begin{pmatrix} a^* & 0 & 0 \\ 0 & d^* & 0 \\ 0 & 0 & f^* \end{pmatrix}\begin{pmatrix} 0 & b & c \\ 0 & 0 & e \\ 0 & 0 & 0 \end{pmatrix}\right]^* \begin{pmatrix} a^* & 0 & 0 \\ 0 & d^* & 0 \\ 0 & 0 & f^* \end{pmatrix} \\ &= \begin{pmatrix} 0 & a^*b & a^*c \\ 0 & 0 & d^*e \\ 0 & 0 & 0 \end{pmatrix}^* \begin{pmatrix} a^* & 0 & 0 \\ 0 & d^* & 0 \\ 0 & 0 & f^* \end{pmatrix} \\ &= \begin{pmatrix} 1 & a^*b & a^*c+a^*bd^*e \\ 0 & 1 & d^*e \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} a^* & 0 & 0 \\ 0 & d^* & 0 \\ 0 & 0 & f^* \end{pmatrix} \\ &= \begin{pmatrix} a^* & a^*bd^* & a^*cf^*+a^*bd^*ef^* \\ 0 & d^* & d^*ef^* \\ 0 & 0 & f^* \end{pmatrix} \ ,\end{aligned}$$

since, setting $m = \begin{pmatrix} 0 & a^*b & a^*c \\ 0 & 0 & d^*e \\ 0 & 0 & 0 \end{pmatrix}$, we have $m^3 = 0$ and hence $m^* = m^0 + m^1 + m^2$.

2.12 (a) *Let E be a matrix with coefficients in $\mathcal{M}$. What do the coefficients of the matrix E^* represent?*

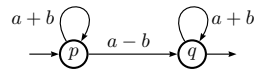
(b) *Answer the same question for a matrix E with coefficients in* $\mathbb{P} = \langle \mathbb{N} \cup \{+\infty\}, \mathsf{max}, \mathsf{min} \rangle$.

Solution: (a) E^* is the matrix of the distances of the shortest paths between the states of $\mathcal{A}$, if we view E as the matrix of distances between states.

(b) If we view the coefficients as 'capacities', and if the capacity of a path is the minimum capacity of the transitions it contains, then the coefficient (p,q) of E^* is the maximum capacity of a path which joins p to q.[72]

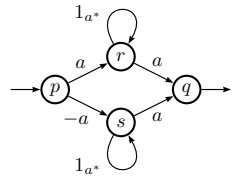
[72] To conclude this exercise, we could say that $\mathcal{M}$ is the semiring of hauliers, and $\mathbb{P}$ that of plumbers.

2.15 *Compute the behaviour of the automaton $\mathcal{D}_2$ in the adjacent figure, when $\mathcal{D}_2$ is*
(a) *considered as a $\mathbb{Z}$-automaton over $\{a,b\}^*$;*
(b) *considered as a $\mathbb{Z}$-automaton over $\mathbb{N}^2$.*



Solution: (a) $|\mathcal{D}_2| = \sum_{f\in A^*} (|f|_a - |f|_b)\, f$; (b) $|\mathcal{D}_2| = \sum_{a^n b^m \in \mathbb{N}^2} (n-m)\, \mathrm{C}_{n+m}^n\, a^n b^m$.

2.16 *Give an example of a $\mathbb{Z}$-automaton over $\{a\}^*$, $\mathcal{A} = \langle I, E, T\rangle$, whose behaviour is undefined and such that the family $\{I\cdot E^n\cdot T\}_{n\in\mathbb{N}}$ is nonetheless summable.*



Solution: In the automaton adjacent, $I\cdot E^n\cdot T = 0$ for all n, yet E^* is undefined since $(E^{<n})_{r,r} = (E^{<n})_{s,s} = n$.

2.17 (a) *Prove (2.15) (Example 2.5); that is, show that*

$$\forall f \in A^* \qquad \langle |\mathcal{C}_1|, f\rangle = \overline{f} \ . \qquad (2.15)$$

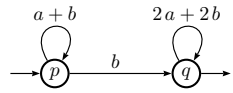


Fig. SE.2.1: The $\mathbb{N}$-automaton $\mathcal{C}_1$

(b) *Show that for all n, there exists an $\mathbb{N}$-automaton $\mathcal{V}'_n$ with $n+1$ states which, for every f in A^*, 'computes' the nth power of the number represented by f: that is, $\langle |\mathcal{V}'_n|, f\rangle = \overline{f}^{\,n}$.*

Solution: (a) By induction on the length of f. The property is true for $|f| = 0$: that is, $\langle |\mathcal{C}_1|, 1_{A^*}\rangle = 0$. The induction can be performed in two different ways.

(i) First method: every word f of length greater than or equal to to 1 can be written $f = x\,g$ with x in A and g in A^*. If $f = a\,g$, then $\overline{f} = \overline{g}$. Now, every successful path of $\mathcal{C}_1$ with label f is of the form $p \xrightarrow{a} p \xrightarrow{g} q$, and thus $\langle |\mathcal{C}_1|, f\rangle = \langle |\mathcal{C}_1|, g\rangle$. If $f = b\,g$, then $\overline{f} = 2^{|g|} + \overline{g}$. A successful path in $\mathcal{C}_1$ with label f is either the path $p \xrightarrow{b} q \xrightarrow{g} q$, whose coefficient is $2^{|g|}$, or in the form $p \xrightarrow{b} p \xrightarrow{g} q$. Thus $\langle |\mathcal{C}_1|, f\rangle = 2^{|g|} + \langle |\mathcal{C}_1|, g\rangle$. In both cases, $\langle |\mathcal{C}_1|, f\rangle = \overline{f}$.

(ii) Second method: every word f of length greater than or equal to 1 can be written $f = g\,x$ with x in A and g in A^*. If $f = g\,a$, then $\overline{f} = 2\,\overline{g}$. Now, every successful path in $\mathcal{C}_1$ with label f is of the form $p \xrightarrow{g} q \xrightarrow{a} p$, and thus $\langle |\mathcal{C}_1|, f\rangle = \langle |\mathcal{C}_1|, g\rangle \times 2$. If $f = g\,b$, then $\overline{f} = 2\,\overline{g} + 1$. A successful path in $\mathcal{C}_1$ with label f is either the path $p \xrightarrow{g} p \xrightarrow{b} q$, whose coefficient is 1, or of the form $p \xrightarrow{g} q \xrightarrow{b} q$. Thus $\langle |\mathcal{C}_1|, f\rangle = 1 + 2\langle |\mathcal{C}_1|, g\rangle$. In both cases, $\langle |\mathcal{C}_1|, f\rangle = \overline{f}$.

(b) We perform the computation using just the second method. The first would be equally possible; it leads to slightly more complex formulas and to automata which we shall see later, via another procedure. *cf. Exer. 3.12, p. 437*

For every integer n, we have

$$\begin{array}{lll} \text{if} \quad f = g\,a \ , & \overline{f}^{\,n} = \overline{g\,a}^{\,n} = (2\,\overline{g})^n = 2^n\,\overline{g}^{\,n} \ , \\ \text{if} \quad f = g\,b \ , & \overline{f}^{\,n} = \overline{g\,b}^{\,n} = (2\,\overline{g}+1)^n = \sum_{i=0}^{i=n} 2^i\,\mathrm{C}_n^i\,\overline{g}^{\,i} \ . \end{array} \qquad (\mathsf{SE}.2.1)$$

Hence the construction of an $\mathbb{N}$-automaton $\mathcal{V}'$: the set of states is $\mathbb{N}$ ($\mathcal{V}'$ is hence infinite), the initial state is $\{0\}$ and the transitions are defined by

$$\forall i, j \in \mathbb{N} \qquad i \xrightarrow{2^i\,a} i \qquad \text{and} \qquad i \xrightarrow{2^i\,\mathrm{C}_j^i\,b} j \ .$$

The formulas (SE.2.1) then prove, by induction on the length of f, that the coefficient of a computation with source 0, destination n and label f is $\overline{f}^{\,n}$. The (finite) automaton $\mathcal{V}'_n$ obtained by keeping the first $n+1$ states of $\mathcal{V}'$ and taking n as the unique terminal state answers the question (*cf.* Figure SE.2.2).

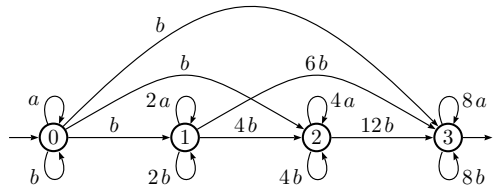


Figure SE.2.2: The 'restriction' of $\mathcal{V}'$ to $\mathcal{V}'_3$

2.18 *Show that every automaton over A with spontaneous transitions is equivalent to an automaton without spontaneous transitions.*[73]

Solution: We write $\mathcal{A} = \langle I, E, T \rangle$ where E is a matrix whose entries are finite unions of letters of A, and possibly of 1_{A^*}. We write $E = Z + G$ where the entries of Z are either 0 or 1_{A^*}, and where those of G are finite unions of letters. We can consider Z to be a boolean matrix, and Z^* is defined. We then have

$$|\mathcal{A}| = I \cdot E^* \cdot T = I \cdot (Z + G)^* \cdot T = I \cdot (Z^* \cdot G)^* \cdot Z^* \cdot T \ .$$

The automaton $\mathcal{A}' = \langle I, Z^* \cdot G, Z^* \cdot T \rangle$ is, like $\mathcal{A}$, an automaton over A without spontaneous transitions and is equivalent to $\mathcal{A}$.

2.20 *Show that the behaviour of the $\mathcal{M}$-automaton $\mathcal{S}_2$ is a series whose support is all of $\{a, b\}^*$ and whose coefficients are less than or equal to 4.*[74]

Solution: The support of $|\mathcal{S}_2|$ is the whole of $\{a, b\}^*$, because of the loop at p. The rest is more subtle.

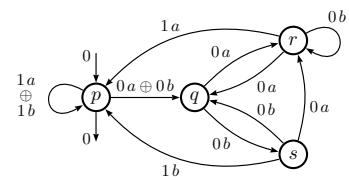
We recognise in $\mathcal{S}_2$ the *forward* closure of the $\mathcal{M}$-automaton $\mathcal{S}_3$ whose form itself shows not only that the support of the $\mathcal{M}$-series that it recognises is the *star* of the language S_3: that is,

$$S_3 = (a + b)(b\,b + a\,b^*a + b\,a\,b^*a)^* \ ,$$

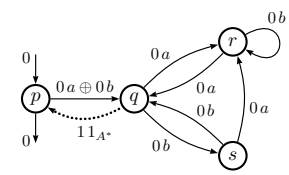
but also that the coefficient of each word f in ${S_3}^*$ gives the minimum number of times the spontaneous transition

$$q \xrightarrow[\mathcal{S}_3]{1\,1_{A^*}} p \ ,$$

must be followed to accept f; that is, the smallest n such that $f \in {S_3}^n$. A case-by-case examination then shows that, whatever f is, this smallest n is at most 4; that is, ${S_3}^* = {S_3}^{\leqslant 4}$.



$\mathcal{M}$-automaton $\mathcal{S}_2$



$\mathcal{M}$-automaton $\mathcal{S}_3$

[73]This exercise anticipates Proposition 2.14, of which it states a special case.

[74]This exercise and its solution, whose material was borrowed from Imre Simon [236], enables me to mention a very beautiful problem in automata theory. We say that a language L has the *finite power property* if there exists an integer n such that $L^* = L^n$. The problem, posed in 1966 by John Brzozowski, is to know if it is decidable whether a rational language has the finite power property. It was answered positively in 1978 by Kosaburo Hashiguchi [113] and independently by Simon himself [235]. This problem is obviously related to that of star height, and Hashiguchi's solution to that problem, mentioned on p. 216, is based on the study of $\mathcal{M}$-automata. For an accessible presentation of these problems and their relationships, as well as comprehensive references, see [236].

2.23 *Using the constructions from the proof of Theorem 2.2, give a $\mathbb{Q}$-automaton over $\{a,b\}^*$ whose behaviour is equal to* $\mathsf{E}_1 = (\frac{1}{6}\,a^* + \frac{1}{3}\,b^*)^*$.

Solution: The *trim* automaton for $\frac{1}{6}\,a^*$ is shown in Figure SE.2.3 on the left, from which we obtain the automaton for E_1, on the right.

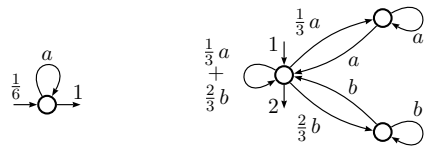


Figure SE.2.3: $\mathbb{Q}$-automata for $\frac{1}{6}\,a^*$ and E_1

2.24 ***Standard $\mathbb{K}$-automata.***[75]

(a) *Write out the calculation that gives Proposition 2.17.*

(b) *Give matrix formulas equivalent to (2.16), (2.20), (2.21), and (2.24) which, applied to standard $\mathbb{K}$-automata, give standard $\mathbb{K}$-automata whose behaviours are respectively the product by a constant and the sum, product and star of the behaviours of the initial $\mathbb{K}$-automata. Deduce Proposition 2.18.*

(c) *Apply these formulas to construct a standard $\mathbb{Q}$-automaton whose behaviour is equal to* $\mathsf{E}_1 = \big((\frac{1}{6}\,a^*) + (\frac{1}{3}\,b^*)\big)^*$.

Solution: (a) Let $\mathcal{A} = \langle\, I, E, T \,\rangle$ be an arbitrary automaton (whose behaviour is defined). Set $I = X + K$, $T = Y + V$ and $E = W + G$ where X, Y and W are scalar and the coefficients of K, V and G are proper. The standard (!) computation of forward closure gives the proper standard automaton equivalent to $\mathcal{A}$:

$$\left\langle \begin{pmatrix} 1 & \boxed{0} & 0 \end{pmatrix}, \begin{pmatrix} 0 & \boxed{B} & C \\ \boxed{0} & \boxed{H} & \boxed{D} \\ 0 & \boxed{0} & 0 \end{pmatrix}, \begin{pmatrix} R \\ \boxed{S} \\ 1 \end{pmatrix} \right\rangle .$$

where $B = J + X \cdot W^* \cdot G$, $C = X \cdot W^* \cdot U$, $H = W^* \cdot G$, $D = W^* \cdot U$, $R = X \cdot W^* \cdot Y$ and $S = W^* \cdot Y$.

(b) Let $\mathcal{A}$ and $\mathcal{A}'$ be two proper standard automata:

$$\mathcal{A} = \left\langle \begin{pmatrix} 1 & \boxed{0} \end{pmatrix}, \begin{pmatrix} 0 & \boxed{J} \\ \boxed{0} & \boxed{F} \end{pmatrix}, \begin{pmatrix} c \\ \boxed{U} \end{pmatrix} \right\rangle$$

$$\text{and} \qquad \mathcal{A}' = \left\langle \begin{pmatrix} 1 & \boxed{0} \end{pmatrix}, \begin{pmatrix} 0 & \boxed{J'} \\ \boxed{0} & \boxed{F'} \end{pmatrix}, \begin{pmatrix} c' \\ \boxed{U'} \end{pmatrix} \right\rangle .$$

The constructions on the proper standard automata of Section I.5.1, adapted to $\mathbb{K}$-automata, give the following matrix formulas. The series $k\,|\mathcal{A}|$ and $|\mathcal{A}|\,k$ (with k in $\mathbb{K}$) are realised respectively by *cf. p. 145*

$$\left\langle \begin{pmatrix} 1 & \boxed{0} \end{pmatrix}, \begin{pmatrix} 0 & \boxed{k\,J} \\ \boxed{0} & \boxed{F} \end{pmatrix}, \begin{pmatrix} k\,c \\ \boxed{U} \end{pmatrix} \right\rangle \quad \text{and} \quad \left\langle \begin{pmatrix} 1 & \boxed{0} \end{pmatrix}, \begin{pmatrix} 0 & \boxed{J} \\ \boxed{0} & \boxed{F} \end{pmatrix}, \begin{pmatrix} c\,k \\ \boxed{U\,k} \end{pmatrix} \right\rangle ;$$

[75] Adapted from [49].

the series $|\mathcal{A}| + |\mathcal{A}'|$ is realised by

$$\left\langle \begin{pmatrix} 1 & \boxed{0} & \boxed{0} \end{pmatrix}, \begin{pmatrix} 0 & \boxed{J} & \boxed{J'} \\ \boxed{0} & \boxed{F} & \boxed{0} \\ \boxed{0} & \boxed{0} & \boxed{F'} \end{pmatrix}, \begin{pmatrix} c+c' \\ \boxed{U} \\ \boxed{U'} \end{pmatrix} \right\rangle ;$$

the series $|\mathcal{A}|\,|\mathcal{A}'|$ is realised by

$$\left\langle \begin{pmatrix} 1 & \boxed{0} & \boxed{0} \end{pmatrix}, \begin{pmatrix} 0 & \boxed{J} & \boxed{c\,J'} \\ \boxed{0} & \boxed{F} & \boxed{U \cdot J'} \\ \boxed{0} & \boxed{0} & \boxed{F'} \end{pmatrix}, \begin{pmatrix} c\,c' \\ \boxed{U\,c'} \\ \boxed{U'} \end{pmatrix} \right\rangle ,$$

and the series $|\mathcal{A}|^*$, which is defined if and only if c^* is defined, is realised by

$$\left\langle \begin{pmatrix} 1 & \boxed{0} \end{pmatrix}, \begin{pmatrix} 0 & \boxed{c^*J} \\ \boxed{0} & \boxed{G} \end{pmatrix}, \begin{pmatrix} c^* \\ \boxed{U\,c^*} \end{pmatrix} \right\rangle$$

(with $G = U \cdot c^* J + F$).

As the standard $\mathbb{K}$-automata for $0_{\mathbb{K}}$, $1_{\mathbb{K}}$ and m are of dimension 1, 1 and 2, the above constructions prove by induction that the dimension of $\mathcal{S}_{\mathsf{E}}$ is $\ell(\mathsf{E}) + 1$.

(c) *Cf.* Figure SE.2.4.

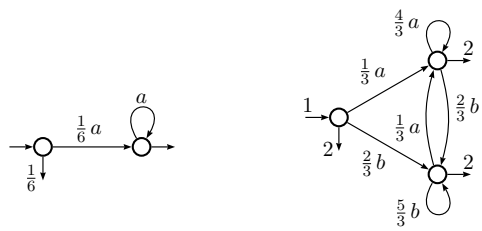
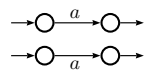
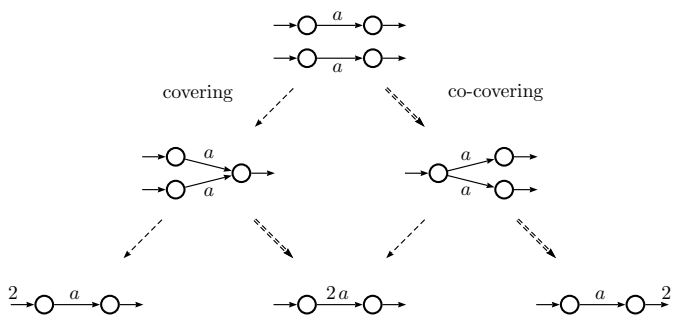


Figure SE.2.4: (Proper) standard $\mathbb{Q}$-automata for $\frac{1}{6}\,a^*$ and for E_1

2.28 *Compute all the $\mathbb{N}$-quotients and $\mathbb{N}$-co-quotients of the automaton in the adjacent figure.*



Solution:



2.29 ***Characteristic automata.***[76] (a) *Show that* every $\mathbb{N}$-automaton over M is equivalent to a characteristic $\mathbb{N}$-automaton. *cf. Prop. 2.24, p. 423*

(b) *Compute the characteristic* $\mathbb{N}$*-automaton equivalent to the automaton* $\mathcal{C}_1$. *cf. Exam. 2.5, p. 406*

Solution: (a) The idea of the proof is to replace each transition whose label is $n\,m$ (n in $\mathbb{N}$ and m in M) by n transitions with label m. For this, we must 'split' each state into a number of states equal to the maximum coefficient which appears in the automaton.

Let $\mathcal{A} = \langle I, E, T \rangle$ be an $\mathbb{N}$-automaton over M, of dimension Q. We can assume that $\mathcal{A}$ is *proper* and hence that I and T are vectors of integers. If T is not characteristic (that is, if its coefficients are not all 0 or 1) we can write $T = T_1 + T_2 + \cdots + T_k$, where all the T_i are characteristic. Then if we take k distinct copies of $\mathcal{A}$ and in each one replace T by one of the T_i, we obtain an co-$\mathbb{N}$-covering of $\mathcal{A}$. In what follows, we therefore suppose that T is characteristic.

Let k be the maximum of the coefficients in $\mathcal{A}$ (hence in E and I). If $\mathcal{A}$ is not characteristic, k is greater than or equal to 2. We write $\mathbf{k}$ for the interval $\mathbf{k} = \{0, \ldots, k-1\}$ in $\mathbb{N}$. Let $\mathcal{A}' = \langle I', E', T' \rangle$ be the $\mathbb{N}$-automaton of dimension $Q \times \mathbf{k}$ defined in the following manner. For each transition

$$p \xrightarrow[\mathcal{A}]{n\,m} q \ ,$$

and for each i in $\mathbf{k}$, we have a transition

$$(p, i) \xrightarrow[\mathcal{A}']{m} (q, j) \ ,$$

for all j in $\mathbf{k}$ such that

$$j - i = l \bmod k \quad \text{with} \quad 0 \leqslant l < n \ .$$

That is, in each (p, i) there are exactly n distinct transitions labelled m which go to a state which projects on to q in $\mathcal{A}$. Furthermore,

$$\text{if} \quad I_p = n \quad \text{then} \quad I_{(p,i)} = \begin{cases} 1 & \text{for} \quad 0 \leqslant i < n \\ 0 & \text{otherwise} \end{cases}$$

and, for each i in $\mathbf{k}$, $T_{(q,i)} = T_q$. These definitions make $\mathcal{A}'$ a $\mathbb{N}$-covering of $\mathcal{A}$.

We deduce from this proof that if in $\mathcal{A} = \langle I, E, T \rangle$, T (resp. I) is already characteristic, then we can choose an $\mathbb{N}$-covering (resp. an co-$\mathbb{N}$-covering) of $\mathcal{A}$ as the characteristic $\mathbb{N}$-automaton equivalent to $\mathcal{A}$.

(b) *Cf.* Figure SE.2.5.

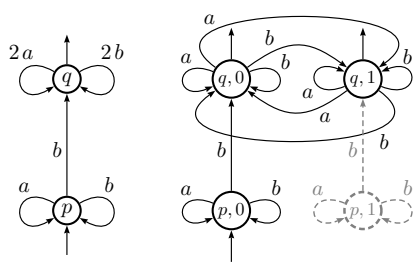


Figure SE.2.5: Characteristic $\mathbb{N}$-covering of $\mathcal{C}_1$

[76] *Cf.* Eilenberg, *op. cit.*, Theor. VI.9.1. Eilenberg, *op. cit.*, calls 'unambiguous' the automata that I call 'characteristic'; I call 'unambiguous' an automaton whose behaviour is a characteristic series, which is obviously a stricter condition.

2.30 ***Equisubtractive semirings.*** *A (commutative) monoid is called* equisubtractive *if for all p, q, r and s such that $p+q=r+s$, there exist x, y, z and t such that $p=x+y$, $q=z+t$, $r=x+z$ and $s=y+t$. A semiring is equisubtractive if it is so as a monoid for addition.*

(a) *Verify that if $\mathbb{K}$ is equisubtractive, then so are $\mathbb{K}\langle M\rangle$ and $\mathbb{K}\langle\langle M\rangle\rangle$.*

(b) *Verify that if $\mathbb{K}$ is equidivisible as a monoid for addition, the same is not necessarily true of $\mathbb{K}\langle M\rangle$.*

(c) *Let $\mathbb{K}$ be an equisubtractive ring and l_1, l_2, …, l_n, k_1, k_2, …, k_m elements of $\mathbb{K}$ such that*

$$l_1+l_2+\cdots+l_n=k_1+k_2+\cdots+k_m \quad . \tag{SE.2.2}$$

Show that there exists an $n\times m$ matrix h such that the sum of the entries of each row i is equal to l_i and the sum of the entries of each column j is equal to k_j.

(d) *Let $\mathbb{K}$ be an* equisubtractive *semiring and let $\mathcal{A}=\langle I,E,T\rangle$, $\mathcal{B}=\langle J,F,U\rangle$ and $\mathcal{C}=\langle K,G,V\rangle$ be three $\mathbb{K}$-automata over M such that $\mathcal{A}$ is a $\mathbb{K}$-covering of $\mathcal{C}$ and $\mathcal{B}$ is a co-$\mathbb{K}$-covering of $\mathcal{C}$. Show that there exists a $\mathbb{K}$-automaton $\mathcal{D}$ which is both a co-$\mathbb{K}$-covering of $\mathcal{A}$ and a $\mathbb{K}$-covering of $\mathcal{B}$ (Proposition 2.25).*

Solution: (a) Obvious, since the addition of polynomials, and of series, is pointwise.

(b) The reason is that the solution of an equidivisibility equation contains an alternative and there is no reason why the choice should come out the same for two different 'points'.

For example, $\mathbb{N}$ is equidivisible, but not $\mathbb{N}\langle a^*\rangle$; consider for example the equality

$$(2\,a+a^2)+(a+4\,a^2)=(a+2\,a^2)+(2\,a+3\,a^2)$$

(c) By induction on n and m. The property is trivial for row or column vectors. The property for $n=m=2$ is exactly the definition of equisubtractivity.

Under the assumption of (SE.2.2), set $k_1+k_2+\cdots+k_{m-1}=k'$. The induction hypothesis ensures the existence of the l'_i and of the $h_{i,m}$ such that $l'_i+h_{i,m}=l_i$ for $1\leqslant i\leqslant n$, $l'_1+l'_2+\cdots+l'_n=k'$ and $h_{1,m}+h_{2,m}+\cdots+h_{n,m}=k_m$, and then ensures the existence of the $n\times(m-1)$ matrix, which completes the solution.

(d) We write φ for the $\mathbb{K}$-covering of $\mathcal{A}$ on $\mathcal{C}$, and ψ for the co-$\mathbb{K}$-covering of $\mathcal{B}$ on $\mathcal{C}$. The dimension of $\mathcal{C}$ is S, that of $\mathcal{A}$, Q and that of $\mathcal{B}$, R.

The $\mathbb{K}$-automaton $\mathcal{D}$ has dimension $Q\times R$, the co-$\mathbb{K}$-covering of $\mathcal{D}$ on $\mathcal{A}$ is the projection on the first component, and the $\mathbb{K}$-covering of $\mathcal{D}$ on $\mathcal{B}$ is the projection on the second. Let u and v be in S, and $z=G_{u,v}$ be in $\mathbb{K}\langle\langle M\rangle\rangle$; let p be in $u\varphi^{-1}$ and s be in $v\psi^{-1}$; let $\{q_1,q_2,\ldots,q_m\}=v\varphi^{-1}$ and $\{r_1,r_2,\ldots,r_n\}=u\psi^{-1}$. For $1\leqslant j\leqslant m$, $k_j=E_{p,q_j}$ and for $1\leqslant i\leqslant n$, $l_i=F_{r_i,s}$. By assumption (φ is a $\mathbb{K}$-covering and ψ a co-$\mathbb{K}$-covering), we have

$$l_1+l_2+\cdots+l_n=k_1+k_2+\cdots+k_m=z \quad .$$

For each i,j, we write $h_{i,j}$ for the label of the transition which goes from (p,r_i) to (q_j,s) in $\mathcal{D}$. For the projection on the first component to be a co-$\mathbb{K}$-covering of $\mathcal{D}$ on $\mathcal{A}$, the sum of the $h_{i,j}$ must be equal to k_j for each j; for the projection on the second component to be a $\mathbb{K}$-covering of $\mathcal{D}$ on $\mathcal{B}$, the sum of the $h_{i,j}$ must be equal to l_i for each i. We know by the assumption that $\mathbb{K}$ is equisubtractive that it is possible to construct an $n\times m$ matrix h which satisfies all these conditions.

3. $\mathbb{K}$-recognisable series

cf. Propty II.2.1, p. 235

3.2 *Prove that* if s is a $\mathbb{K}$-recognisable series on M and N a submonoid of M then $s\odot\underline{N}$ is a $\mathbb{K}$-recognisable series on N *(Property 3.1).*

Solution: If (λ,μ,ν) is a $\mathbb{K}$-representation of M which recognises s, the restriction of μ to N gives a $\mathbb{K}$-representation (of N) which recognises $s\odot\underline{N}$.

3.3 *Show*[77] *that a series s of $\mathbb{K}\langle\langle M\rangle\rangle$ is $\mathbb{K}$-recognisable if and only if there exists an integer n and a morphism $\mu\colon M \to \mathbb{K}^{n\times n}$ such that*

$$\forall m \in M \setminus 1_M \qquad \langle s, m\rangle = (m\mu)_{1,n} \quad .$$

Solution: The condition is obviously sufficient. Let (η, κ, ζ) be a $\mathbb{K}$-representation which recognises s. For all m in M, set

$$m\mu = \begin{pmatrix} 0 & \boxed{\eta\cdot m\kappa} & \eta\cdot m\kappa\cdot\zeta \\ \boxed{0} & \boxed{m\kappa} & \boxed{m\kappa\cdot\zeta} \\ 0 & \boxed{0} & 0 \end{pmatrix}$$

We verify easily that μ is a morphism:

$$\begin{pmatrix} 0 & \eta\cdot m\kappa & \eta\cdot m\kappa\cdot\zeta \\ 0 & m\kappa & m\kappa\cdot\zeta \\ 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & \eta\cdot n\kappa & \eta\cdot n\kappa\cdot\zeta \\ 0 & n\kappa & n\kappa\cdot\zeta \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \eta\cdot m\kappa\cdot n\kappa & \eta\cdot m\kappa\cdot n\kappa\cdot\zeta \\ 0 & m\kappa\cdot n\kappa & m\kappa\cdot n\kappa\cdot\zeta \\ 0 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & \eta\cdot(m\,n)\kappa & \eta\cdot(m\,n)\kappa\cdot\zeta \\ 0 & (m\,n)\kappa & (m\,n)\kappa\cdot\zeta \\ 0 & 0 & 0 \end{pmatrix}$$

and it answers the question by its definition.

3.4 *Show*[78] *that a series s of $\mathbb{K}\langle\langle M\rangle\rangle$ is $\mathbb{K}$-recognisable if and only if there exists an integer n, a matrix z of $\mathbb{K}^{n\times n}$ and a morphism $\mu\colon M \to \mathbb{K}^{n\times n}$ such that*

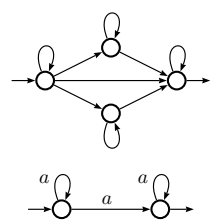
cf. Exer. 0.6.7 for the definition of the trace of a matrix.

$$\forall m \in M \qquad \langle s, m\rangle = \mathsf{Tr}(z\cdot m\mu) \quad .$$

Solution: If s is recognisable, then, as in the last exercise, we take a morphism μ such that $\langle s, m\rangle = m\mu_{1,n}$ (for all m in M). The matrix z all of whose coefficients are zero except $z_{n,1} = 1_{\mathbb{K}}$ gives $\mathsf{Tr}(z\cdot m\mu) = m\mu_{1,n} = \langle s, m\rangle$.

Conversely, if $\mu\colon M \to \mathbb{K}^{n\times n}$ is a morphism, each of the series $s_{p,q}$, $\langle s_{p,q}, m\rangle = m\mu_{p,q}$, is recognisable. The series s, $\langle s, m\rangle = \mathsf{Tr}(z\cdot m\mu) = \sum_{p,q\in[n]} z_{q,p}\, m\mu_{p,q}$, is a linear combination of the $s_{p,q}$.

3.10 *Compute the behaviour of the $\mathbb{N}$-automaton over a^* in the adjacent figure (all the transitions are labelled a).*

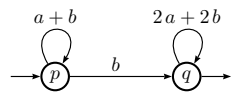


Solution: We note that this automaton is obtained (by the method of Lemma 3.8) from the tensor square of the representation associated with the automaton in the adjacent figure, whose behaviour is $s = \sum_{n\in\mathbb{N}} n\,a^n$. By virtue of Theorem 3.2, $t = \sum_{n\in\mathbb{N}} n^2\,a^n$.

3.12 *Consider the $\mathbb{N}$-automaton $\mathcal{C}_1$ of Example 2.5, and use the conventions given there.*

(a) *Prove that the $\mathbb{N}$-automaton $\mathcal{C}_2$ of Example 2.8 'computes' $\overline{f}^{\,2}$; i.e., is such that for all f, $\langle|\mathcal{C}_2|, f\rangle = \overline{f}^{\,2}$.*

(b) *Construct, using the method of Example 2.8, an $\mathbb{N}$-automaton $\mathcal{V}_3$ which has four states and which 'computes' $\overline{f}^{\,3}$.*



The $\mathbb{N}$-automaton $\mathcal{C}_1$

[77] 'Normalisation' of representations taken from [221].

[78] Another 'normalisation' of representations, also taken from [221].

(c) *Generalise and construct, for each integer n and still with the same method, an $\mathbb{N}$-automaton $\mathcal{V}_n$ which has $n+1$ states and which computes $\overline{f}^{\,n}$:*

$$\forall f \in A^* \qquad \langle |\mathcal{V}_n|, f\rangle = \overline{f}^{\,n} \ .$$

(d) *Compare with the automaton $\mathcal{V}'_n$ constructed in Exercise 2.17.*

Solution: (a) The series $|\mathcal{C}_1|$ is realised by the $\mathbb{N}$-representation $(\lambda_1, \mu_1, \nu_1)$ defined by

$$a\mu_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \ , \quad b\mu_1 = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} \ , \quad \lambda_1 = \begin{pmatrix} 1 & 0 \end{pmatrix} \ , \quad \nu_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \ ,$$

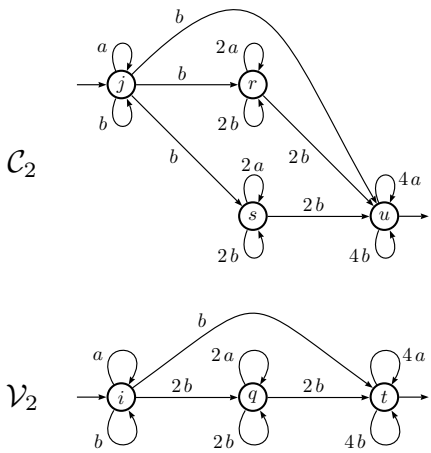


Fig. SE.3.1: The $\mathbb{N}$-automata $\mathcal{C}_2$ and $\mathcal{V}_2$

and computes $\overline{f}$. The Hadamard-squared series of $|\mathcal{C}_1|$ hence computes $\overline{f}^{\,2}$. It is thus realised by the representation $(\lambda_2, \mu_2, \nu_2)$ obtained by forming the tensor product of $(\lambda_1, \mu_1, \nu_1)$ with itself. It follows that

$$a\mu_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix} \ , \quad b\mu_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 4 \end{pmatrix} \ ,$$

$$\lambda_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \ , \qquad \nu_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \ .$$

where we recognise the $\mathbb{N}$-automaton $\mathcal{C}_2$ (Figure SE.3.1) of Example 2.8. We have seen that it is equivalent to the 3-state automaton $\mathcal{V}_2$.

(b) We start from the $\mathbb{N}$-automata $\mathcal{C}_1$ and $\mathcal{V}_2$. The behaviour of $\mathcal{V}_2$ is realised by the representation $(\eta_2, \kappa_2, \xi_2)$ defined by

$$a\kappa_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix} \ , \quad b\kappa_2 = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & 4 \end{pmatrix} \ , \quad \eta_2 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \ , \quad \xi_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \ .$$

By forming the Hadamard product of $(\eta_2, \kappa_2, \xi_2)$ with $(\lambda_1, \mu_1, \nu_1)$ we obtain a representation $(\lambda_3, \mu_3, \nu_3)$ such that

$$\forall f \in A^* \qquad \lambda_3 \cdot f\mu_3 \cdot \nu_3 = \overline{f}^{\,3} \ .$$

We have:

$$\lambda_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$a\mu_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 8 \end{pmatrix}, \qquad b\mu_3 = \begin{pmatrix} 1 & 1 & 2 & 2 & 1 & 1 \\ 0 & 2 & 0 & 4 & 0 & 2 \\ 0 & 0 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 4 & 0 & 4 \\ 0 & 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 0 & 0 & 8 \end{pmatrix}, \qquad \nu_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \ .$$

To prepare for what follows, we adopt the convention that the states of $\mathcal{V}_2$ are $\{0, 1, 2\}$ and those of $\mathcal{C}_1$ are $\{0, 1\}$; thus, the representation $(\lambda_3, \mu_3, \nu_3)$ corresponds to the automaton of Figure SE.3.2.

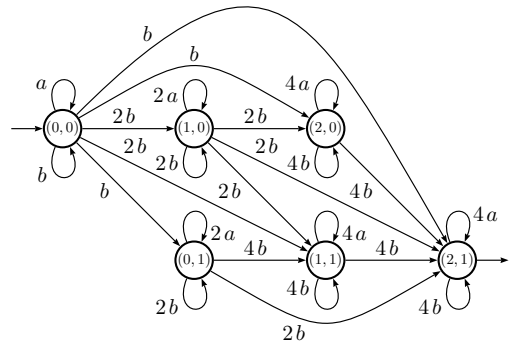


Figure SE.3.2: The $\mathbb{N}$-automaton $\mathcal{C}_3$

We then observe that if we form, in the matrices $a\mu_3$ and $b\mu_3$, the sum of the columns $(0,1)$ and $(1,0)$ on one hand and $(1,1)$ and $(2,0)$ on the other, we obtain matrices whose corresponding rows are equal:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} 1 & 3 & 3 & 1 \\ 0 & 2 & 4 & 2 \\ 0 & 2 & 4 & 2 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 8 \end{pmatrix} .$$

We thus obtain an automaton $\mathcal{V}_3$ (whose states we have renamed 0, 1, 2 and 3) of which $\mathcal{C}_3$ is an $\mathbb{N}$-covering and to which it is therefore equivalent (Figure SE.3.3).

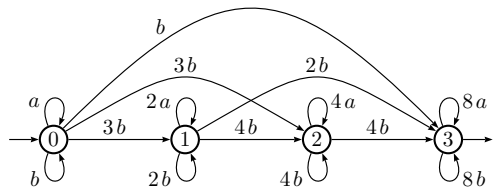


Figure SE.3.3: The $\mathbb{N}$-automaton $\mathcal{V}_3$

(c) Observation of the representations $(\eta_1, \kappa_1, \xi_1)$, $(\eta_2, \kappa_2, \xi_2)$ and $(\eta_3, \kappa_3, \xi_3)$ of the automata $\mathcal{V}_1 = \mathcal{C}_1$, $\mathcal{V}_2$ and $\mathcal{V}_3$:

$$a\kappa_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} , \quad b\kappa_1 = \begin{pmatrix} 1 & 1 \\ 0 & 2 \end{pmatrix} , \quad a\kappa_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix} , \quad b\kappa_2 = \begin{pmatrix} 1 & 2 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & 4 \end{pmatrix} ,$$

$$a\kappa_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix} , \quad b\kappa_3 = \begin{pmatrix} 1 & 3 & 3 & 1 \\ 0 & 2 & 4 & 2 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 8 \end{pmatrix} ,$$

leads to the induction hypothesis

$$(a\kappa_n)_{i,j} = \begin{cases} 2^i & \text{if } i = j , \\ 0 & \text{otherwise,} \end{cases} \qquad (b\kappa_n)_{i,j} = \begin{cases} 2^i \, \mathrm{C}_{n-i}^{j-i} & \text{for } 0 \leqslant i \leqslant j \leqslant n , \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(SE.3.1)}$$

On the other hand, the transformation from $b\mu_3$ (see above) to $b\kappa_3$ suggests the proof of the induction step. The tensor product of $(\eta_n, \kappa_n, \xi_n)$ and $(\lambda_1, \mu_1, \nu_1)$ gives the representation

$(\lambda_{n+1}, \mu_{n+1}, \nu_{n+1})$, of dimension $[0,n]\times\{0,1\}$, of the automaton $\mathcal{C}_{n+1}$ which computes $\overline{f}^{\,n+1}$ for each f of A^* (from now on we will only deal with the matrix $b\mu_{n+1}$; the rest of the representation is treated identically):

$$(b\mu_{n+1})_{(i,0),(j,0)} = (b\kappa_n)_{i,j} = 2^i\,\mathrm{C}_{n-i}^{j-i} \qquad (b\mu_{n+1})_{(i,0),(j,1)} = (b\kappa_n)_{i,j} = 2^i\,\mathrm{C}_{n-i}^{j-i}$$
$$(b\mu_{n+1})_{(i,1),(j,0)} = 0 \qquad (b\mu_{n+1})_{(i,1),(j,1)} = 2\,(b\kappa_n)_{i,j} = 2^{i+1}\,\mathrm{C}_{n-i}^{j-i}\ .$$

We form the sum of the *columns* $(j,1)$ and $(j+1,0)$, $0 \leqslant j < n$, in this matrix and observe that *rows* $(i,1)$ and $(i+1,0)$ are *equal*:

$$\begin{aligned}(b\mu_{n+1})_{(i,1),(j,1)} + (b\mu_{n+1})_{(i,1),(j+1,0)} &= 2^{i+1}\,\mathrm{C}_{n-i}^{j-i} + 0 = 2^{i+1}\,\mathrm{C}_{n+1-(i+1)}^{(j+1)-(i+1)}\\ (b\mu_{n+1})_{(i+1,0),(j,1)} + (b\mu_{n+1})_{(i+1,0),(j+1,0)} &= 2^{i+1}\,\mathrm{C}_{n-(i+1)}^{j-(i+1)} + 2^{i+1}\,\mathrm{C}_{n-(i+1)}^{(j+1)-(i+1)}\\ &= 2^{i+1}\,\mathrm{C}_{n+1-(i+1)}^{(j+1)-(i+1)}\ .\end{aligned}$$

Thus, in the automaton $\mathcal{C}_{n+1}$ the amalgamation of the states $(i,1)$ and $(i+1,0)$, $0 \le i < n$, is legitimate, and gives an $\mathbb{N}$-covering $\mathcal{V}_{n+1}$, with $n+2$ states, whose transitions are well defined by (SE.3.1).

(d) We can obviously compute co-$\mathbb{N}$-coverings instead of $\mathbb{N}$-coverings. We saw that $\mathcal{C}_2$ is an co-$\mathbb{N}$-covering of $\mathcal{V}'_2$. The tensor product of the representation $(\eta'_2, \kappa'_2, \xi'_2)$ by $(\lambda_1, \mu_1, \nu_1)$ gives the representation of an automaton $\mathcal{C}'_3$ which is mapped by an co-$\mathbb{N}$-covering to the 4-state automaton $\mathcal{V}'_3$.

cf. Fig. 2.9, p. 421

cf. Fig. SE.2.2, p. 504

Observation of the three representations $(\eta_1, \kappa_1, \xi_1) = (\eta'_1, \kappa'_1, \xi'_1)$, $(\eta'_2, \kappa'_2, \xi'_2)$ and $(\eta'_3, \kappa'_3, \xi'_3)$ of the automata $\mathcal{V}'_1 = \mathcal{C}_1$, $\mathcal{V}'_2$ and $\mathcal{V}'_3$:

$$a\kappa'_1 = \begin{pmatrix}1&0\\0&2\end{pmatrix}\ ,\quad b\kappa'_1 = \begin{pmatrix}1&1\\0&2\end{pmatrix}\ ,\quad a\kappa'_2 = \begin{pmatrix}1&0&0\\0&2&0\\0&0&4\end{pmatrix}\ ,\quad b\kappa'_2 = \begin{pmatrix}1&1&1\\0&2&4\\0&0&4\end{pmatrix}\ ,$$

$$a\kappa'_3 = \begin{pmatrix}1&0&0&0\\0&2&0&0\\0&0&4&0\\0&0&0&8\end{pmatrix}\ ,\quad b\kappa'_3 = \begin{pmatrix}1&1&1&1\\0&2&4&6\\0&0&4&12\\0&0&0&8\end{pmatrix}\ ,$$

leads to the induction hypothesis

$$(a\kappa_n)_{i,j} = \begin{cases}2^i & \text{if} \quad i=j\ ,\\ 0 & \text{otherwise}\ ;\end{cases} \qquad (b\kappa_n)_{i,j} = \begin{cases}2^i\,\mathrm{C}_i^j & \text{for} \quad 0 \leqslant i \leqslant j \leqslant n\ ,\\ 0 & \text{otherwise}\ .\end{cases} \tag{SE.3.2}$$

The proof of (SE.3.2) proceeds, by induction on n, in the same way as that of (SE.3.1), and we find there the formulas that allowed us to define the automaton $\mathcal{V}'$ in the solution to Exercise 2.17.[79]

cf. p. 503

cf. p. 436

3.15 ***Shuffle product.*** (a) *Prove Proposition 3.15.*

(b) *Prove the following identities:*[80]

(i) $a^* \between (-a)^* = 1$; (ii) $(1+a) \between a^* = (a^*)^2$; (iii) $(a\,b)^* \between (-a\,b)^* = (-4\,a^2b^2)^*$.

(c) *Spell out the construction of the* infiltration product *of two* $\mathbb{K}$*-automata* $\mathcal{A}'$ *and* $\mathcal{A}''$ *over* A^**; i.e., of the automaton* $\mathcal{A}' \uparrow \mathcal{A}''$ *such that* $|\mathcal{A}' \uparrow \mathcal{A}''| = |\mathcal{A}'| \uparrow |\mathcal{A}''|$.

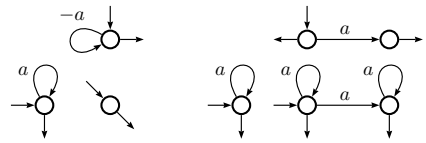
---

[79] The automata $\mathcal{V}_n$ and $\mathcal{V}'_n$ were inspired by automata constructed in [70, 74]. The point of view of these two articles is different from ours since it consists of defining real functions with automata that read the *infinite* binary representation of numbers in the interval $[0,1]$ and studying, among other things, their continuity. There is nevertheless a common basis, and it is no surprise that the automaton from which $\mathcal{V}_n$ is taken recognises the function x^n.

[80] These identities were shown to me by M. Waldschmidt; the third one is due to M. Petitot (personal communications). I found it instructive to give them a purely 'automatic' proof.

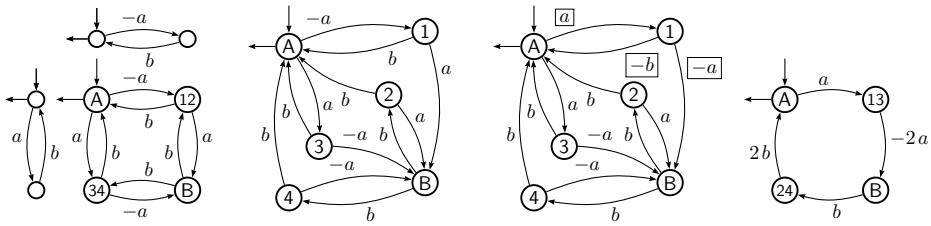
Solution: (a) The construction of $\mathcal{A}' \between \mathcal{A}''$ described in Proposition 3.15 corresponds to the transcription into an automaton of the representation $\chi[\mu' \otimes \mu'']$ where χ is the morphism defined in Proposition 3.7 (and where μ' and μ'' are the representations of $\mathcal{A}'$ and $\mathcal{A}''$ respectively). *cf. Th. 3.3, p. 431*

(b) The first two identities are immediate from the adjacent figures; the third is proved by the sequence of constructions shown in Figure SE.3.4.



We begin by constructing a covering of the $\mathbb{Z}$-automaton which accepts $(a\,b)^* \between (-a\,b)^*$, shown in Figure SE.3.4 (a). The transitions that arrive at 12 (resp. at 34) are the sum of those which *arrive* at 1 and 2 (resp. at 3 and 4), understanding that the same transitions *leave* from these states.

In (c), we make the coefficient -1 'circulate' through the state 1: the transition label which *arrives* at 1 is multiplied by -1 and those of the transitions which *leave* from there by -1 (which is its own inverse). The automaton thus obtained is a co-covering of that shown in (d) which accepts $(-4\,a^2b^2)^*$: the transitions that leave from 13 (resp. from 24) are the sum of those which *leave* from 1 and 3 (resp. from 2 and 4), understanding that the same transitions *arrive* at these states.



(a) shuffle (b) covering (c) circulation (d) co-quotient

Figure SE.3.4: The identity $(a\,b)^* \between (-a\,b)^* = (-4\,a^2b^2)^*$

(c) Let $\mathcal{A}' = \langle\, Q', A, E', I', T' \,\rangle$ and $\mathcal{A}'' = \langle\, Q'', A, E'', I'', T'' \,\rangle$ be two proper $\mathbb{K}$-automata over A (the initial and final vectors are scalar and the transitions are labelled by letters equipped with coefficients). In imitation of Proposition 3.15, the infiltration of $|\mathcal{A}'|$ and $|\mathcal{A}''|$ is realised by the $\mathbb{K}$-automaton written $\mathcal{A}' \uparrow \mathcal{A}''$ and defined by

$$\mathcal{A}' \uparrow \mathcal{A}'' = \langle\, Q' \times Q'', A, E, I' \otimes I'', T' \otimes T'' \,\rangle \ ,$$

where the set E of transitions is described by

$$\begin{aligned} E = \big\{ \big((p', p''), k'\,a, (q', p'')\big) \;\big|\; (p', k'\,a, q') \in E' \quad \text{and} \quad p'' \in Q'' \big\} \\ \bigcup \big\{ \big((p', p''), k''\,a, (p', q'')\big) \;\big|\; p' \in Q' \quad \text{and} \quad (p'', k''\,a, q'') \in E'' \big\} \\ \bigcup \big\{ \big((p', p''), (k'k'')\,a, (q', q'')\big) \;\big|\; (p', k'\,a, q') \in E' \quad \text{and} \quad (p'', k''\,a, q'') \in E'' \big\} \ . \end{aligned}$$

This construction corresponds to the transcription into an automaton of the representation $\psi[\mu' \otimes \mu'']$ where ψ is the morphism defined in Exercise 3.7 (and where μ' and μ'' are the representations of $\mathcal{A}'$ and $\mathcal{A}''$ respectively). *cf. p. 432*

4. Series on a free monoid

4.2 *Prove:* $\forall f \in A^* \quad f^{-1}T = f\mu \cdot T$ *using vectorial notation.* *cf. Equat. (4.8), p. 441*

Solution: $\forall f, g \in A^* \quad \langle f^{-1}T, g\rangle = (f\,g)\mu \cdot \nu = f\mu \cdot g\mu \cdot \nu = f\mu \cdot \langle T, g\rangle$.

cf. p. 440, p. 428 **4.3** *Use Theorem 4.1 to prove Theorem 3.2:*

Let $\mathbb{I}$ and $\mathbb{L}$ be two commutable sub-semirings of a semiring $\mathbb{K}$. Let s be in $\mathbb{I}\mathrm{Rec}\, A^*$ and t in $\mathbb{L}\mathrm{Rec}\, A^*$; show that $s \odot t$ is $\mathbb{K}$-recognisable.

Solution:[81] The series s is $\mathbb{I}$-recognisable: s belongs to a stable submodule U of $\mathbb{I}\langle\langle A^*\rangle\rangle$ of finite type, generated by $\{s_q \mid q \in Q\}$. The series t is $\mathbb{L}$-recognisable: t belongs to a stable submodule V of $\mathbb{L}\langle\langle A^*\rangle\rangle$ of finite type, generated by $\{t_r \mid r \in R\}$. Let W be the submodule of $\mathbb{K}\langle\langle A^*\rangle\rangle$ generated by the set

$$U \odot V = \{u \odot v \mid u \in U\,,\ v \in V\}\ ;$$

W contains $s \odot t$ and is stable since, for all f,

$$f^{-1}(u \odot v) = (f^{-1}u) \odot (f^{-1}v)\ .$$

It remains to show that W is of finite type. Let u be in U and v in V:

$$u = \sum_{q \in Q} h_q\, s_q \quad \text{and} \quad v = \sum_{r \in R} l_r\, t_r$$

with the h_q in $\mathbb{I}$ and the l_r in $\mathbb{L}$. For all f in A^*, we have

$$\begin{aligned}
\langle u \odot t, f\rangle = \langle u, f\rangle\langle v, f\rangle &= \left(\sum_{q \in Q} h_q \langle s_q, f\rangle\right)\left(\sum_{r \in R} l_r \langle t_r, f\rangle\right) \\
&= \sum_{q \in Q, r \in R} h_q \underbrace{\langle s_q, f\rangle l_r}_{\text{which commute}} \langle t_r, f\rangle = \sum_{q \in Q, r \in R} h_q l_r \langle s_q, f\rangle\langle t_r, f\rangle \\
&= \sum_{q \in Q, r \in R} h_q l_r \langle s_q \odot t_r, f\rangle\ .
\end{aligned}$$

Hence $u \odot t = \sum_{q \in Q, r \in R} h_q l_r\, s_q \odot t_r$, and W is generated, as a $\mathbb{K}$-module, by the series $s_q \odot t_r$.

4.7 *Verify that, if $\mathbb{K}$ is not a field, it is not true that the set of quotients of a recognisable series generates a sub-$\mathbb{K}$-module of $\mathbb{K}\langle\langle A^*\rangle\rangle$ of finite type.*

cf. p. 439 *Solution*: The example $t_2 = (a^*)^2$ answers the question: $(a^n)^{-1}t_2 = t_2 + n\,a^*$ and these series generate a sub-$\mathbb{N}$-module, which is not finitely generated, of the $\mathbb{N}$-module generated by t_2 and a^*.

4.8 *Show that the rank of the series $|\mathcal{C}_n| = \sum_{f \in A^*} \overline{f}^{\,n} f$ is $n+1$.*

Solution:[82] In Exercises 2.17 and 3.12 we computed the $\mathbb{N}$-automata $\mathcal{V}_n$ and $\mathcal{V}'_n$ which realise c_n and which have $n+1$ states: this series is of rank at most $n+1$.

cf. p. 453 As in Example 4.5, let M_n be the $(n+1)\times(n+1)$ matrix extracted from the Hankel matrix of c_n by keeping the rows and columns with indices b^i, $0 \leqslant i \leqslant n$; that is,

$$M_n = \big(m_{i,j}\big) \qquad \text{with} \qquad m_{i,j} = \big(2^{i+j} - 1\big)^n\ , \quad 0 \leqslant i, j \leqslant n\ .$$

We write

$$m_{i,j} = \sum_{k=0}^{k=n} \mathrm{C}_n^k \left(2^{i+j}\right)^k (-1)^{n-k} = \sum_{k=0}^{k=n} \left((-1)^{n-k}\, \mathrm{C}_n^k\, 2^{i\,k}\right) 2^{k\,j}$$

[81] Proof due to M. Fliess [91] (*cf.* also [31, Exerc. I.5.2]).

[82] Proof due to J. Siquier, private communication.

and hence $M_n = X_n \cdot Y_n$ where $X_n = \big(x_{i,j}\big)$ and $Y_n = \big(y_{i,j}\big)$ with

$$x_{i,j} = (-1)^{n-j}\,\mathrm{C}_n^j\,2^{i\,j} \quad , \qquad y_{i,j} = 2^{i\,j} \quad , \qquad 0 \leqslant i,j \leqslant n \quad .$$

We recognise in Y_n a *Vandermonde matrix* whose determinant is non-zero. The determinant of X_n, equal to that of Y_n multiplied by $\prod_{k=0}^{k=n}(-1)^{n-k}\mathrm{C}_n^k$, is also non-zero. Hence the determinant of M_n is non-zero.

4.10 *Give a direct proof of Corollary 4.19:*

Corollary 4.19 *Let $\mathbb{K}$ be a sub-semiring of a skew field $\mathbb{F}$. A series s on A^* realised by a $\mathbb{K}$-representation (λ,μ,ν) of dimension n is uniformly zero if and only if it is zero for every word of length less than n; that is,*

$$\forall f \in A^* \quad \lambda \cdot f\mu \cdot \nu = 0 \quad \Longleftrightarrow \quad \forall f \in A^*\,,\ |f| < n \quad \lambda \cdot f\mu \cdot \nu = 0 \quad .$$

Solution:[83] We can assume that neither λ nor ν is a null vector. Set

$$W = \{X \mid X \in \mathbb{F}^n \quad X \cdot \nu = 0\} \quad ,$$

the subspace of $\mathbb{F}^n$ orthogonal to ν and, for each $k \leqslant n-1$,

$$V_k = \langle\{\lambda \cdot f\mu \mid f \in A^* \quad |f| \leqslant k\}\rangle$$

is the subspace of $\mathbb{F}^n$ generated by all the vectors $\lambda \cdot f\mu$ where the length of f is less than or equal to k. By assumption, we have

$$V_0 \subseteq V_1 \subseteq \cdots \subseteq V_{n-1} \subseteq W \quad .$$

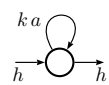
Since ν is non-null, W is a subspace of dimension $n-1$ of $\mathbb{F}^n$. Since λ is non-null, V_0 is a subspace of dimension 1. The dimensions of the V_i, increasing in i, cannot all be distinct, and there therefore exists a k, $0 \leqslant k < n-1$, such that *cf. Cor. 7.1, p. 492*

$$V_k = V_{k+1} \quad .$$

The subspace V_{k+2} is generated by all the vectors X of $V_{k+1} = V_k$ and the vectors $X \cdot a\mu$ with a in A and X in $V_{k+1} = V_k$. Thus $V_{k+2} = V_{k+1} = V_k$. By induction, $V_{k+p} = V_k \subseteq W$ for every integer p in $\mathbb{N}$. Hence $\lambda \cdot f\mu \cdot \nu = 0$ for every word f in A^*.

4.12 *Show with a simple example that if $\mathbb{K}$ is not a sub-semiring of a skew field, we can construct a $\mathbb{K}$-recognisable series not uniformly zero, but zero on all the words of length less than or equal to n and realised by a $\mathbb{K}$-representation of dimension d strictly less than n.*

Solution: We take $\mathbb{K} = \mathbb{Z}^{n\times n}$. The series s is recognised by the very simple automaton in the adjacent figure, where k is the matrix of the permutation $(1\ 2 \cdots n)$ and h is that of the (partial) function which sends 1 to 2. We have $\langle s, a^r\rangle = h\,k^r\,h = 0_{\mathbb{K}}$ for all $r < n-1$ and $\langle s, a^{n-1}\rangle = h\,k^{n-1}\,h = h \neq 0_{\mathbb{K}}$.



4.13 ***Discriminating length II.*** *Recall that we defined the* discriminating length *between two non-equivalent automata $\mathcal{A}$ and $\mathcal{B}$ to be the length of the shortest word which is accepted by one and not the other. We write $L_{\mathsf{d}}(n,m)$ (resp. $L_{\mathsf{nd}}(n,m)$) for the maximum of the discriminating lengths when $\mathcal{A}$ and $\mathcal{B}$ have respectively n and m states and are deterministic (resp. non-deterministic).* *cf. Exer. I.3.10, p. 111*

(a) *Compute $L_{\mathsf{d}}(n,m)$;* (b) *Evaluate $L_{\mathsf{nd}}(n,m)$.*

[83] Proof given by Eilenberg, *op. cit.*

Solution: (a) We consider the deterministic automata $\mathcal{A}$ and $\mathcal{B}$ as weighted automata in $\mathbb{Z}$. The series $z = |\mathcal{A}| - |\mathcal{B}|$ is $\mathbb{Z}$-rational and is zero for all words of length less than $L_{\mathsf{d}}(n,m)$. Corollary 4.19 implies that $L_{\mathsf{d}}(n,m) \leqslant n+m-1$ and Exercise 4.11 that $L_{\mathsf{d}}(n,m) = n+m-1$.

cf. p. 462

(b) If $\mathcal{A}$ and $\mathcal{B}$ are not deterministic, we can make them so to reduce this to the previous problem. In the determinisation process, we can identify the two sink states, which gives $L_{\mathsf{nd}}(n,m) \leqslant 2^n + 2^m - 2$. It is an open problem to find a better majoration.[84]

Remark SE.4.1 This problem of finding a word which distinguishes between the behaviour of two automata goes back to the very beginning of automata theory, to a similarly stated problem in the article by E. Moore, 'Gedanken-experiments on sequential machines' which appeared in the mythical and seminal collection *Automata Studies* edited by C. Shannon and J. McCarthy [232].[85] Moore developed the notion of indistinguishable states and the reduction algorithm that bears his name from such considerations, but the search for discriminating words of minimal length attracted other authors. The solution to the problem for deterministic automata (taking into account the cardinality of the alphabet) occupies Chapters 2 and 3 of J. H. Conway's book [66].

The same problem was considered for Mealy machines – that is, automata whose transitions are labelled by a pair of incoming letter and outgoing letter – in particular by P. Starke and his students. In this context a word is discriminating if, when the two automata read it as input, they produce two distinct words (or two distinct sets of words) as output. In order to have only one parameter, we seek rather to distinguish two sets of initial states in the same n-state automaton. If we write $M_{\mathsf{nd}}(n)$ for the bound of these discriminating lengths, then that $2^{n-1} - 1 \leqslant M_{\mathsf{nd}}(n) \leqslant 2^n - 2$; the majoration is found in Starke [240], the minoration is due to H.-D. Burkhard [47], with an automaton whose input alphabet varies with n (it has $2^{n-1} - 1$ letters).[86] □

4.14 ***Syntactic algebras and ideals.***

Solution: *cf.* the work of J. Berstel and Ch. Reutenauer [31], Chap. II.

4.17 *Let $t \in \mathbb{N}\mathrm{Rat}\, A^*$ and $f \in \mathsf{supp}\, t$. Give an elementary proof that $u = t - \underline{f}$ belongs to $\mathbb{N}\mathrm{Rat}\, A^*$.*

Solution: The language $K = (\mathsf{supp}\, t) \setminus \{f\}$ is rational and $u = t \odot \underline{K} + (\langle t, f\rangle - 1)f$ is an $\mathbb{N}$-rational series.

6. Rational subsets in free products

cf. Lem. 6.1, p. 477

6.1 *Prove that if M and N have no divisors of the identity, the inverse image of an element u of $M*N$ is the product of the inverse images of the factors of its canonical factorisation.*

Solution: The product of the inverse images of factors of the canonical factorisation is contained in the inverse image of $u = u_1 u_2 \cdots u_n$; that is, if $\mathrm{MI}(u) = M$ and $\mathrm{MF}u = N$, then

$$u_1\varphi^{-1}\, u_2\psi^{-1}\, u_3\varphi^{-1} \cdots u_{n-1}\varphi^{-1}\, u_n\psi^{-1} \subseteq u\,[\varphi*\psi]^{-1} \quad , \tag{SE.6.1}$$

since $\varphi*\psi$ is a morphism. Conversely, let f be in $(A \cup B)^*$ and let $f = f_1 f_2 \cdots f_n$ be its canonical factorisation as an element of $A^* * B^*$. Since M and N are without divisors of the

[84] A problem posed by Eilenberg, *op. cit.* [80, Exer. VI.4.2, p. 218] which to my knowledge remains unsolved even though everything leads one to suspect that this bound can be improved.

[85] See the history of the origins of automata theory by D. Perrin, already cited [184].

[86] I thank K. Peeva for having drawn my attention to these last references.

identity, which is the same as saying that $1_M \notin (A^+)\varphi$ and $1_N \notin (B^+)\psi$, the canonical factorisation of $f\,[\varphi * \psi]$ is

$$f\,[\varphi * \psi] = f_1\varphi\, f_2\psi\, f_3\varphi \cdots f_{n-1}\varphi\, f_n\psi$$

(still in the case where $\mathrm{MI}(f) = A^*$ and $\mathrm{MF}f = B^*$) which proves the inverse inclusion of (SE.6.1).

6.3 *Show that $(\mathbb{Z}/2\mathbb{Z}) * (\mathbb{Z}/2\mathbb{Z})$ contains a subgroup of index 2 isomorphic to $\mathbb{Z}$.*

Solution: The group $\mathbb{Z}/2\mathbb{Z}$ is presented by $\langle\, \{a\}\,;\, a\,a = 1 \,\rangle$. Hence $(\mathbb{Z}/2\mathbb{Z}) * (\mathbb{Z}/2\mathbb{Z})$ is presented by $\langle\, \{a, b\}\,;\, a\,a = 1, b\,b = 1 \,\rangle$. It is a simplification monoid. Its elements are in bijection with the language *cf. Sec. II.6.3, p. 314*

$$S = A^* \setminus A^*\{a\,a, b\,b\}A^* \ .$$

Since every power of $a\,b$ is a reduced word, $a\,b$ generates a subgroup H which is isomorphic to $\mathbb{Z}$: $H = \{(a\,b)^n \mid n \in \mathbb{N}\} \cup \{(b\,a)^m \mid m \in \mathbb{N}\}$. We have $S = H \cup H\,b$ and H is of index 2.

6.5 *Let E be a square matrix which can be decomposed into the following blocks*

$$E = \begin{pmatrix} 0 & V \\ X + Z & 0 \end{pmatrix} \ ,$$

where V, X and Z are square matrices (hence of the same dimension) and X is a boolean matrix. Show the two equalities (Lemma 6.3): *cf. p. 480*

$$\text{(i)} \qquad E^* = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \begin{pmatrix} 0 & (V\,X)^*V \\ Z & 0 \end{pmatrix}^* \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \ ;$$

$$\text{(ii)} \qquad E^* = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \begin{pmatrix} 0 & (V\,X)^*V \\ X + Z & 0 \end{pmatrix}^* \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \ .$$

Solution:[87] (i)

$$E^* = \begin{pmatrix} 0 & V \\ X + Z & 0 \end{pmatrix}^* = \left[\begin{pmatrix} 0 & 0 \\ X & 0 \end{pmatrix} + \begin{pmatrix} 0 & V \\ Z & 0 \end{pmatrix}\right]^* \qquad \text{hence, since } \begin{pmatrix} 0 & 0 \\ X & 0 \end{pmatrix}^* = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \ ,$$

$$= \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \left[\begin{pmatrix} 0 & V \\ Z & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix}\right]^* = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \begin{pmatrix} V\,X & V \\ Z & 0 \end{pmatrix}^* = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \left[\begin{pmatrix} V\,X & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & V \\ Z & 0 \end{pmatrix}\right]^*$$

$$= \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \left[\begin{pmatrix} (V\,X)^* & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & V \\ Z & 0 \end{pmatrix}\right]^* \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} (V\,X)^*V\,X & 0 \\ 0 & 0 \end{pmatrix}\right]$$

$$= \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \begin{pmatrix} 0 & (V\,X)^*V \\ Z & 0 \end{pmatrix}^* \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & (V\,X)^*V \\ Z & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ X & 0 \end{pmatrix}\right]$$

$$= \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \begin{pmatrix} 0 & (V\,X)^*V \\ Z & 0 \end{pmatrix}^+ \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \begin{pmatrix} 0 & (V\,X)^*V \\ Z & 0 \end{pmatrix}^* \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix}$$

since $X + X = X$ and hence $\begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix}$.

(ii) We expand the right hand side of (ii) using (i), obtaining

$$\begin{pmatrix} 0 & (V\,X)^*V \\ X + Z & 0 \end{pmatrix}^* = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix} \begin{pmatrix} 0 & ((V\,X)^*V\,X)^*(V\,X)^*V \\ Z & 0 \end{pmatrix}^* \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix}$$

and since $((V\,X)^*V\,X)^*(V\,X)^* = (V\,X)^*$ and $\begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix}^2 = \begin{pmatrix} 1 & 0 \\ X & 1 \end{pmatrix}$, we obtain (ii) by applying (i) again, but the other way around.

[87] Taken from [209].

6.8 *Characterise the subsets of $M*N$ whose minimal bipartite automaton is finite.*

Solution: $\mathrm{Rat}\,(\mathfrak{P}(M)\cup\mathfrak{P}(N))$, the rational closure of the *arbitrary* subsets of M and N.

7. A non-commutative linear algebra primer

7.2 ***Dual vector space.*** *In this exercise, $\mathbb{K}$ is a skew field and V a left $\mathbb{K}$-vector space.*

(a) *Verify that the set of $\mathbb{K}$-functionals on V is a right $\mathbb{K}$-vector space.*

(b) *Verify that there exists a canonical* injection *from V to $V^{\star\star}$, the dual of the dual of V.*

cf. p. 493 *Solution*: (a) The verification that Equation (7.4) indeed makes $V^\star$ a right $\mathbb{K}$-vector space is immediate, *but there is something to verify.* Among other things, for example,

$$\begin{aligned}\langle v+w,\sigma+\tau\rangle &= \langle v+w,\sigma\rangle+\langle v+w,\tau\rangle\\ &= \langle v,\sigma\rangle+\langle w,\sigma\rangle+\langle v,\tau\rangle+\langle w,\tau\rangle\\ &= \langle v,\sigma\rangle+\langle v,\tau\rangle+\langle w,\sigma\rangle+\langle w,\tau\rangle\\ &= \langle v,\sigma+\tau\rangle+\langle w,\sigma+\tau\rangle\end{aligned}$$

and the rest follow in the same way.

(b) The map $v\mapsto\langle v,\sigma\rangle$ is a functional on $V^\star$ since

$$\langle v,\sigma+\tau\rangle=\langle v,\sigma\rangle+\langle v,\tau\rangle \qquad \text{and} \qquad \langle v,\sigma\, l\rangle=\langle v,\sigma\rangle l\ .$$

We also verify just as easily as before that this map is a *morphism* (of left vector spaces) from V to $V^{\star\star}$. It remains to show that the morphism is injective: that is, its kernel reduces to 0_V.

cf. proof of Th. 7.2, p. 492 Every non-zero element x of V generates a subspace for which we can choose a supplementary subspace Y. Every v in V can be written uniquely in the form $v=v_x x+y$ with y in Y and the map $v\mapsto v_x$ is a functional on V which takes the value $1_{\mathbb{K}}$ for x and is hence non-null. This shows that there is no non-zero x such that $\langle x,\sigma\rangle=0_{\mathbb{K}}$ for all σ in $V^\star$.

7.3 ***Morphisms and matrices.*** *In this exercise, $\mathbb{K}$ is a skew field, V and U are two left $\mathbb{K}$-vector spaces and $\alpha\colon V\to U$ is a morphism.*

cf. p. 495 (a) *Verify that, for all v in V, $\underline{(v)\alpha}=\underline{v}\cdot\underline{\alpha}$ (Equation (7.6)).*

(b) *Let $\alpha^\star\colon U^\star\to V^\star$ be the dual morphism of α. Verify that the matrix of $\alpha^\star$ with bases $\mathcal{C}^\star$ and $\mathcal{B}^\star$ is equal to $\underline{\alpha}$, the matrix of α with bases $\mathcal{B}$ and $\mathcal{C}$.*

Solution: (a) By definition, $v=\sum_{i\in[r]}\underline{v}_i\,b^i$ and, for all i in $[r]$, $(b^i)\alpha=\sum_{j\in[s]}\underline{\alpha}^i_j\,c^j$. Hence

$$v\alpha=\sum_{i\in[r]}\underline{v}_i\left(\sum_{j\in[s]}\underline{\alpha}^i_j\,c^j\right)=\sum_{j\in[s]}\left(\sum_{i\in[r]}\underline{v}_i\,\underline{\alpha}^i_j\right)c^j\ ,$$

which says precisely that, for all j in $[s]$, $\left(\underline{(v)\alpha}\right)_j=(\underline{v}\cdot\underline{\alpha})_j$.

(b) By definition we have

$$\langle b^i\alpha,\gamma_j\rangle=\underline{\alpha}^i\cdot\underline{\gamma}_j=\underline{\alpha}^i_j=\langle b^i,\alpha^\star(\gamma_j)\rangle=\langle\underline{b}^i,\underline{\alpha^\star}_j\rangle=\underline{\alpha^\star}^i_j\ .$$

NOTES & REFERENCES

Several now classic works deal with series over the free monoid.

S. Eilenberg, *Automata, Languages and Machines*, volume A. Academic Press, 1974. [80]

A. Salomaa and M. Soittola, *Automata-Theoretic Aspects of Formal Power Series.* Springer, 1977. [218]

J. Berstel and Ch. Reutenauer, *Les séries rationnelles et leurs langages.* Masson, 1984. Translation: *Rational Series and Their Languages.* Springer, 1986. [31]

W. Kuich and A. Salomaa, *Semirings, Automata, Languages.* Springer, 1986. [142]

1. Formal power series on a graded monoid

The construction of the topology on $\mathbb{K}\langle\langle M\rangle\rangle$ is standard, even if the object $\mathbb{K}\langle\langle M\rangle\rangle$ itself is not particularly; see [228] for example.

2. $\mathbb{K}$-automata and $\mathbb{K}$-rational series

I was led to define 'strong' semirings, a terminology suggested to me by J. van der Hoeven, to be able to prove the equivalence between the existence of the star of an arbitrary series and that of the star of its constant term. This equivalence is necessary to be able next to construct a theory of rational series which runs parallel to that of rational languages.

In a sense the Fundamental Theorem is what Kleene showed for automata over A, or that which is often called the Kleene–Schützenberger Theorem for weighted automata. However, because these results apply to automata over free monoids, their standard form states the identity between rational languages and recognisable languages, which *no longer holds* for automata (weighted or otherwise) over arbitrary monoids.

Kleene's Theorem was therefore split in two, as it were: one part which holds for automata over arbitrary monoids and which, considering what the proof involves, concentrates the substance of the theorem; and one part which holds only for automata over free monoids and which is nearly a formality. In summary, the Fundamental Theorem is indeed Kleene's Theorem, but we cannot give it that name since its statement is inconsistent with the usual statements of Kleene's Theorem.

I already quoted the similarity between $\mathbb{K}$-coverings and *state amalgamation* and *state splitting* (In- and Out-) in *symbolic dynamical systems* [150, § 2.4]. If $\mathcal{B}$ is obtained from $\mathcal{A}$ by an In-splitting (resp. Out-splitting) then $\mathcal{B}$ is an $\mathbb{N}$-covering (resp. co-$\mathbb{N}$-covering) of $\mathcal{A}$. But the converse is not true. Roughly speaking, $\mathcal{B} = \langle J, F, U \rangle$ is a $\mathbb{K}$-covering of $\mathcal{A}$ if the rows with 'equivalent' indices of the matrix $F \cdot H_\varphi$ are equal while $\mathcal{A}$ is obtained by amalgamation from $\mathcal{B}$ if the rows with 'equivalent' indices of the matrix F are equal.[88]

cf. Rem. 2.12, p. 423

cf. notations of Def. 2.10, p. 419

It follows that the composition of two (In-)amalgamations is not necessarily an amalgamation, contrary to the situation with $\mathbb{K}$-coverings. Theorem 2.3 is the equivalent for $\mathbb{K}$-coverings of the theorem of R. Williams for amalgamations. For symbolic dynamical systems and their relationship with finite automata, see also [17] and [22].

3. $\mathbb{K}$-recognisable series

The definition of representations in the form (λ, μ, ν) is due to Fliess [90]. Theorem 3.1 is due to Schützenberger [222], including the more general formulation of Theorem 3.2. Proposition 3.15 is also due to Fliess [91].

[88] *Note added in translation:* In recent work ([20, 21]), M.-P. Béal, S. Lombardy and I gave the $\mathbb{K}$-coverings a definition much more in line with symbolic dynamics.

4. Series on a free monoid

The first subsection follows [31]. The original work is due to Schützenberger [220, 221]. The characterisation of recognisable series (Theorem 4.1) is a generalisation, due to Jacob [126], of the property (Theorem 4.5) stated by Fliess for the case of series on a field [90].

The derivation of weighted expressions, the next subsection, is a generalisation of V. Antimirov's work on derived terms [8], described in Chapter I. We note once more that the introduction of weights clarifies and structures a result on languages, even if having to take into account not necessarily positive semirings adds a certain complexity. The material of this subsection was taken from [153]. With somewhat different techniques, Rutten [203, 204] also proved Theorems 4.3 and 4.4.[89]

The two final subsections return to the path taken by [31]. We present the theory of the reduction of representations in a slightly different way to make it apply to skew fields, but without recourse Dieudonné's theory of determinants (*cf.* [65, Vol. III, Ch. 11]), a possibility noted by Fliess. In [92], it was also observed that Schützenberger's reduction algorithm applies to the case of series on a skew field, with a reference to a previous theory of non-commutative determinants [200]. The complexity of this algorithm is $O(k\,n^3)$, where k is the cardinality of the alphabet and n is the dimension of the representation which is to be reduced [48].

Theorem 4.14 is due to Schützenberger [220], and was generalised to commutative rings by Sontag [237]; the proof is adapted from [31] to enable the generalisation to skew fields. Proposition 4.23 is also from [220]. Theorem 4.15 is due to A. Restivo and Ch. Reutenauer [195]. The Rational Skimming Theorem is also due to Schützenberger [223]; however, the name given to it and the present proof are original.[90]

5. Series on an arbitrary monoid

We followed the method of J. H. Conway [66] to define complete semirings. Kleene's Theorem for weighted series in a continuous semiring was proved by Kuich [141].

Most authors have not considered semirings that are *a priori* complete but semirings in which a star operation is arbitrarily defined, and the problem is to determine the axioms which the star operation must satisfy in order to have a reasonable theory; see for example [137, 36, 85].

6. Rational subsets in free products

This section was adapted from my articles [209, 210]. The initial aim of this work was the characterisation of *free partially commutative monoids* M for which $\mathrm{Rat}\,M$ is a Boolean algebra, and $\mathrm{Rat}\,M = \mathrm{URat}\,M$ or $\mathrm{Rec}\,M$ is decidable in $\mathrm{Rat}\,M$. The results in this area were obtained independently in [2, 32].

'So,' [said the doctor]. 'Now vee may perhaps to begin. Yes?'

Philip Roth

[89] *Note added in translation:* In [156], Sylvain Lombardy and I gave a slightly different formal definition of derived terms, in order to cope more satisfactorily with the phantom ones.

[90] *Note added in translation:* Note 50 of Chapter II (*cf.* p. 301) is the awkward trace of a planned last subsection which I eventually gave up writing, for it was not ready when I decided, at last, to finish off the first (French) edition. It would have been called 'Rational series and their coefficients' and have given a much simplified version of the chapter with the same name in the work of Berstel and Reutenauer (*op. cit.*). I intended in particular to present the so-called 'Fatou Lemma': *A rational series of* $\mathbb{Q}[x]$ *whose coefficients are all integers: that is, all belong to* $\mathbb{Z}$*, is a rational series of* $\mathbb{Z}[x]$ which explain the terminology for the property that I have called the 'Fatou property'.

Part II

Rationality in relations

This page intentionally left blank

Chapter IV

The richness of transducers

Relations realised by finite automata

In this chapter we undertake the study of *relations realised by finite automata*. It will be followed in the next chapter by the particularly fruitful special case of *functional relations*, or functions, which are also realised by finite automata. To this end, we shall be led to use several of the notions and results worked out in the previous two chapters, about automata over the direct products of free monoids (which are not free monoids) and about automata over free monoids but with weights in semirings of suitable coefficients.

This chapter also reproduces, in brief, the structure of the preceding three chapters. In the first section we shall begin by taking a 'set' point of view of the free monoid: automata relate words to words. We will thus construct a theory which has its roots in the origins of automata theory, and which has been more or less elaborated in many works.[1] Its main aim has been the classification of families of non-rational languages, principally of sub-families of algebraic languages, by means of relations between words thus defined. This aspect will not be tackled here, and we shall stick to the study of the properties of these relations for their own sake.

In the second and third sections we shall continue this study in the general context of weighted automata and series. We shall first consider the problems inherent in the definition of (additive) maps which relate series to series, then study those which are recognised by finite automata. The subject is less often presented in this form.[1]

This two-pass mode of exposition involves some repetition, I am aware, and I am sorry for it; but it is, I believe, a price worth paying – in any case, I have not the slightest idea how to pay less – so that on one hand simple computations can be explained in an elementary fashion and on the other the deep nature of these same computations can be presented in the general context to which they really belong and in which they take on their true significance.

The study of relations recognised by finite automata, with or without weights, is for the most part organised around two main results: the Evaluation Theorem and the Composition Theorem, which we will call the 'pivotal theorems'. In the case of weighted relations, their importance is even greater because they dictate our choices when we have to make assumptions about the semiring of coefficients.

The problem of the equivalence of transducers occupies the fourth section and

[1] See the bibliographic notes.

brings to light the fundamental difference caused by whether or not we consider the *multiplicity* with which pairs of words are related by automata. Equivalence of transducers *without* weights can be transformed more or less easily into the Post Correspondence Problem and turns out thus to be undecidable; the subtle case is that of relations with values in a one-generator free monoid, and is due independently to O. Ibarra and L. Lisovik. Equivalence of transducers *with* weights (in a field) is on the other hand decidable. The proof of this decidability is without doubt the most beautiful example of the algebraic nature of 'automata objects' that we study. It is in fact based, as T. Harju and J. Karhumäki showed, on two standard but not elementary results in algebra, of which we shall be keen to give a thorough demonstration in the supplementary section 'Malcev–Neumann series'.

Also in the second part of the chapter, two families of relations will attract our attention by the variety of situations in which they appear: the *deterministic* relations and, even more, the *synchronous* relations. The definitions of both these families are based on Rabin and Scott's model of k-tape automata already described at the end of Section 1. In fact, these two families are studied with the same 'set' point of view as that of Section 1, and can be tackled without reference to Sections 2 to 4.

A group of results about rational relations, namely those related to the 'Rational Uniformisation Theorem',[2] is deferred to the next chapter, where it takes its natural place with the Weak (and Strong) Decomposition Theorems.

Contents of the chapter

1. Rational relations: an introduction
2. $\mathbb{K}$-relations
3. Rational $\mathbb{K}$-relations
4. Equivalence of finite $\mathbb{K}$-transducers

Deeper & broader

5. Deterministic rational relations
6. Synchronisation of transducers
7. Malcev–Neumann series

Solutions to the exercises

Notes & references

[2] Or that which Eilenberg calls the 'Cross-Section Theorem' in [80].

1 Rational relations: an introduction

Before any generalisation, we shall study those relations between elements of free monoids realised by finite automata. As between two sets, a *relation between words* – that is, one between two free monoids, say A^* and B^* – can be seen from two complementary points of view.

The first point of view, which might be called *static*,[3] considers relations *according to their graphs*, hence as subsets of $A^* \times B^*$. The two sets A^* and B^* then play perfectly symmetric roles, which is not the least important feature of this point of view. If the relations are realised by automata, the automata corresponding to this point of view are labelled with elements of the monoid $A^* \times B^*$ (a *non-free* monoid).

In the second, which we call *dynamic* in contrast to the first, a relation from A^* to B^* is considered as a *map from* A^* *to* $\mathfrak{P}(B^*)$. The symmetry is broken: there is one *argument* and one or more *results*, or, using the standard terminology for automata, there is an *input* word, or a word *read*, and one or more *output* words. If the relations are realised by automata, these automata are then labelled with words in A^*, in fact, with *letters* of A, which have *coefficients* which are words (or subsets) of B^*, that enable the computation of the image of the input word. This second point of view fits naturally with series; it will therefore be just as well, if not better, to adopt this approach, which will only be touched on in this section, before being treated fully in Section 3.

Even if the aim is to treat here, in the most elementary fashion, relations *between words* realised by finite automata, the techniques used are those which were developed in Chapter II, where we were dealing with rational subsets of arbitrary monoids (in any case, not necessarily free). We can also note that it is, at least in part, in order to give clear and efficient proofs of results about relations between words that the notions of rational subsets in non-free monoids were developed. It is therefore entirely natural that we should treat almost simultaneously relations between arbitrary monoids in this section. We only restrict the statements of a few standard propositions.

1.1 Rational relations

1.1.1 Rational relations between free monoids

In this section, A and B are two finite alphabets, τ is a relation from A^* to B^*, and $\widehat{\tau} \subseteq A^* \times B^*$ is its graph. A relation is completely defined by its graph; the following definition is thus natural.

Definition 1.1 A relation $\tau \colon A^* \to B^*$ is called *rational* if its graph is a rational subset of $A^* \times B^*$. □

It follows from this definition, since the rational subsets correspond in the isomorphism that sends $A^* \times B^*$ to $B^* \times A^*$; that is:

[3] To use an expression of J. Berstel in [27].

Property 1.1 *The inverse of a rational relation from A^* to B^* is a rational relation from B^* to A^*.* ■

cf. Prop. II.1.7, p. 223

Since the image of a rational set under a morphism is a rational set, and the projections of $A^* \times B^*$ onto A^* and B^* respectively are morphisms, and $\mathsf{Dom}\,\tau$ and $\mathsf{Im}\,\tau$ are the images of $\widehat{\tau}$ under these projections, we also have:

Property 1.2 *The* domain *and the* image *of a rational relation are rational sets.* ■

Examples 1.1 Rational substitutions.

(i) The *identity* map $\iota\colon A^* \to A^*$ is rational. Its graph $\widehat{\iota}$, the diagonal of $A^* \times A^*$, can be given the rational expression

$$\widehat{\iota} = \{(f, f) \mid f \in A^*\} = \mathcal{D}^* \qquad \text{where} \qquad \mathcal{D} = \bigcup_{a \in A} (a, a) \ .$$

(ii) More generally, *every morphism* $\alpha\colon A^* \to B^*$ *is a rational relation*: an expression for its graph $\widehat{\alpha}$ is

$$\widehat{\alpha} = \{(f, f\alpha) \mid f \in A^*\} = \left(\bigcup_{a \in A} (a, a\alpha) \right)^* \ .$$

cf. Def. II.2.4, p. 252

(iii) A further step in the generalisation: a *substitution* from A^* to B^* (with σ a morphism $\sigma\colon A^* \to \mathfrak{P}(B^*)$) is called *rational* if $a\sigma$ belongs to $\mathrm{Rat}\, B^*$ for each a in A. It then follows that *every rational substitution* $\sigma\colon A^* \to B^*$ *is a rational relation* since the expression of its graph $\widehat{\sigma}$ is

$$\widehat{\sigma} = \{(f, f\sigma) \mid f \in A^*\} = \left(\bigcup_{a \in A} (a, a\sigma) \right)^* = \left(\bigcup_{a \in A} (a, 1_{A^*})(1_{A^*}, a\sigma) \right)^* \ .$$

It follows from (ii) that *inverse morphisms* and, more generally, from (iii) that *inverse rational substitutions* are *rational relations.* □

Example 1.2 Factors. The relation $\tau\colon A^* \to A^*$ which relates each word of A^* to the set of its factors is rational: its graph is given by the expression

$$\widehat{\tau} = (A^* \times 1_{A^*})\, \mathcal{D}^* (A^* \times 1_{A^*}) \ ,$$

where $\mathcal{D}$ is the set defined in Example 1.1 (i). □

cf. Exam. II.1.1, p. 224

Remark 1.1 In Chapter II, the example we gave of a monoid in which the rational subsets are not closed under intersection, and hence under complement, is the product of two free monoids $M = \{a, b\}^* \times \{c\}^*$. The intersection

$$(a, c)^*(b, 1)^* \cap (a, 1)^*(b, c)^* = \{(a^n b^n, c^n) \mid n \in \mathbb{N}\}$$

is not rational. It follows that rational relations are closed neither under intersection nor under complement (the *complement* of a relation τ is the relation whose graph is the complement of $\widehat{\tau}$). □

Example 1.3 Intersection with a rational language. Let $K \subseteq A^*$. The map *intersection with* K, written $\iota_K \colon A^* \to A^*$, and defined by

$$\forall f \in A^* \qquad f\iota_K = \{f\} \cap K = \begin{cases} f & \text{if} \quad f \in K \\ \emptyset & \text{otherwise} \end{cases}$$

is a rational relation if and only if $K \in \operatorname{Rat} A^*$.

We have $\mathsf{Dom}\,\iota_K = \mathsf{Im}\,\iota_K = K$ and the condition is necessary by Property 1.2. Let E be a rational expression over A^* that denotes K. We can view E as a rational expression over $\mathbf{A}$, the set of variables in bijection with A, and K is the result of E under the interpretation τ which maps each a of $\mathbf{A}$ to $\mathsf{a}\tau = a$. *cf. § I.4.1.2, p. 127*

It should be obvious that the graph $\widehat{\iota_K}$ is the result of E under the interpretation τ' which maps each a of $\mathbf{A}$ to $\mathsf{a}\tau' = (a, a)$. □

We will see many other examples in the exercises.

1.1.2 Rational relations between arbitrary monoids

Like that of a rational set, the notion of a rational relation extends to monoids that are not necessarily free. In what follows, M and N are two *a priori* arbitrary monoids.

Definition 1.2 A relation $\tau \colon M \to N$ is *rational* if its graph $\widehat{\tau}$ is a rational subset of $M \times N$. □

Properties 1.1 and 1.2 extend naturally to the case of relations from M to N; that is, the inverse of a rational relation from M to N is a rational relation from N to M; the domain and the image of a rational relation are rational sets of M and N respectively. Corollary II.1.4 then gives: *cf. p. 223*

Property 1.3 *Let τ be a rational relation from M to N. There exists a finitely generated submonoid M' of M and a finite generated submonoid N' of N such that $\widehat{\tau}$ is contained in $M' \times N'$.* ■

Example 1.1 (continued) **Morphisms.** If M is a *finitely generated* monoid (by a generating set G), *every morphism* $\alpha \colon M \to N$ *is a rational relation*: its graph $\widehat{\alpha}$ has the expression

$$\widehat{\alpha} = \{(m, m\alpha) \mid m \in M\} = \left(\bigcup_{m \in G} (m, m\alpha) \right)^*$$

which is a rational set since G is finite. It follows that an *inverse morphism* is a *rational relation.* □

Example 1.4 Rectangles. If R is a rational subset of M and S is a rational subset of N, then the relation from M to N whose graph is $R \times S$ is a rational relation; the same is true of a relation whose graph is a finite union of such 'rectangles' of $M \times N$. □

Example 1.5 Product. The map $\pi\colon A^*\times A^* \to A^*$ defined, for all f and g in A^*, by $(f,g)\pi = f\,g$ is a rational relation:

$$\widehat{\pi} = \left(\bigcup_{a\in A}\big((a,1_{A^*}),a\big)\right)^* \left(\bigcup_{a\in A}\big((1_{A^*},a),a\big)\right)^* \quad . \qquad \square$$

This last example suggests that we can just as well consider relations not only between two monoids but among an arbitrary (finite) number of monoids M_1, M_2, ... , M_k; such a relation, or rather its graph, is then a set of k-tuples. Thus, the above relation π is a relation on $A^*\times A^*\times A^*$ and

$$\widehat{\pi} = \{(f,g,f\,g) \mid f,g \in A^*\} \quad .$$

The definitions and results of this entire subsection thus extend directly, *mutatis mutandis*, to relations on an arbitrary number of monoids, free or otherwise.

Exercises Examples of rational relations. In the following exercises, A and B are two finite alphabets; K and L are two rationals of A^* and B^* respectively. Show that the following relations from A^* to A^*, or from A^* to B^*, are rational.

• 1.1 **Operations on constants.**
(a) The constant relation γ_L defined by $\forall f \in A^* \quad f\gamma_L = L$.
(b) The left (or right) product with a fixed rational: $f \longmapsto K\,f$.
(c) The left (or right) quotient by a fixed rational: $f \longmapsto K^{-1}f$.

• 1.2 **Factors and subwords.** (a) The relation which makes a word in A^* correspond to the set of its prefixes, respectively of its suffixes.

for factors, see Exam. 1.2

(b) The relation which makes a word A^* correspond to the set of its subwords.

cf. Sec. 0.3, p. 18

• 1.3 **Orders: predecessors and successors.** Suppose that the alphabet $A = \{a_1,\dots,a_n\}$ is *totally ordered*: $a_1 < a_2 < \cdots < a_n$, an order which we extend to A^* by the definition of the *lexicographic order* $\preccurlyeq$:

cf. Exer. 0.3.10, p. 24

$$f \preccurlyeq g \iff \begin{cases} g = f\,h & \text{with} \quad h \in A^* \quad \text{or} \\ f = u\,a_i\,v\,, \quad g = u\,a_j\,w & \text{with} \quad i < j \ . \end{cases}$$

(a) Show that the relations which make a word of A^* correspond respectively to its predecessor, successor, set of predecessors set of successors in this lexicographic order are rational.

(b) Do the same considering the *radix order*[4] on A^*, $\sqsubseteq$:

$$f \sqsubseteq g \iff \begin{cases} |f| < |g| & \text{or} \\ |f| = |g| & \text{and} \quad f \preccurlyeq g \ . \end{cases}$$

1.4 **Distances.** We define the *Hamming distance* on A^*: two words $f = f_1\,f_2\cdots f_n$ and $g = g_1\,g_2\cdots g_n$ of the same length n have a Hamming distance between them of k if $f_i \neq g_i$ for exactly k values of the index i.

Show that the relations which make a word in A^* correspond to the set of words of distance k (and of distance less than k, and distance greater than k) from this word are rational.

[4] Also called *genealogical order*, or, in French, *ordre militaire* (*military order*).

• 1.5 **Shuffle and infiltration.** Recall that *shuffle* and *infiltration* are two relations from $A^* \times A^*$ to A^*: *cf. Def. I.1.7 and Def. I.1.8, p. 66*

$$f \between g = \{h \in A^* \mid \exists I, J \subset [|h|]\,,\ I \cup J = [|h|]\,,\ I \cap J = \emptyset \qquad f = h_I \quad \text{and} \quad g = h_J\}$$
$$f \uparrow g = \{h \in A^* \mid \exists I, J \subset [|h|]\,,\ I \cup J = [|h|] \qquad f = h_I \quad \text{and} \quad g = h_J\}$$

(where h_I denotes the subword of h obtained by deleting those letters whose index is not in I).

Show that shuffle and infiltration are *rational* relations from $A^* \times A^*$ to A^*.

• 1.6 Show that the function $\theta_1 \colon \{a\}^* \to \{x\}^*$, defined by

$$(a^{2p})\theta_1 = x^{4p} \qquad \text{and} \qquad (a^{2p+1})\theta_1 = x^{2p+1}$$

is a rational relation.

• 1.7 Let $\tau \colon A^* \times A^* \to B^*$ be a rational relation; show by means of an example that the 'diagonal' of τ, $\delta_\tau \colon A^* \to B^*$, defined by $f\delta_\tau = (f, f)\tau$ for all f in A^*, is not necessarily a rational relation.

1.2 Realisation by automata

The set $(A \times \{1_{B^*}\}) \cup (\{1_{A^*}\} \times B)$ is a generating set of $A^* \times B^*$; it is its *minimal* generating set. We will say that an automaton over $A^* \times B^*$ is *in normal form* if all its transitions are labelled with elements of this set. The Fundamental Theorem of Finite Automata,[5] applied to automata over the monoid $A^* \times B^*$, gives: *cf. Th. II.1.2, p. 227*

Theorem 1.1 *A relation $\tau \colon A^* \to B^*$ is rational if and only if its graph is the behaviour of a finite automaton over $A^* \times B^*$ in normal form.* ■

Another consequence of the same result is that the behaviour of a finite automaton over $A^* \times B^*$ is a rational relation; or, in other words, that every finite automaton over $A^* \times B^*$ is equivalent to a finite automaton over $A^* \times B^*$ in normal form.

Thus, the rational relations from A^* to B^* are the relations realised by the finite automata over $A^* \times B^*$. Despite an unfortunate collision with other areas of mathematics – and it will get worse in the case of functions – we stick to the shorter first name. However, our lexical problems are not over. We call an automaton over a product of free monoids a *transducer (over $A^* \times B^*$)*. In itself, this new word adds nothing: a transducer is an automaton, nothing more; it is usage which justifies it.[6] Equipped with this terminology, we shall be able to speak in the same sentence of a transducer and an automaton and keep their input monoids implicit ($A^* \times B^*$ for the one, A^* for the other) without descending into incomprehensibility (or in any case, complete ambiguity).

The *underlying input automaton* of a transducer over $A^* \times B^*$ is the automaton over A^* obtained by taking the projection of each label on its first component.

[5] This instance of the Fundamental Theorem is due to Elgot and Mezei (*cf.* [84]).

[6] Elgot & Mezei, *art. cit.*, call rational relations *transductions* and the automata that realise them *transducers*.

Example 1.3 (continued) Let $K \in \operatorname{Rat} A^*$ and let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be an automaton which accepts K. It should be clear that the transducer $\mathcal{I}_K$ over $A^* \times A^*$,

$$\mathcal{I}_K = \langle Q, A^* \times A^*, F, I, T \rangle \qquad \text{and} \qquad F = \{(p, (a, a), q) \mid (p, a, q) \in E\} \ ,$$

is such that $\widehat{\iota}_K = |\mathcal{I}_K|$. The underlying input automaton of $\mathcal{I}_K$ is $\mathcal{A}$. □

Example 1.6 We saw with the automaton $\mathcal{P}_2$ that we can compute the remainder of the integer division by 3 of a number written in binary (in the alphabet $\{a, b\}$). It is not difficult to transform this automaton into a transducer $\mathcal{Q}_2$ which computes the *quotient* of a number divided by 3 (*cf.* Figure 1.1). For example, $(b\,a\,b\,a\,b, a\,a\,b\,b\,b)$ is a pair accepted by $\mathcal{Q}_2$, $\overline{b\,a\,b\,a\,b} = 21$ and $\overline{a\,a\,b\,b\,b} = 7$. The underlying input automaton of $\mathcal{Q}_2$ is $\mathcal{P}_2$. □

cf. Exam. I.1.2, p. 52

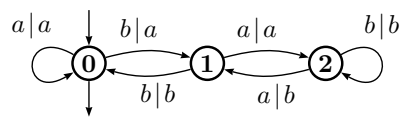


Figure 1.1: The transducer $\mathcal{Q}_2$ computes the quotient by 3 of a number written in binary.

Finding the *complement of the identity* is another computation to which the use of transducers is well suited.

Proposition 1.1 *The relation 'complement of the identity' from a free monoid to itself (that is, the relation whose graph is the complement of the diagonal $\mathcal{D}^*$) is a rational relation.*

Proof. Let $\chi_A \colon A^* \to A^*$ be the complement of the identity on A^*:

$$\widehat{\chi_A} = \{(f, g) \mid f, g \in A^* \quad f \neq g\} \ ,$$

and consider the transducer in Figure 1.2. We easily convince ourselves that if a successful computation ends in p (resp. in s), its label is of the form $(f\,h, f)$ (resp. $(f, f\,h)$), with h in A^+, and that if it ends in t, its label is of the form $(f\,x\,g, f\,y\,h)$, with x and y in A and distinct. ■

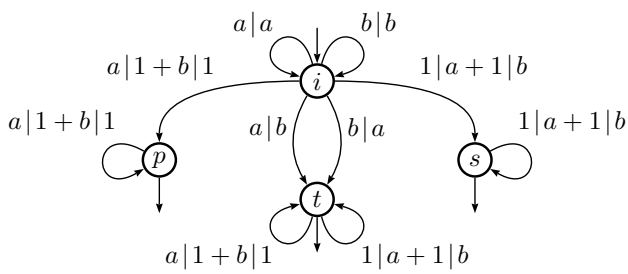


Figure 1.2: A transducer whose behaviour is the complement of the identity

The Fundamental Theorem of Finite Automata, applied to automata over $M \times N$, gives:

Theorem 1.1 *A relation $\tau\colon M \to N$ is rational if and only if its graph is the behaviour of a finite automaton over $M \times N$. It is always possible to choose an automaton whose transitions have labels in G, where G is a generating set of $M \times N$.* ■

Exercises

- 1.8 Let $\varphi_1\colon \{a,b,c\}^* \to \{x,y\}^*$ be the morphism defined by

$$a\varphi_1 = x \ , \quad b\varphi_1 = y\,x \ , \quad c\varphi_1 = x\,y \ .$$

Give a transducer in normal form which realises φ_1.

- 1.9 Give a transducer in normal form which realises the substitution σ_1 defined by $a\sigma_1 = a^+$ and $b\sigma_1 = b^+$.

- 1.10 Give a transducer which realises:

(a) the relation 'lexicographic order' (that is, which relates a word to all those which are greater than it);

(b) the relation 'radix order';

(c) the relation α_1 from $\{a,b\}^*$ to itself which in each word replaces every second a by b.

1.11 Verify that the transducer $\mathcal{Q}_2$ of Example 1.6 indeed computes the quotient in the division by 3 of a number written in binary.

1.3 Realisation by morphisms

We return to the definition of rational relations by their graphs. To make it more useful, and in particular in order to determine the effect of such relations on rational sets and on the nature of their composition product, we prove one of their characteristic properties, a procedure for their construction. This characterisation still belongs to the static point of view, and holds for rational relations between arbitrary monoids. It allows us to prove the two pivotal theorems.

1.3.1 Realisation

If $\varphi\colon Z^* \to M$ and $\psi\colon Z^* \to N$ are two morphisms, then the map

$$\varphi\times\psi\colon Z^* \longrightarrow M\times N \qquad \text{defined by} \qquad (u)\,\varphi\times\psi = (u\varphi, u\psi) \ ,$$

for all u in Z^*, is a morphism.

Lemma 1.2 *Let $\varphi\colon Z^* \to M$ and $\psi\colon Z^* \to N$ be two morphisms and K an arbitrary subset of Z^*. The graph of the relation $\tau\colon M \to N$ defined by*

$$\forall m \in M \qquad m\tau = (m\varphi^{-1} \cap K)\psi : \quad \textit{that is,} \qquad \tau = \varphi^{-1} \circ \iota_K \circ \psi \ ;$$

is the set $\widehat{\tau} = (K)\,\varphi\times\psi$.

Proof. We have the sequence of equalities:

$$\begin{aligned}(K)\,\varphi\times\psi &= \{(u\varphi, u\psi) \mid u \in K\} \;=\; \{(m,n) \mid \exists u \in K \quad u\varphi = m, u\psi = n\}\\ &= \{(m,n) \mid \exists u \in m\varphi^{-1} \cap K \quad u\psi = n\}\\ &= \{(m,n) \mid n \in (m\varphi^{-1} \cap K)\psi\} \;=\; \widehat{\tau}\ .\end{aligned}$$

■

The promised characterisation is therefore:

Theorem 1.2 [Nivat] *A relation $\tau\colon M \to N$ is rational if and only if there exists a finite alphabet Z, two morphisms $\varphi\colon Z^* \to M$ and $\psi\colon Z^* \to N$, and a subset K of* $\operatorname{Rat} Z^*$ *such that*

$$\text{(i)}\quad \widehat{\tau} = (K)\,\varphi\times\psi \qquad \textit{or} \qquad \text{(ii)}\quad \tau = \varphi^{-1}\circ\iota_K\circ\psi\ .$$

If M (resp. N) is a free monoid, φ (resp. ψ) can be chosen to be alphabetic. If $M = A^*$ and $N = B^*$ are two disjoint *free monoids, we can choose*

$$Z = A \cup B\ , \qquad \varphi = \pi_A \qquad \textit{and} \qquad \psi = \pi_B\ ,$$

where π_A and π_B are the projections of $(A \cup B)^*$ onto A^* and B^* respectively.

Lemma 1.2 states that the two conditions (i) and (ii) are equivalent. We shall prove (i), but we will use the theorem in the form (ii).

Prop. II.1.7 (a), p. 223 *Proof.* The condition is sufficient: if K is rational, then so is $(K)\,\varphi\times\psi$.

The condition is necessary. Let $\widehat{\tau}$ be a rational subset of $M\times N$. By Property 1.3, p. 527 we can assume that M and N are finitely generated.

Let A' and B' be two *disjoint* finite alphabets and $\varphi\colon A'^* \to M$ and $\psi\colon B'^* \to N$ be two *surjective* morphisms. Set $Z = A' \cup B'$ and extend φ and ψ to all of Z^* by setting $a\psi = 1_N$ for all a in A' and $b\varphi = 1_M$ for all b in B'. The morphism

$$\varphi\times\psi\colon Z^* \longrightarrow M\times N\ ,$$

Prop. II.1.7 (b), p. 223 is *surjective* since φ and ψ are surjective and A' and B' are disjoint. Hence, there exists a rational K in Z^* such that $(K)\,\varphi\times\psi = \widehat{\tau}$.

If $M = A^*$ is a free monoid, we take A' to be in bijection with A; if $N = B^*$, we take B' to be in bijection with B. We go to the trouble of specifying that A' *is in bijection* with A (resp. that B' *is in bijection* with B) and not that $A' = A$ (resp. that $B' = B$) to be sure that A' and B' are *disjoint*, even if $A = B$. ■

$$\begin{tikzcd}
(A' \cup B')^* \arrow[r, "\iota_K"] \arrow[d, "\varphi"'] & (A' \cup B')^* \arrow[d, "\psi"] \\
M \arrow[r, "\tau"'] & N
\end{tikzcd}$$

Figure 1.3: Realisation by morphisms of rational relations (Theorem 1.2)

Example 1.7 Let $\chi\colon \{a,b\}^* \to \{a,b\}^*$ be the morphism which changes the a's to b's and the b's to a's. We set $A' = \{a', b'\}$ and $A'' = \{a'', b''\}$, $Z = A' \cup A''$ and $K = (a'b'' \cup b'a'')^*$. The morphisms φ and ψ are the composition of the projections onto A' and A'' and of the deletion of the $'$ and the $''$ respectively. □

1.3.2 Evaluation Theorem

Theorem 1.3 *Let τ be a rational relation from M to N. The image under τ of a recognisable subset of M is a rational subset of N.*

Proof. Let Z be an alphabet, K be in $\operatorname{Rat} Z^* = \operatorname{Rec} Z^*$, and $\varphi\colon Z^* \to M$ and $\psi\colon Z^* \to N$ be the realisation by morphisms of τ given by Theorem 1.2. Let R be in $\operatorname{Rec} M$. We have successively:

$R\varphi^{-1} \in \operatorname{Rec} Z^*$, since the recognisable subsets are closed under inverse morphism; *Cor. II.2.2, p. 237*

$(R\varphi^{-1}) \cap K \in \operatorname{Rec} Z^*$, since the recognisable subsets are closed under intersection and also $(R\varphi^{-1}) \cap K \in \operatorname{Rat} Z^* = \operatorname{Rec} Z^*$ by Kleene's Theorem, since Z is finite; *Cor. II.2.8, p. 244*

$((R\varphi^{-1}) \cap K)\psi \in \operatorname{Rat} N$, since the rational subsets are closed under morphism. ■ *Prop. II.1.7, p. 223*

In the case where τ is a relation between free monoids, Theorem 1.3 becomes:

Corollary 1.3 *The image of a rational language in A^* under a rational relation from A^* to B^* is a rational language in B^*.* ■

This fundamental property of rational relations is far from being *characteristic* of rational relations. We have already seen several examples of functions and relations which preserve rationality yet which are not rational relations and we shall see several more.[7]

Example 1.8 The image of every rational of $\widetilde{A}^*$ under the Dyck reduction:

$$\delta_A\colon \widetilde{A}^* \to \widetilde{A}^* \ ,$$

is rational, but the relation δ_A is not itself rational. If it were, its inverse would also be and $(1_{\widetilde{A}^*})\delta_A^{-1} = D_A^*$ would be a rational language. □ *Cor. II.6.10, p. 310*

Remark 1.2 We could also recall at this point that a map which relates every rational language in A^* to a rational language in B^* is not only not necessarily a rational relation, but is not even necessarily a *relation from A^* to B^** in the sense that it is not an *additive* map from $\mathfrak{P}(A^*)$ to $\mathfrak{P}(B^*)$, defined by its graph, a subset of $A^* \times B^*$.

One example is the *complement operation*. For another, consider $\sigma\colon B^* \to A^*$, an arbitrary substitution; then the map $\breve{\sigma}$ from $\mathfrak{P}(A^*)$ to $\mathfrak{P}(B^*)$, defined by *cf. Cor. II.2.18, p. 253 and Th. II.4.5, p. 291*

$$\forall L \in \mathfrak{P}(A^*) \qquad L\breve{\sigma} = \{f \in B^* \mid f\sigma \subseteq L\} \ ,$$

is such that $L\breve{\sigma}$ is rational when L is rational but *is not a relation*. □

[7] J. H. Conway calls a relation which preserves rationality a *regulator*, and one whose inverse is also a regulator, a *biregulator* (*cf.* [66]).

1.3.3 Composition Theorem

A fundamental property of rational relations between free monoids is that they are closed under composition.

Theorem 1.4 [**Elgot–Mezei**] *Let $\theta\colon A^* \to B^*$ and $\sigma\colon B^* \to C^*$ be two rational relations. Then the composition $\theta\sigma\colon A^* \to C^*$ is a rational relation.*

This result can be proved in two different ways, which correspond in fact to weaker but different assumptions. We give the first here and touch on the second in Subsection 1.5. We return to both in Section 3. We shall first prove:

Theorem 1.5 *Let $\theta\colon M \to B^*$ and $\sigma\colon B^* \to P$ be two rational relations. Then the composition $\theta\sigma\colon M \to P$ is a rational relation.*

We start by proving a lemma.

Lemma 1.4 *Let A, B and C be three disjoint alphabets, and ψ_A, ψ_C, φ_A and φ_C, the projections*[8]

$$\psi_A\colon (A \cup B \cup C)^* \to (B \cup C)^* \quad , \quad \psi_C\colon (A \cup B \cup C)^* \to (A \cup B)^* \quad ,$$
$$\varphi_A\colon (A \cup B)^* \to B^* \quad \text{and} \quad \varphi_C\colon (B \cup C)^* \to B^* \quad .$$

We have:
$$\varphi_A \circ \varphi_C^{-1} = \psi_C^{-1} \circ \psi_A \quad .$$
More precisely, setting $Z = A \cup B \cup C$ and $T = Z^* \setminus Z^* C A Z^*$, we have:
$$\varphi_A \circ \varphi_C^{-1} = \psi_C^{-1} \circ \iota_T \circ \psi_A \quad .$$

The two equations of Lemma 1.4 are shown in the commutative diagrams of Figure 1.4.

$$\begin{tikzcd}
 & (A \cup B \cup C)^* \arrow[dl, "\psi_C"'] \arrow[dr, "\psi_A"] & \\
(A \cup B)^* \arrow[rr, dashed, "\psi_C^{-1} \circ \psi_A = \varphi_A \circ \varphi_C^{-1}"] \arrow[dr, "\varphi_A"'] & & (B \cup C)^* \arrow[dl, "\varphi_C"] \\
 & B^* &
\end{tikzcd}
\qquad
\begin{tikzcd}
(A \cup B \cup C)^* \arrow[r, "\iota_T"] \arrow[d, "\psi_C"'] & (A \cup B \cup C)^* \arrow[d, "\psi_A"] \\
(A \cup B)^* \arrow[r, dashed, "\psi_C^{-1} \circ \iota_T \circ \psi_A = \varphi_A \circ \varphi_C^{-1}"] \arrow[dr, "\varphi_A"'] & (B \cup C)^* \arrow[dl, "\varphi_C"] \\
 & B^*
\end{tikzcd}$$

Figure 1.4: The two equations of Lemma 1.4

Proof of Lemma 1.4. By definition, the graph of $\varphi_A \circ \varphi_C^{-1}$ is

$$\widehat{\varphi_A \circ \varphi_C^{-1}} = \{(u, v) \mid u \in (A \cup B)^*\,,\ v \in (B \cup C)^* \quad u\varphi_A = v\varphi_C\} \quad .$$

Let w be in $(A \cup B \cup C)^*$; then $w\psi_C\varphi_A = w\psi_A\varphi_C$, hence $(w\psi_C, w\psi_A) \in \widehat{\varphi_A \circ \varphi_C^{-1}}$.

[8] To improve legibility, I have marked the composition of morphisms explicitly with a $\circ$.

Conversely, let (u,v) be in $\widehat{\varphi_A \circ \varphi_C^{-1}}$; that is, $u\varphi_A = v\varphi_C = f$ is in B^*. We show, by induction on the length of f, that there exists w in T such that

$$w\psi_C = u \qquad \text{and} \qquad w\psi_A = v \ .$$

If $f = 1_{B^*}$, then $u \in A^*$ and $v \in C^*$ and the word $u\,v$ is what we want:

$$(u\,v)\psi_C = u \ , \quad (u\,v)\psi_A = v \qquad \text{and} \qquad u\,v \notin Z^*\,C\,A\,Z^* \ .$$

Suppose that $f = b\,f$; then $u = h\,b\,u'$, with h in A^* and $u'\varphi_A = f'$, and $v = k\,b\,v'$, with k in C^* and $v'\varphi_C = f'$. By the induction hypothesis there exists w' such that $w'\psi_C = u'$ and $w'\psi_A = v'$; the word $h\,k\,b\,w'$ is what we want:

$$(h\,k\,b\,w')\psi_C = u \ , \quad (h\,k\,b\,w')\psi_A = v \qquad \text{and} \qquad h\,k\,b\,w' \notin Z^*\,C\,A\,Z^* \ .$$

We have thus shown $(Z^*)\psi_C \times \psi_A \subseteq \widehat{\varphi_A \circ \varphi_C^{-1}} \subseteq (T)\psi_C \times \psi_A$. As $(T)\psi_C \times \psi_A$ is trivially contained in $(Z^*)\psi_C \times \psi_A$, we have

$$(T)\psi_C \times \psi_A = (Z^*)\psi_C \times \psi_A = \widehat{\varphi_A \varphi_C^{-1}} \ .$$

Lemma 1.2 then implies $\varphi_A \circ \varphi_C^{-1} = \psi_C^{-1} \circ \psi_A = \psi_C^{-1} \circ \iota_T \circ \psi_A$. ■

It will be useful to have stated as a property and as commutative diagrams two trivial equivalences: let E and F be two sets, K a subset of F, and $\alpha\colon E \to F$ a *function*. For all x in E and all y in F, we obviously have

$$x \in K\alpha^{-1} \iff x\alpha \in K \qquad \text{and} \qquad y\alpha^{-1} \in K\alpha^{-1} \iff y \in K \ ,$$

which can be rewritten:

Property 1.4 *With the above notations,*

$$\alpha \circ \iota_K = \iota_{(K)\alpha^{-1}} \circ \alpha \qquad \textit{and} \qquad \alpha^{-1} \circ \iota_{(K)\alpha^{-1}} = \iota_K \circ \alpha^{-1} \ ;$$

that is, the two diagrams of Figure 1.5 are commutative. ■

$$\begin{tikzcd} E \arrow[r, "\iota_{(K)\alpha^{-1}}"] \arrow[d, "\alpha"'] & E \arrow[d, "\alpha"] \\ F \arrow[r, "\iota_K"'] & F \end{tikzcd} \qquad \begin{tikzcd} E \arrow[r, "\iota_{(K)\alpha^{-1}}"] & E \\ F \arrow[r, "\iota_K"'] \arrow[u, "\alpha^{-1}"] & F \arrow[u, "\alpha^{-1}"'] \end{tikzcd}$$

Figure 1.5: The two equations of Property 1.4

Proof of Theorem 1.4. It is enough now to use the above results, written as commutative diagrams, to construct, as in Figure 1.6, one large diagram, commutative because each of the constituent subdiagrams is, on which we 'read' the result.

We start with what we are given: the two relations θ and σ, represented by a double arrow. Theorem 1.2 gives the two diagrams with solid arrows, Lemma 1.4 gives the diagram in dashed lines, and Property 1.4 the two diagrams in dotted lines. Hence we see that

$$\theta \circ \sigma = (\psi_C \circ \chi)^{-1} \circ \iota_{(K\psi_C^{-1} \cap T \cap L\psi_A^{-1})} \circ (\psi_A \circ \omega) \ ,$$

which is a rational relation by the forward direction of Theorem 1.2. ■

$$\begin{tikzcd}
(A\cup B\cup C)^* \arrow[r, dotted, "\iota_{(K)\psi_C^{-1}}"] \arrow[d, dotted, "\psi_C"'] & (A\cup B\cup C)^* \arrow[r, dashed, "\iota_T"] \arrow[d, dashed, "\psi_C"'] & (A\cup B\cup C)^* \arrow[r, dotted, "\iota_{(L)\psi_A^{-1}}"] \arrow[d, dashed, "\psi_A"] & (A\cup B\cup C)^* \arrow[d, dotted, "\psi_A"] \\
(A\cup B)^* \arrow[r, "\iota_K"] \arrow[d, "\chi"'] & (A\cup B)^* \arrow[dr, "\varphi_A"] & (B\cup C)^* \arrow[r, "\iota_L"] \arrow[dl, "\varphi_C"'] & (B\cup C)^* \arrow[d, "\omega"] \\
M \arrow[rr, Rightarrow, "\theta"'] & & B^* \arrow[rr, Rightarrow, "\sigma"'] & P
\end{tikzcd}$$

Figure 1.6: The proof of Theorem 1.4

cf. p. 530 By Proposition 1.1, we then deduce:

Corollary 1.5 *If M is a finitely generated monoid and B is a finite alphabet, the complement of a rational* function *from M to B^* is a rational relation.*

Proof. Write χ_B for the complement of the identity on B^*. The complement of the graph of a rational function $\sigma\colon M \to B^*$ can be written as

$$\complement\widehat{\sigma} = \left(\complement\mathsf{Dom}\,\sigma \times B^*\right) \cup \left(\widehat{\sigma \circ \chi_B}\right) \quad . \qquad \blacksquare$$

Remark 1.3 Let $\tau\colon A^* \to A^*$ be a rational relation. We have just shown that $\tau^2 = \tau \circ \tau$, and also, by induction, $\tau^n = \tau \circ \tau^{n-1}$, are rational relations from A^* to itself. However, the graph of τ^2 *is not* the square of the graph of τ (as a subset of $A^* \times A^*$) and the graph of τ^n is not the nth power of the graph of τ. To avoid confusing the iteration of a relation and the star of its graph, we shall use the following notations. □

cf. Rem. 0.1.1, p. 12

cf. Prop. 0.1.2 and Note 0.7, p. 13

Definition 1.3 Let τ be a relation from A^* to itself. We write

$$\tau^{\leqslant n} = \bigcup_{i=0}^{i=n} \tau^i \ , \qquad \overset{\star}{\tau} = \bigcup_{i=0}^{i=\infty} \tau^i \ , \qquad \text{and} \qquad \overset{+}{\tau} = \bigcup_{i=1}^{i=\infty} \tau^i \ .$$

The relation $\overset{\star}{\tau}$ is the reflexive and transitive closure of τ and is called the *iteration* of τ, while the relation $\overset{+}{\tau}$ is the transitive closure of τ and is called the *proper iteration* of τ. □

Fact 1.1 *If $\tau\colon A^* \to A^*$ is a rational relation, $\tau^{\leqslant n}$ is a rational relation but, in the general case, neither $\overset{\star}{\tau}$ nor $\overset{+}{\tau}$ is a rational relation.*

cf. Sec. II.6.2, p. 307

Example 1.9 Let $A = \{a, b\}$; we write $\underset{\Pi}{\longmapsto}$ for the regular closure in $A^* \times A^*$ of the pair $(a\,b, 1_{A^*})$; that is, for f and g in A^*, we have

$$f \underset{\Pi}{\longmapsto} g \quad \Longleftrightarrow \quad \exists u, v \in A^* \qquad f = u\,a\,b\,v \quad \text{and} \quad g = u\,v \ ;$$

in other words, g is obtained from f by deleting a factor $a\,b$. The relation $\underset{\Pi}{\longmapsto}$ is rational, but its iteration $\overset{\star}{\underset{\Pi}{\longmapsto}}$ is not:

$$\{f \in A^* \mid f \overset{\star}{\underset{\Pi}{\longmapsto}} 1_{A^*}\} \cap a^*b^* = \{a^n b^n \mid n \in \mathbb{N}\} \ . \qquad \square$$

1.3.4 Star Lemma

From their realisation by morphisms we also obtain a star lemma for rational relations.

Proposition 1.6 *Let $\theta\colon A^* \to B^*$ be a rational relation. There exists an integer N such that, for every pair (f,g) in the graph of θ the sum of whose lengths is greater than N, there exists a factorisation*

$$(f,g) = (h,k)(u,v)(s,t)$$

such that (i) $1 \leqslant |u| + |v| < N$ *and* (ii) $(h,k)(u,v)^*(s,t) \subseteq \widehat{\theta}$.

Proof. We can assume that A and B are disjoint. Let $\theta = \varphi^{-1} \circ \iota_K \circ \psi$ be the realisation by morphisms of θ, where $Z = A \cup B$ and φ and ψ are the projections of Z^* on to A^* and B^* respectively. The set K is a rational of Z^* such that $\widehat{\theta} = (K)\,\varphi \times \psi$. Thus,

$$(f,g) \in \widehat{\theta} \Longleftrightarrow \exists w \in K \quad f = w\varphi \quad \text{and} \quad g = w\psi \qquad \text{and} \qquad |w| = |f| + |g| \ .$$

The Star Lemma applied to K entails the existence of an integer N such that, for all w in K of length greater than N, there exists a factorisation $w = x\,y\,z$ such that $1 \leqslant |y| < N$ and $x\,y^*\,z \subseteq K$, from which we deduce

$$(x\,\varphi \times \psi)(y\,\varphi \times \psi)^*(z\,\varphi \times \psi) \subseteq \widehat{\theta}$$

which is the result we seek with

$$(h,k) = x\,\varphi \times \psi \quad , \quad (u,v) = y\,\varphi \times \psi \quad \text{and} \quad (s,t) = z\,\varphi \times \psi \quad . \qquad \blacksquare$$

Example 1.10 The transposition, or 'mirror image' function, $\rho\colon A^* \to A^*$,

$$(a_1\,a_2 \cdots a_n)\rho = a_n\,a_{n-1} \cdots a_1 \quad ,$$

is not a rational relation. Suppose the contrary. Take A' in bijection with and disjoint from A. There exists K in $\mathrm{Rat}\,(A' \cup A)^*$ such that $\widehat{\rho} = (K)\,\varphi \times \psi$ (in the notation of the proof above).

We consider the pair $(a^{2N}b^{2N}, b^{2N}a^{2N}) \in \widehat{\rho}$, where N is the integer given by Proposition 1.6, and, also given by the proposition, a factorisation

$$(a^{2N}b^{2N}, b^{2N}a^{2N}) = (h,k)(u,v)(s,t) \ .$$

It corresponds to a factorisation $w = x\,y\,z$ of a word of K.

Up to symmetry of reasoning, we can assume that $|u|_b > 0$. This implies on one hand that $|h| > N$ and on the other, since ρ leaves the number of occurrences of a and b unchanged, that $|v|_b > 0$ and hence that $|k|_a = 0$.

There therefore exists a factor t of x of length at least equal to N whose projection by φ contains no letter b. The Star Lemma *applied to the language* K allows us to find in this factor t a factor l which we can iterate and such that $|l\varphi|_b = 0$ and $|l\psi|_a \leqslant |k|_a = 0$. Contradiction.

Thus, transposition is another example of a relation which preserves rationality (and hence whose inverse also preserves rationality) without being rational. $\square$ *cf. Prop. I.1.1, p. 58*

Exercises & Problems

- 1.12 **Quotient.** Show that the left (and right) quotient of a fixed rational, $f \longmapsto f^{-1}K$, for $f \in A^*$ and $K \in \mathrm{Rat}\, A^*$, is a rational relation from A^* to itself.

- 1.13 **Büchi systems.** Let K be a *Büchi system* on A^*. Show that the relation β_K from A^* to itself, defined by this system, $f\beta_K = \{g \mid f \overset{\star}{\underset{\mathcal{B}\ K}{\longmapsto}} g\}$, is rational.

cf. Sec. II.6.4, p. 319

- 1.14 Verify that the adjacent commutative diagram (that is, $\alpha\gamma = \beta\delta$) does not imply in general that $\gamma\delta^{-1} = \alpha^{-1}\beta$.

$$\begin{tikzcd} & E \arrow[dl, "\alpha"'] \arrow[dr, "\beta"] & \\ F \arrow[dr, "\gamma"'] & & G \arrow[dl, "\delta"] \\ & H & \end{tikzcd}$$

- 1.15 Give an example of rational relations $\theta\colon A^* \to B^* \times C^*$ and $\sigma\colon B^* \times C^* \to D^*$ where $\theta \circ \sigma$ is not a rational relation.

- 1.16 **On the Star Lemma.** Show that we *cannot strengthen* the statement of Proposition 1.6 to the following proposition, analogous to that which we can make from the Star Lemma for rational *languages*:

cf. Lem. I.1.13, p. 71

There exists an integer N such that for all (r, w) in $\widehat{\theta}$ and for every factorisation

$$(r, w) = (r', w')(f, g)(r'', w'')$$

such that $|f| + |g| > N$ (or even $|f| > N$ and $|g| > N$), there exists a factorisation

$$(f, g) = (h, k)(u, v)(s, t)$$

such that (i) $1 \leqslant |u| + |v| < N$ *and* (ii) $(r', w')(h, k)(u, v)^*(s, t)(r'', w'') \subseteq \widehat{\theta}$.

- 1.17 **Conjugacy.** Show that the relation $\mathsf{Conj}\colon A^* \to A^*$ which maps a word f on to the set of its *conjugates* $\mathsf{Conj}(f) = \{v\,u \mid u, v \in A^* \quad u\,v = f\}$ is not a rational relation.

- 1.18 **Relation between commutative monoids.** Show that a relation $\tau\colon M \to N$ between commutative monoids is rational if and only if there exists an integer k, two morphisms $\varphi\colon \mathbb{N}^k \to M$ and $\psi\colon \mathbb{N}^k \to N$, and a rational K of $\mathbb{N}^k$ such that

$$\text{(i)}\ \ \widehat{\tau} = (K)\,\varphi \times \psi \qquad \text{or} \qquad \text{(ii)}\ \ \tau = \varphi^{-1} \circ \iota_K \circ \psi \ .$$

- 1.19 **Iterations of rational relations.** Determine whether the iteration of each of the following rational relations (from $\{a, b\}^*$ to itself) is rational or not:
  - (a) $\varphi : f\varphi$ is the set of words obtained by replacing *one* factor $a\,a$ of f by 1_{A^*} ;
  - (b) $\chi : f\chi$ is the set of words obtained by replacing *one* factor $a\,b$ of f by $b\,a$;
  - (c) $\psi : f\psi$ is the set of words obtained by replacing *one* factor $a\,b\,b$ of f by $b\,a\,a$;
  - (d) $\omega : f\omega$ is the set of words obtained by replacing *one* factor $a\,b\,b$ of f by $b\,a$.

- 1.20 **Recurrent and adult words.** Let $\sigma\colon A^* \to A^*$ be a substitution. We say that a word g is a *descendant* of f if $g \in (f)\sigma^n$ for some positive n; that is, g is in $\overset{+}{\sigma}$, and that a word f is *recurrent*[9] for σ if it is a descendant of any of its own descendants; that is,

$$f \in \bigcap_{g \in (f)\overset{+}{\sigma}} (g)\,\overset{+}{\sigma} \ .$$

A word is *adult*[10] if $f\sigma = \{f\}$. Prove:

Proposition 1.7 *If σ is a rational substitution, the relation that associates a word in A^* with the set of its recurrent (resp. adult) descendants is rational.*

---

[9] A term which has nothing to do with the terminology used in symbolic dynamical systems.

[10] A term that comes from the theory of Lindenmayer systems.

1.4 Recognisable relations

As at the beginning of the section, the following definition comes naturally.

Definition 1.4 A relation $\tau\colon M \to N$ is *recognisable* if $\widehat{\tau} \in \operatorname{Rec} M \times N$. □

From this we deduce immediately that the *inverse* of a recognisable relation is recognisable and that, since $\operatorname{Rec} M$ is a Boolean algebra for all M, we have: *cf. Rem. 1.1, p. 526*

Property 1.5 *The complement of a recognisable relation and the intersection of two recognisable relations are recognisable relations.* ■

The other properties of recognisable relations are derived from the characterisation of recognisable subsets of a direct product of monoids: *cf. Cor. II.2.20, p. 254*

A relation $\tau\colon M \to N$ is recognisable if and only if there exists a finite set I of recognisable subsets R_i of M and S_i of N such that the graph of τ is written

$$\widehat{\tau} = \bigcup_{i \in I} R_i \times S_i \quad .$$

Property 1.6 *The image of an arbitrary subset of M under a recognisable relation (from M to N) is a recognisable subset of N.*

Proof. Let $\tau\colon M \to N$ be a recognisable relation with a decomposition as above and let K be an arbitrary subset of M. It follows that

$$K\tau = \bigcup_{j \in I,\, K \cap R_j \neq \emptyset} S_j \quad ,$$

a recognisable set, since $\operatorname{Rec} N$ is closed under union. ■

Property 1.7 *The composition of two recognisable relations is a recognisable relation.*

Proof. Let τ be as above and let $\sigma\colon N \to P$ be a recognisable relation whose graph is given by

$$\widehat{\sigma} = \bigcup_{j \in J} T_j \times U_j \quad ,$$

with J finite, the T_j in $\operatorname{Rec} N$ and the U_j in $\operatorname{Rec} P$. The graph of $\tau\sigma$ is given by

$$\widehat{\tau\sigma} = \bigcup_{(i,j) \in I \times J,\, S_i \cap T_j \neq \emptyset} R_i \times U_j \quad . \qquad ■$$

Proposition 1.8 *Let M be a finitely generated monoid, $\tau\colon M \to N$ a recognisable relation and $\sigma\colon N \to P$ a rational relation, then the composition $\tau\sigma\colon M \to P$ is a rational relation.*

Proof. The graph of $\tau \circ \sigma$ is, with the above notation, *cf. Prop. II.2.7, p. 243 and Th. 1.3*

$$\widehat{\tau\sigma} = \bigcup_{i \in I} R_i \times (S_i)\sigma \quad ,$$

which is a rational subset of $M \times P$, since the R_i and the $(S_i)\sigma$ are rational. ■

The case of the composition of a rational relation with a recognisable relation is treated by looking at the inverse. Recall also the instance of Proposition II.2.6 for relations, a property which is heavily used to 'normalise' rational relations: that is, give them particular forms:

cf. p. 239

Property 1.8 *The intersection of a recognisable relation and a rational relation is a rational relation.* ■

Property 1.9 *A recognisable function has a finite image.*

Proof. The graph of a recognisable function $\alpha\colon M \to N$ is written $\widehat{\alpha} = \bigcup_{i\in I} R_i \times S_i$ where I is finite and where each of the S_i must be a singleton $\{s_i\}$ so that α is a function. It then follows that $\mathsf{Im}\,\alpha = \bigcup_{i\in I} s_i$ is a finite set. ■

Property 1.10 *A rational function from A^* to B^* (with A finite) is recognisable if and only if it has a finite image.*

Proof. Let $\alpha\colon A^* \to B^*$ be a rational function. For each w in $\mathsf{Im}\,\alpha$, $w\alpha^{-1}$ is a rational subset, hence recognisable, of A^*. Each w in B^* is also a recognisable subset of B^*. Thus

$$\widehat{\alpha} = \bigcup_{w\in \mathsf{Im}\,\alpha} w\alpha^{-1} \times \{w\}$$

is a recognisable subset of $A^*\times B^*$ when $\mathsf{Im}\,\alpha$ is finite. The necessary condition follows from the previous property. ■

Exercises

• 1.21 **Iteration of a recognisable relation.** Show that the proper iteration of a recognisable relation is a recognisable relation. What can be said about the iteration?

1.22 Explain how, when a recognisable relation is written in the form

$$\widehat{\tau} = \bigcup_{i\in I} R_i \times S_i \ ,$$

we can always assume that the R_i (or the S_i) are pairwise disjoint.

1.5 Realisation by representation

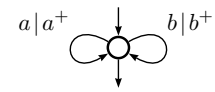
In the case of relations from a free monoid A^* to an arbitrary monoid M we can give a characterisation of rational relations based on Kleene's Theorem in A^*. This characterisation operates from the *dynamic point of view*: we relate each word f of A^* to its image $f\tau$ in the relation τ; τ is a map from A^* to $\mathfrak{P}(B^*)$, and is precisely what we called a *series on A^* with coefficients in $\mathfrak{P}(B^*)$* in the previous chapter.

I would like however to present it without all the mathematical apparatus used in Chapter III, and without all the precautions we shall be led to take in the next section in order to treat relations between series in full generality. It is in fact fundamental, in particular when one wants to compute on, or with, rational relations, and I think that it should be understood by all readers, including those who are not mathematically inclined.

1.5.1 Real-time transducers

A *real-time transducer*[11] over $A^* \times B^*$ or rather, *from* A^* *to* B^*, is an automaton $\mathcal{A} = \langle Q, A, B^*, E, I, T \rangle$ – note that we now have a *sextuple* – whose transitions are labelled with elements of $A \times \mathfrak{P}(B^*)$ and where I and T are functions from Q to $1_{A^*} \times \mathfrak{P}(B^*)$; the real-time transducer is *finite* if Q is finite, if the labels are in $A \times \operatorname{Rat} B^*$, and if the values of I and T are in $1_{A^*} \times \operatorname{Rat} B^*$.

Example 1.11 The transducer $\mathcal{S}_1$ in the adjacent figure is a real-time transducer which realises the substitution σ_1 defined by $a\sigma_1 = a^+$ and $b\sigma_1 = b^+$. □



Theorem 1.6 *A relation $\tau\colon A^* \to B^*$ is rational if and only if it is realised by a finite real-time transducer.*

Proof. The condition is sufficient; that is, let us show that a finite real-time transducer $\mathcal{A} = \langle Q, A, B^*, E, I, T \rangle$ realises a rational relation. We note first that we can always suppose I and T to be boolean vectors. We then use the technique of relabelling automata: let Z be a finite alphabet, in bijection with the transitions of $\mathcal{A}$, and $\mathcal{C} = \langle Q, Z, D, I, T \rangle$ be that automaton over Z^* such that $\mathcal{A}$ is the image of $\mathcal{C}$ in the relabelling η which preserves the label *cf. Th. II.1.1, p. 225 and Lem. II.1.8.*

$$D = \{\big(p, (p, (a, K_{a,p,q}), q), q\big) \mid (p, (a, K_{a,p,q}), q) \in E = Z\}$$

and we have $|\mathcal{A}| = |\mathcal{C}|\eta$. Then let $\varphi\colon Z^* \to A^*$ be the morphism and $\psi\colon Z^* \to B^*$ the substitution defined respectively by

$$(p, (a, K_{a,p,q}), q)\varphi = a \qquad \text{and} \qquad (p, (a, K_{a,p,q}), q)\psi = K_{a,p,q} \ .$$

We have $\eta = \varphi \times \psi$, or $|\mathcal{A}| = |\mathcal{C}|\varphi \times \psi$ and hence $\tau = \varphi^{-1} \circ \iota_{|\mathcal{C}|} \circ \psi$. By Theorem 1.4, τ is rational if the substitution ψ is: that is, if each of the $K_{a,p,q}$ is a rational language of B^*. *cf. Lem. 1.2, p. 531*

Conversely, let τ be a rational relation: τ is recognised by a finite transducer in normal form, which we can write in a matricial way as $\langle I, E, T \rangle$. The matrix E can be written as the sum $E = F + G$ where the coefficients of F are sums of elements of $(a, 1_{B^*})$ and those of G are sums of elements $(1_{A^*}, b)$. A now familiar calculation gives *cf. Th. 1.1, p. 529* *cf. Sec. 3.2, p. 562*

$$\begin{aligned} \widehat{\tau} = |\langle I, E, T \rangle| &= I \cdot E^* \cdot T = I \cdot (F + G)^* \cdot T \\ &= I \cdot (G^* \cdot F)^* \cdot (G^* \cdot T) = (I \cdot G^*) \cdot (F \cdot G^*)^* \cdot T \ . \end{aligned}$$

The coefficients of G^* are all in the form $(1_{A^*}, K)$ where K is a rational of B^*. The coefficients of $G^* \cdot F$ (and of $F \cdot G^*$) are hence finite sums of elements of the form (a, K) with a in A and K in $\operatorname{Rat} B^*$: $\langle I, G^* \cdot F, G^* \cdot T \rangle$ and $\langle I \cdot G^*, F \cdot G^*, T \rangle$ are two finite real-time transducers that recognise τ. ■ *cf. Prop. III.2.10, p. 401*

[11] This name refers implicitly to the Rabin–Scott model that we describe in the next subsection; it says that at each transition of the transducer one letter is read from the input tape; there is a correspondence between the length of the word read and that of the computation that deals with it.

Example 1.11 (continued) A transducer in normal form which recognises the substitution σ_1 can be written

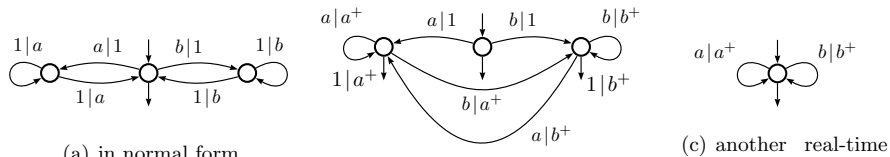
$$I = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \quad , \quad E = \begin{pmatrix} 0 & (a,1) & (b,1) \\ (1,a) & (1,a) & 0 \\ (1,b) & 0 & (1,b) \end{pmatrix} \quad , \quad T = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} .$$

The above calculation gives

$$F = \begin{pmatrix} 0 & (a,1) & (b,1) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad , \quad G = \begin{pmatrix} 0 & 0 & 0 \\ (1,a) & (1,a) & 0 \\ (1,b) & 0 & (1,b) \end{pmatrix} \quad ,$$

$$G^* = \begin{pmatrix} (1,1) & 0 & 0 \\ (1,a^+) & (1,a^*) & 0 \\ (1,b^+) & 0 & (1,b^*) \end{pmatrix} \quad ,$$

hence the two real-time transducers of Figures 1.7 (b) and (c), which correspond to computations $\langle I, G^* \cdot F, G^* \cdot T \rangle$ and $\langle I \cdot G^*, F \cdot G^*, T \rangle$ respectively. □



(a) in normal form

(b) a real-time transducer

(c) another real-time transducer

Figure 1.7: Transducers for σ_1

Remark 1.4 The proof of Theorem 1.6 (necessary condition) applies verbatim to transducers which are not in normal form: it suffices that the labels be (finite) sums of elements of the form (a, K) or $(1_{A^*}, K)$, with a in A and K in $\mathrm{Rat}\, B^*$. □

cf. Exer. 1.23

1.5.2 From real-time transducers to representations

A real-time transducer $\langle I, E, T \rangle$ of dimension Q naturally defines a morphism $\mu\colon A^* \to \mathrm{Rat}\, B^{*Q \times Q}$: for each a in A, $a\mu$ is the matrix of the second components of elements whose first component is a. As all their coefficients are of the form $(1_{A^*}, K)$, with K in $\mathrm{Rat}\, B^*$, I and T *can be identified* with vectors whose coefficients are in $\mathrm{Rat}\, B^*$. Modulo this identification, we have just related $\langle I, E, T \rangle$ to what we can call a representation (I, μ, T) from A^* to $\mathrm{Rat}\, B^*$. For instance, the real-time transducer of Figure 1.7(b) corresponds to the representation

$$I = \begin{pmatrix} 1_{B^*} & 0 & 0 \end{pmatrix} \quad , \quad a\mu = \begin{pmatrix} 0 & 1_{B^*} & 0 \\ 0 & a^+ & 0 \\ 0 & a^+ & 0 \end{pmatrix} \quad , \quad b\mu = \begin{pmatrix} 0 & 0 & 1_{B^*} \\ 0 & 0 & b^+ \\ 0 & 0 & b^+ \end{pmatrix} \quad , \quad T = \begin{pmatrix} 1_{B^*} \\ a^+ \\ b^+ \end{pmatrix} \quad .$$

cf. Lem. III.3.8, p. 433

Conversely, and still modulo the above-mentioned identification, we can see every $\mathrm{Rat}\, B^*$-representation (λ, μ, ν) as a finite real-time transducer $\langle \lambda, E, \nu \rangle$, by setting $E = \sum_{a \in A} (a, a\mu)$.

It is then elementary to verify, by writing τ for the relation recognised by $\langle I, E, T \rangle$ (and μ for the representation associated with E), that, for all f in A^*, we have

$$f\tau = \lambda \cdot f\mu \cdot \nu \,;$$

that is, (λ, μ, ν) realises τ. It is important to realise that 'finite real-time transducer' (from A^* to B^*) and '$\mathrm{Rat}\, B^*$-representation of A^*' denote *one and the same object*, written in two different ways. Theorem 1.6 can[12] also be rewritten:

Theorem 1.7 *A relation $\tau\colon A^* \to B^*$ is rational if and only if it is realised by a representation from A^* to* $\mathrm{Rat}\, B^*$. ■

1.5.3 Theorem of Evaluation and Composition of Representations

Continuing to anticipate what we shall develop in Section 3, the realisation of rational relations by representations enables us to give another expression and proof of the two pivotal theorems.

Lemma 1.9 *Let $\mu\colon A^* \to \mathrm{Rat}\, B^{*Q\times Q}$ be a $\mathrm{Rat}\, B^*$-representation of A^* and K a rational of A^*. Then the coefficients of $K\mu$ are rationals of B^*.*

Proof. Let E be a rational expression over A^* which denotes K. We can see E as a rational expression over $\mathbf{A}$, a set of variables in bijection with A, and K is the result of E under the interpretation τ which to each a in $\mathbf{A}$ relates $\mathsf{a}\tau = a$. *cf. Exam. 1.3, p. 527*

Let μ' be the interpretation taking values in $\mathrm{Rat}\, B^{*Q\times Q}$ which relates each a in $\mathbf{A}$ to $\mathsf{a}\mu' = a\mu$. By induction on the depth of E, we prove both that $L\mu = [\mathsf{E}]\,\mu'$ and the result of the lemma (we use Proposition III.2.10 for the operator $*$). ■ *cf. p. 401*

The Evaluation Theorem follows directly: if (λ, μ, ν) is a $\mathrm{Rat}\, B^*$-representation which realises $\tau\colon A^* \to B^*$ and L is a rational language in A^*, then $L\tau = \lambda \cdot L\mu \cdot \nu$.

We then define the composition product $\pi = \mu \circ \kappa$ of two morphisms $\mu\colon A^* \to \mathrm{Rat}\, B^{*Q\times Q}$ and $\kappa\colon B^* \to \mathrm{Rat}\, C^{*R\times R}$. It is a morphism

$$\pi\colon A^* \to \mathrm{Rat}\, C^{*(Q\times R)\times(Q\times R)} \quad ,$$

obtained by replacing each coefficient in $a\mu$ by its image under κ:

$$\forall a \in A\,,\ \forall q, s \in Q\,,\ \forall r, t \in R \qquad a\pi_{(q,r),(s,t)} = (a\mu_{q,s})\kappa_{r,t} \quad . \tag{1.1}$$

Lemma 1.9 ensures that every entry of $a\pi$, for all a in A, is in $\mathrm{Rat}\, C^*$. We then verify that the same Equation (1.1) continues to hold when the letter a is replaced by any word f in A^*: that is, that π is indeed a $\mathrm{Rat}\, C^*$-representation. Then we define the composition product $(\zeta, \pi, \omega) = (\lambda, \mu, \nu) \circ (\eta, \kappa, \chi)$ by the formulas *cf. Lem. 3.10, p. 566*

$$\pi = \mu \circ \kappa \quad ,$$

$$\forall q \in Q \qquad \zeta_{(q,R)} = \eta \cdot (\lambda_q)\kappa \qquad \text{and} \qquad \omega_{(q,R)} = (\nu_q)\kappa \cdot \chi \quad . \tag{1.2}$$

We then prove that the composition product of representations realises the composition of the relations: *cf. Th. 3.6, p. 567*

[12] It is more in this form that it is known as the Kleene–Schützenberger Theorem (for rational relations).

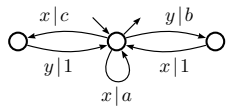
Theorem 1.8 *Let $\theta\colon A^* \to B^*$ and $\sigma\colon B^* \to C^*$ be two rational relations that are realised respectively by the representations (λ, μ, ν) and (η, κ, χ). Then the composition $\theta\sigma\colon A^* \to C^*$ is realised by the composition product $(\zeta, \pi, \omega) = (\lambda, \mu, \nu) \circ (\eta, \kappa, \chi)$, which implies that it is a rational relation.* ■

cf. Exer. 1.8 **Example 1.12** Let $\varphi_1\colon \{a,b,c\}^* \to \{x,y\}^*$ be the morphism defined by

$$a\varphi_1 = x \ , \quad b\varphi_1 = y\,x \ , \quad c\varphi_1 = x\,y \ ,$$

and let us compute a transducer for the relation $\varphi_1\varphi_1^{-1}$. The representation of φ_1 is clearly $(1, \varphi_1, 1)$. From the graph of φ_1, $\widehat{\varphi_1} = [(a,x) + (b,y\,x) + (c,x\,y)]^*$, we deduce the graph of φ_1^{-1}:

$$\begin{aligned}\widehat{\varphi_1^{-1}} &= [(x,a) + (y\,x,b) + (x\,y,c)]^* \\ &= [(x,a) + (y,b)(x,1) + (x,c)(y,1)]^* \ ,\end{aligned}$$



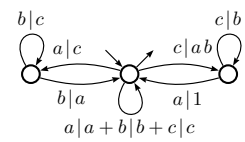
which leads to the real-time transducer above and thence to the representation

$$\eta_1 = \begin{pmatrix}1 & 0 & 0\end{pmatrix} \ , \quad x\kappa_1 = \begin{pmatrix} a & c & 0 \\ 0 & 0 & 0 \\ 1_{A^*} & 0 & 0 \end{pmatrix} \ , \quad y\kappa_1 = \begin{pmatrix} 0 & 0 & b \\ 1_{A^*} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \ , \quad \chi_1 = \begin{pmatrix}1\\0\\0\end{pmatrix} \ .$$

A realisation by representation of the composition $\varphi_1\varphi_1^{-1}$ is given by the composition product $(\zeta_1, \pi_1, \omega_1) = (1, \varphi_1, 1) \circ (\eta_1, \kappa_1, \chi_1)$:

$$\eta_1 = \begin{pmatrix}1 & 0 & 0\end{pmatrix} \ , \quad a\pi_1 = x\kappa_1 = \begin{pmatrix} a & c & 0 \\ 0 & 0 & 0 \\ 1_{A^*} & 0 & 0 \end{pmatrix} \ , \quad b\pi_1 = (y\,x)\kappa_1 = \begin{pmatrix} b & 0 & 0 \\ a & c & 0 \\ 0 & 0 & 0 \end{pmatrix} \ ,$$

$$c\pi_1 = (x\,y)\kappa_1 = \begin{pmatrix} c & 0 & a\,b \\ 0 & 0 & 0 \\ 0 & 0 & b \end{pmatrix} \ , \qquad \chi_1 = \begin{pmatrix}1\\0\\0\end{pmatrix} \ .$$



which corresponds to the real-time transducer opposite. □

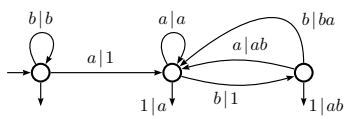
Exercises

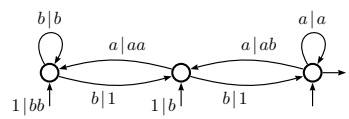
1.23 Give a *real-time* transducer which recognises:

(a) the complement of the identity; (b) the lexicographic order relation; (c) the radix order relation.

• 1.24 Give a rational expression for the graph of $\varphi_1\varphi_1^{-1}$ (*cf.* Example 1.12). What is the interest of this exercise?

• 1.25 Give a transducer which realises the composition of the relations realised by the transducers below (the transducer on the left by the transducer on the right).





• 1.26 Do the same for the transducers below.

• 1.27 If τ is a relation from A^* to M, we define τ_{f} and τ_∞ to be the *finite* and *infinite* 'parts' of τ respectively as follows:

if $\|f\tau\|$ is finite, then $f\tau_{\mathsf{f}} = f\tau$ and $f\tau_\infty = \emptyset$, otherwise $f\tau_{\mathsf{f}} = \emptyset$ and $f\tau_\infty = f\tau$.

Show that if τ is rational, then τ_{f} and τ_∞ are rational and effectively computable from τ.

1.6 The Rabin–Scott model

Modelling a finite automaton as a constrained Turing machine leads naturally to a generalisation of the model to several tapes. The device still comprises a finite control with a finite number of states and no longer one but *several* input tapes. The finite control is linked to each of these tapes by a read head (see Figure 1.8). *cf. Sec. I.7, p. 171*

At each step of the computation, the control, according to its current state p, *chooses a tape* which it will read, and, according to the symbol a read from that tape, changes to a state q and *moves the read head of that tape* one square to the right.[13] Because the read head always moves *in the same direction*, this type of machine is called a *1-way k-tape automaton*.

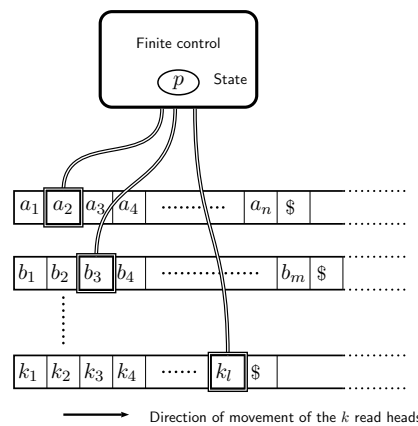


Fig. 1.8: A Rabin–Scott k-tape automaton

If the tape to be read is uniquely determined by the state p and if, having chosen this tape, the state q is uniquely determined by p and the symbol read from the tape, the automaton is *deterministic*; otherwise it is non-deterministic.

At the start of a computation, a word is written on each of the k input tapes, the finite control is in a distinguished *initial* state, and each read head is placed on the first square of each of the k tapes. After a series of steps like that described above, a computation *terminates* if all the read heads have reached the square on their respective tapes in which is written the *end of tape symbol*. The computation is *successful* if the state in which the finite control finds itself at the end of the

[13] We can imagine other definitions of the mode of operation; in particular the choice of the read head being moved and of the next state could depend not only on the current state but also on the symbols read from all the tapes. In fact, these definitions turn out to be equivalent.

computation is a *final* state. A k-tuple of words is accepted by the automaton if it can be read by a successful computation.

The equivalence of this machine model with transducers, labelled graphs over $A^* \times B^*$ for 2-tape automata, or labelled graphs over $A_1^* \times A_2^* \times \cdots \times A_k^*$ for k-tape automata, is not as trivial as that between labelled graphs over A^* and 1-tape automata, but does not pose particular problems.

cf. p. 173 On the other hand, Theorem I.7.1 (two-way automata accept rational languages) cannot be extended to k-tape automata and, if even only one read head can move in both directions, the set of k-tuples accepted is no longer necessarily a rational set of $A_1^* \times A_2^* \times \cdots \times A_k^*$.

We will return to the k-tape model of automata to define the classes of rational relations studied in Sections 5 and 6.

Exercises

1.28 Verify the equivalence of the Rabin–Scott k-tape model with transducers. Verify that the model would still be equivalent if the choice of the tape read and the new state depended not only on the previous state but also the symbols read on all tapes (*cf.* Note 13).

• 1.29 Give a simple example of a 2-tape automaton which can move its read head in both directions on one of two tapes and which accepts a non-rational subset of $A^* \times B^*$.

2 $\mathbb{K}$-RELATIONS

After *series* on a monoid M with coefficients in a semiring $\mathbb{K}$, a generalisation of subsets of M, and after *relations* from M to N, which extend *additively* to maps from $\mathfrak{P}(M)$ to $\mathfrak{P}(N)$, we now consider maps from $\mathbb{K}\langle\langle M\rangle\rangle$ to $\mathbb{K}\langle\langle N\rangle\rangle$ which generalise relations from M to N and which we shall call $\mathbb{K}$-*relations*. We impose the double constraint that $\mathbb{K}$-relations should be to series what relations are to subsets, *hence additive*, and to relations what series are to subsets: that is, *relations with coefficients*. Before launching into more precise and necessarily formal definitions, let us describe the situation, and the problems, with an example.

Example 2.1 Let $M = \{a\}^*$ and $N = \{x\}^*$ be two free monoids with one generator and let us consider the relation θ_1 from a^* to x^* defined by

$$a^n\theta_1 = \{x^m \mid 0 \leqslant m \leqslant n\}$$

whose graph $\widehat{\theta_1} = \{(a^n, x^m) \mid 0 \leqslant m \leqslant n\} = \{(a, 1_{x^*}), (a, x)\}^*$ is shown in Figure 2.1. We then compute naturally that

$$a\theta_1 = \{1_{x^*}, x\} \ , \quad \{a, a^2\}\theta_1 = a\theta_1 \cup a^2\theta_1 = \{1_{x^*}, x, x^2\} \quad \text{and} \quad \{a^*\}\theta_1 = x^* \ .$$

We see clearly that θ_1 is a map from $\mathfrak{P}(a^*)$ to $\mathfrak{P}(x^*)$, totally defined by the value of θ_1 on each element of M, or, which comes to the same thing, by the graph of θ_1, an element of $\mathfrak{P}(a^* \times x^*)$.

Now let us look at a situation where the basic objects are no longer subsets of $\mathfrak{P}(a^*)$, or of $\mathfrak{P}(x^*)$, but series on a^*, or on x^*, with coefficients in $\mathbb{N}$: i.e. $\mathbb{N}\langle\langle a^*\rangle\rangle$ replaces $\mathfrak{P}(a^*)$, and $\mathbb{N}\langle\langle x^*\rangle\rangle$ replaces $\mathfrak{P}(x^*)$. The map that will take the place of θ_1, which we shall write σ_1, will associate each a^n with an element of $\mathbb{N}\langle\langle x^*\rangle\rangle$:

$$a\sigma_1 = 1_{x^*} + x \ , \quad a^2\sigma_1 = 1_{x^*} + x + x^2 \qquad \text{and, more generally,}$$
$$a^n\sigma_1 = 1_{x^*} + x + x^2 + \cdots + x^n = \sum_{0\leqslant m\leqslant n} x^m \ .$$

We compute, by additivity,

$$(a + a^2)\sigma_1 = a\sigma_1 + a^2\sigma_1 = 2\,1_{x^*} + 2\,x + x^2 \ ,$$

where we note a first difference from θ_1. The same computation gives, for all n,

$$(a^{<n})\sigma_1 = n\,1_{x^*} + (n-1)\,x + (n-2)\,x^2 + \cdots + x^{n-1} \ . \tag{2.1}$$

We see clearly that σ_1 is a map from $\mathbb{N}\langle a^*\rangle$ to $\mathbb{N}\langle x^*\rangle$, totally defined by the value of σ_1 for each element of M or, which comes to the same thing, by an element of $\mathbb{N}\langle\langle a^*\times x^*\rangle\rangle$, which we shall call the *graph* of σ_1. We deduce from (2.1) a second difference from θ_1 and our first difficulty, that the image of the whole characteristic series of a^*, $(\underline{a^*})\sigma_1$, *cannot be defined.*

$\vdots$					$\iddots$
x^3				$\times$	$\cdots$
x^2			$\times$	$\times$	$\cdots$
x		$\times$	$\times$	$\times$	$\cdots$
1_{x^*}	$\times$	$\times$	$\times$	$\times$	$\cdots$
	1_{a^*}	a	a^2	a^3	$\cdots$

(a) $\widehat{\theta_1}$

$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\iddots$
x^3	0	0	0	1	$\cdots$
x^2	0	0	1	1	$\cdots$
x	0	1	1	1	$\cdots$
1_{x^*}	1	1	1	1	$\cdots$
	1_{a^*}	a	a^2	a^3	$\cdots$

(b) $\widehat{\sigma_1}$

$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\iddots$
x^3	0	0	0	1	$\cdots$
x^2	0	0	1	3	$\cdots$
x	0	1	2	3	$\cdots$
1_{x^*}	1	1	1	1	$\cdots$
	1_{a^*}	a	a^2	a^3	$\cdots$

(c) $\widehat{\tau_1}$

Figure 2.1: The graphs of the relations θ_1, σ_1 and τ_1

The graph of σ_1 is the characteristic series of the graph of θ_1; if we define the $\mathbb{N}$-relations by series on $a^*\times x^*$, there is no reason to limit ourselves to characteristic series. Let τ_1 be the $\mathbb{N}$-relation whose graph $\widehat{\tau_1}$ is defined by

$$\widehat{\tau_1} = \{(a, 1_{x^*}) + (a, x)\}^*$$

which is shown in Figure 2.1(c). We then compute, for example,

$$a\tau_1 = 1_{x^*} + x \ , \quad a^2\tau_1 = 1_{x^*} + 2\,x + x^2 \quad \text{and} \quad (a + a^2)\tau_1 = 2\,1_{x^*} + 3\,x + x^2 \ .$$

This natural, even necessary, extension brings out a difficulty which has been concealed by the commutativity of $\mathbb{N}$. Suppose for a moment that we are considering a $\mathbb{K}$-relation μ from a^* to x^* with $\mathbb{K}$ non-commutative: $a\mu$ is a series on $\mathbb{K}\langle\langle x^*\rangle\rangle$; we obviously want, for all k in $\mathbb{K}$,

$$(k\,a)\mu = k\,(a\mu) \qquad \text{and} \qquad (a\,k)\mu = (a\mu)\,k \ . \tag{2.2}$$

However, $k\,a = a\,k$ in $\mathbb{K}\langle\langle x^*\rangle\rangle$ and there is no reason why it should be true that $k\,(a\mu) = (a\mu)\,k$, unless *every coefficient* of $a\mu$ commutes with *all the elements* of $\mathbb{K}$. With this example under our belts, we can now proceed to general definitions. □

2.1 Definitions

For the rest of the *chapter* M and N denote two monoids and $\mathbb{K}$ is a semiring, chosen in such a way that the semialgebras $\mathbb{K}\langle\langle M\rangle\rangle$, $\mathbb{K}\langle\langle N\rangle\rangle$, and $\mathbb{K}\langle\langle M\times N\rangle\rangle$ are well defined, as are the rational operations in these semialgebras: $\mathbb{K}$ is a (topological) strong semiring and M and N are *graded*, or else $\mathbb{K}$ is *continuous* and M and N are arbitrary. We shall thus treat two very different situations together. In the first we shall have to ensure at each step that the objects we construct are well defined, while in the second, on the contrary, this is ensured by the nature of the definitions.

cf. Def. III.1.3, p. 383 and Def. III.5.1, p. 471

The *centre* of $\mathbb{K}$ (that is, the set of elements of $\mathbb{K}$ which commute with every element of $\mathbb{K}$) is a sub-semiring of $\mathbb{K}$, written $\mathbb{K}_{\mathsf{c}}$.

2.1.1 The canonical isomorphism

When we consider polynomials or series in several (commutative) variables, it is well known that we can order them with respect to one or another variable. It is a purely formal exercise to verify that these manipulations generalise to polynomials, or to series, over a product of monoids.

Proposition 2.1 *The three $\mathbb{K}$-semialgebras*

$$\mathbb{K}\langle\langle M\times N\rangle\rangle \quad , \qquad [\mathbb{K}\langle\langle M\rangle\rangle]\,\langle\langle N\rangle\rangle \qquad \textit{and} \qquad [\mathbb{K}\langle\langle N\rangle\rangle]\,\langle\langle M\rangle\rangle$$

are isomorphic. Under these isomorphisms the three $\mathbb{K}$-sub-semialgebras

$$\mathbb{K}\langle M\times N\rangle \quad , \qquad [\mathbb{K}\langle M\rangle]\,\langle N\rangle \qquad \textit{and} \qquad [\mathbb{K}\langle N\rangle]\,\langle M\rangle$$

correspond.

Proof. The semialgebras $\mathbb{K}\langle\langle M\rangle\rangle$ and $\mathbb{K}\langle\langle N\rangle\rangle$ are canonically isomorphic to two $\mathbb{K}$-sub-semialgebras of $\mathbb{K}\langle\langle M\times N\rangle\rangle$: we identify m with $(m, 1_N)$ and n with $(1_M, n)$. This identification enables us to make the two canonical isomorphisms explicit. If α is in $\mathbb{K}\langle\langle M\times N\rangle\rangle$,

$$\alpha = \sum_{(m,n)\in M\times N} <\alpha,(m,n)>(m,n) = \sum_{(m,n)\in M\times N} <\alpha,(m,n)>(m,1_N)(1_M,n)$$

can be written, by pulling out the factors $(m, 1_N)$, or $(1_M, n)$ (which is allowed since the families $\{(m,1_N)\}_{m\in M}$ and $\{(1_M,n)\}_{n\in N}$ are summable),

$$\begin{aligned}\alpha &= \sum_{m\in M}\left(\sum_{n\in N} <\alpha,(m,n)>(1_M,n)\right)(m,1_N)\\ &= \sum_{n\in N}\left(\sum_{m\in M} <\alpha,(m,n)>(m,1_N)\right)(1_M,n)\ .\end{aligned}$$

Thus α corresponds to α' in $[\mathbb{K}\langle\langle N\rangle\rangle]\,\langle\langle M\rangle\rangle$:

$$\alpha' = \sum_{m\in M}\left(\sum_{n\in N} <\alpha,(m,n)>n\right) m \ ,$$

and α'' in $[\mathbb{K}\langle\langle M\rangle\rangle]\,\langle\langle N\rangle\rangle$:

$$\alpha'' = \sum_{n\in N}\left(\sum_{m\in M} <\alpha,(m,n)>m\right) n \ .$$

This correspondence is formal and obviously bijective; just as obviously, it maps polynomials over $M\times N$ to polynomials over M whose coefficients are polynomials over N. We see that it is indeed a morphism on $\mathbb{K}$-semialgebras, whether $\mathbb{K}$ is commutative or not. ■ *cf. Exer. 2.1*

Remark 2.1 Modulo this identification *and if* $\mathbb{K}$ *is commutative*, then every element of $\mathbb{K}\langle\langle M\rangle\rangle$ commutes with every element of $\mathbb{K}\langle\langle N\rangle\rangle$. □

The coefficients of a series can be seen as the result of a *scalar product* of the series with the elements of the underlying monoid, identified with their characteristic series. The scalar product is thus well suited to express computations over the coefficients of the series. *Even if* $\mathbb{K}$ *is not commutative*, but if s *or* t belong to $\mathbb{K}_{\mathsf{c}}\langle\langle M\rangle\rangle$, we have

$$<s,t> = <t,s>$$

and we shall use this notational trick freely. In this case, the scalar product is even compatible with left and right multiplication by arbitrary elements of $\mathbb{K}$:

$$k<s,t> = <k\,s,t> \ , \qquad <s,t>k = <t,s>k = <t,s\,k> = <s\,k,t> \ .$$

2.1.2 $\mathbb{K}$-relations

The definition of weighted relations, or $\mathbb{K}$-*relations*, repeats point for point the definition of a relation from M to N from its graph. Example 2.1 showed that we must expect a $\mathbb{K}$-relation not to be totally defined; what is important is that the elementary properties hold on the domain of definition.

Definition 2.1 Let $\mathbb{K}$ be a semiring and $\mathbb{K}_{\mathsf{c}}$ its centre; let M and N be two monoids. A $\mathbb{K}$-*relation from* M *to* N, written $\theta\colon \mathbb{K}\langle\langle M\rangle\rangle \to \mathbb{K}\langle\langle N\rangle\rangle$, or more often $\theta\colon M\to N$, is any series θ of $[\mathbb{K}_{\mathsf{c}}\langle\langle N\rangle\rangle]\,\langle\langle M\rangle\rangle$.

(i) The *graph* of θ is the series $\widehat{\theta}$, an element of $\mathbb{K}_{\mathsf{c}}\langle\langle M\times N\rangle\rangle$ which corresponds to θ under the canonic isomorphism.

(ii) The *inverse* of θ, namely θ^{-1}, is the $\mathbb{K}$-relation from N to M, and hence a series of $[\mathbb{K}_{\mathsf{c}}\langle\langle M\rangle\rangle]\,\langle\langle N\rangle\rangle$, which has the same graph $\widehat{\theta}$ as θ .

(iii) For all m in M, the *image* of m under θ is the series $<\theta,m>$, written more simply $m\theta$, a series of $\mathbb{K}\langle\langle N\rangle\rangle$ whose coefficient at each n of N is $<\widehat{\theta},(m,n)>$. Symmetrically, for each n in N, the image of n under θ^{-1} is the series $<\theta^{-1},n>$, written more

simply $n\theta^{-1}$, a series of $\mathbb{K}\langle\langle M\rangle\rangle$ whose coefficient at each m of M is also $<\widehat{\theta},(m,n)>$ and will be written $<m,n\theta^{-1}>$ rather than $<n\theta^{-1},m>$. The following equations

$$\forall (m,n)\in M\times N \qquad <m\theta,n> = <\widehat{\theta},(m,n)> = <m,n\theta^{-1}> \ , \tag{2.3}$$

$$m\theta = \sum_{n\in N} <\widehat{\theta},(m,n)>n \qquad \text{and} \qquad n\theta^{-1} = \sum_{m\in M} <m,n\theta^{-1}>m \ , \tag{2.4}$$

summarise the various possible points of view on a $\mathbb{K}$-relation. □

Remark 2.2 Instead of assuming that the semiring of coefficients is *commutative*,[14] we have 'only' imposed the condition – to deal with the difficulty raised in Example 2.1 and to ensure Equation (2.2) – that the coefficients of the relation, $\widehat{\theta}$, belong to the *centre* of this semiring. This could seem a rather weak generalisation; in fact it allows us first and foremost to consider, as $\mathbb{K}$-relations from M to N, the *characteristic relations* of relations from M to N, even if $\mathbb{K}$ is not commutative. □

Definition 2.2 Let θ be a $\mathbb{K}$-relation from M to N and s a series of $\mathbb{K}\langle\langle M\rangle\rangle$. The image of s under θ, namely $s\theta$, is defined if and only if the family $\{<s,m>m\theta\}_{m\in M}$ is a *summable family* of series of $\mathbb{K}\langle\langle N\rangle\rangle$. In that case,

$$s\theta = \sum_{m\in M} <s,m>m\theta \ . \qquad \square$$

cf. Propty III.1.2, p. 388

By definition of the topology of $\mathbb{K}\langle\langle N\rangle\rangle$, $\{<s,m>m\theta\}_{m\in M}$ is summable if and only if for all n in N, the family $\{<s,m><m\theta,n>\}_{m\in M}$ is summable in $\mathbb{K}$ and we then have

$$\forall n\in N \qquad <s\theta,n> = \sum_{m\in M} <s,m><m\theta,n> \ . \tag{2.5}$$

This is the same as saying that if $s\theta$ is defined we can, as in the canonical isomorphism, 'formally invert' the summations over M and N:

$$s\theta = \sum_{m\in M} <s,m>\left(\sum_{n\in N} <m\theta,n>n\right) = \sum_{n\in N}\left(\sum_{m\in M} <s,m><m\theta,n>\right) n \ .$$

Example 2.2 For every series t in $\mathbb{K}_{\mathsf{c}}\langle\langle M\rangle\rangle$, the Hadamard product with t (or $\mathbb{K}$-*intersection with* t) is a $\mathbb{K}$-relation from M to itself, written ι_t:

$$\forall s\in\mathbb{K}\langle\langle M\rangle\rangle \qquad s\iota_t = s\odot t \ .$$

It is immediate to verify that the graph of ι_t is the series

$$\widehat{\iota_t} = \sum_{m\in M} <t,m>(m,m) \ ,$$

and that $s\iota_t$ is defined for all s in $\mathbb{K}\langle\langle M\rangle\rangle$. If t is a characteristic series,[15] ι_t is a $\mathbb{K}$-relation, for arbitrary $\mathbb{K}$. □

[14] As usual, if we can say that: weighted relations have not been much studied; see the bibliographic notes.

[15] A typical case where it is useful not to have to assume that $\mathbb{K}$ is commutative.

Example 2.3 The relation 'sum of coefficients' $\sigma\colon \mathbb{K}\langle\langle M\rangle\rangle \to \mathbb{K}$ defined by $m\sigma = 1_{\mathbb{K}}$ for all m in M, from which, if the sum is defined,

$$\forall s \in \mathbb{K}\langle\langle M\rangle\rangle \qquad s\sigma = \sum_{m\in M} <s,m> = <s,\underline{M}> \ ,$$

is a $\mathbb{K}$-relation from M to the trivial monoid with one element or alternatively, from M to *any* monoid N:

$$\widehat{\sigma} = \underline{M\times 1_N} \ .$$

In general, and if $\mathbb{K}$ is not complete, σ is not defined outside $\mathbb{K}\langle M\rangle$. □

Property 2.1 *Let θ be a $\mathbb{K}$-relation from M to N and s in $\mathbb{K}\langle\langle M\rangle\rangle$. The image $s\theta$ is defined if and only if $<s,n\theta^{-1}>$ is defined for all n in N, and we have*

$$<s\theta,n> = <s,n\theta^{-1}> \ . \tag{2.6}$$

Proof. The series $s\theta$ is defined if, for all n in N, the family $\{<s,m><m\theta,n>\}_{m\in M}$ *cf. Equat. (2.5)* is summable, or if the family $\{<s,m><m,n\theta^{-1}>\}_{m\in M}$ is summable, which means exactly that the scalar product $<s,n\theta^{-1}>$ is defined. ■

Equation (2.6) is the extension 'by linearity' of the equality (2.3) which defines $\mathbb{K}$-relations, linearity extended to infinite sums thanks to the topology on $\mathbb{K}\langle\langle M\times N\rangle\rangle$. It will be used frequently in what follows. Following the same line of thought, we can easily prove:

Properties 2.2 *Let θ be a $\mathbb{K}$-relation from M to N. For all k in $\mathbb{K}$ and all s and t in $\mathbb{K}\langle\langle M\rangle\rangle$, the following holds.*

(i) *If $s\theta$ is defined, then $(k\,s)\theta$ and $(s\,k)\theta$ are defined and we have:*

$$(k\,s)\theta = k\,(s\theta) \qquad \textit{and} \qquad (s\,k)\theta = (s\theta)\,k \ .$$

(ii) *If $s\theta$ and $t\theta$ are defined, then $(s+t)\theta$ is defined and we have:*

$$(s+t)\theta = s\theta + t\theta \ .$$

■

2.1.3 Support of relations – characteristic relations

If $\mathbb{K} = \mathbb{B}$, the Boolean semiring, we will simply write 'relation', rather than '$\mathbb{B}$-relation', and the reader can verify that we have said nothing and will add nothing that will contradict the material developed in the previous section.

$\mathbb{K}$-relations from M to N can, via their graph, be dealt with by the concepts defined for series. Thus we can speak of the *support* of a $\mathbb{K}$-relation from M to N: it is a subset of $M\times N$, i.e. a relation, in the usual sense of the term, from M to N.

We shall say of a $\mathbb{K}$-relation θ that it is *of finite image*[16] if that is the case *for its support*, i.e. if $m\theta$ is a *polynomial* for all m; dually, we shall say that θ is *of finite co-image*[17] if $n\theta^{-1}$ is a *polynomial* for all n. Then, from Property 2.1:

[16] Some authors call them 'faithful $\mathbb{K}$-relations'; see [13] for example.

[17] G. Jacob [126, 127], and others following him, call them 'regulated $\mathbb{K}$-relations'.

Property 2.3 *The image of a series under a $\mathbb{K}$-relation of finite co-image is always defined.* ■

Definition 2.3 A $\mathbb{K}$-relation is *characteristic*, or *unambiguous*, if its graph is a characteristic series of $\mathbb{K}\langle\langle M \times N\rangle\rangle$. □

We can associate each relation τ from M to N with its *characteristic* $\mathbb{K}$*-relation* $\underline{\tau}$, also from M to N, whose graph is the characteristic series of the graph of τ. This definition is consistent with that of the inverse of a ($\mathbb{K}$-)relation:

$$(\underline{\tau})^{-1} = \underline{(\tau^{-1})} \tag{2.7}$$

and we simply write $\underline{\tau}^{-1}$.

Example 2.2 (continued) Let t be in $\mathbb{K}_{\mathsf{c}}\langle\langle M\rangle\rangle$; the $\mathbb{K}$-relation ι_t is characteristic if the series t is. Conversely, if P is a subset of M, $\iota_{\underline{P}}$ is the characteristic $\mathbb{K}$-relation of the relation 'intersection with P'. □

Along the same lines, we say:

Definition 2.4 A $\mathbb{K}$-relation θ from M to N is a $\mathbb{K}$*-function* if its support is a function from M to N, i.e. if $m\theta$ is *a monomial* for all m in M. A $\mathbb{K}$-relation is *injective* if its support is an injective relation; that is, if the supports of the images of two distinct elements of M are disjoint. □

cf. Exer. 0.1.8, p. 14

We then easily show:

Proposition 2.2 *A $\mathbb{K}$-relation is injective if and only if its inverse is a $\mathbb{K}$-function.* ■

2.1.4 Continuity

A $\mathbb{K}$-relation *is not necessarily continuous*; that is, a sequence s_n of series of $\mathbb{K}\langle\langle M\rangle\rangle$ can converge to a series s, and $s\theta$ and $(s_n)\theta$ can be defined for all n, yet the $(s_n)\theta$ may converge to a limit other than $s\theta$.

Take for example $\mathbb{K} = \mathbb{N}$, $M = a^*$, and the relation σ, 'sum of coefficients'. The sequence $s_n = a^n$ converges to the zero series z, but $(s_n)\sigma = 1$ hence has the limit 1 when n tends to infinity, and $z\sigma = 0$.

Proposition 2.3 *Let M and N be two graded monoids. A $\mathbb{K}$-relation $\theta\colon M \to N$ with finite co-image is continuous.*

Proof. Let $\{s_n\}_{n\in\mathbb{N}}$ be a sequence of series of $\mathbb{K}\langle\langle M\rangle\rangle$ which converges to a series s; that is, for every m in M the sequence $\{\langle s_n, m\rangle\}_{n\in\mathbb{N}}$ converges to $\langle s, m\rangle$.

cf. Propty III.1.2, p. 388

For all p in N, $\langle s_n\theta, p\rangle = \langle s_n, p\theta^{-1}\rangle$ and the sequence $\{\langle s_n, p\theta^{-1}\rangle\}_{n\in\mathbb{N}}$ is a linear combination (hence finite) of convergent sequences, and converges to corresponding linear combination of limits, $\langle s, p\theta^{-1}\rangle = \langle s\theta, p\rangle$. By definition of the topology of $\mathbb{K}\langle\langle N\rangle\rangle$, the sequence $\{s_n\theta\}$ converges to $s\theta$. ■

Exercises

2.1 Verify that the identification between $\mathbb{K}\langle\langle M\times N\rangle\rangle$ and $[\mathbb{K}\langle\langle N\rangle\rangle]\,\langle\langle M\rangle\rangle$ proposed in the proof of Proposition 2.1 is indeed an isomorphism of $\mathbb{K}$-semialgebras.

• 2.2 **Uniform convergence.** (a) Give an example of a uniformly convergent sequence $\{s_n\}_{n\in\mathbb{N}}$ of series of $\mathbb{Q}\langle\langle M\rangle\rangle$ and of a $\mathbb{Q}$-relation θ from M to N such that $(s_n)\theta$ is defined for all n and the sequence $\{s_n\theta\}_{n\in\mathbb{N}}$ *does not converge* to a series t in $\mathbb{Q}\langle\langle N\rangle\rangle$. *cf. Exer. III.1.20, p. 392*

(b) Prove:

Proposition 2.4 *Let θ be a $\mathbb{K}$-relation from M to N and $\{s_n\}_{n\in\mathbb{N}}$ a sequence of series of $\mathbb{K}\langle\langle M\rangle\rangle$ that is* uniformly convergent *to a series s, and such that $(s_n)\theta$ is defined for all n and so that the sequence $\{s_n\theta\}_{n\in\mathbb{N}}$ converges to a series t in $\mathbb{K}\langle\langle N\rangle\rangle$. Then $s\theta$ is defined and is equal to t.*

2.3 Show that if $\mathbb{K}$ is such that $1_{\mathbb{K}}+1_{\mathbb{K}}$ is not equal to $1_{\mathbb{K}}$ or $0_{\mathbb{K}}$, a characteristic $\mathbb{K}$-relation τ is injective if and only if the image under τ of every characteristic series on M is a characteristic series on N.

2.2 Composition

The composition of $\mathbb{K}$-relations is not defined as easily as the composition of relations. A simple case gives an idea of the difficulty: let θ be a function from M to N; for all n in N, we have $(n)\theta^{-1}\theta = n$ *but we do not have* $((\underline{n})\underline{\theta^{-1}})\underline{\theta} = \underline{n}$ and even, in the case where the pre-image of n is infinite and where $\mathbb{K}$ is not complete, $((\underline{n})\underline{\theta^{-1}})\underline{\theta}$ can easily be *undefined*.

For the rest of this subsubsection, $\theta\colon M\to N$ and $\sigma\colon N\to P$ are two $\mathbb{K}$-relations. To compose θ and σ, we shall compute $(m\theta)\sigma$ for each m in M. But if, according to the definition of a $\mathbb{K}$-relation, $m\theta$ is a series of $\mathbb{K}\langle\langle N\rangle\rangle$, possibly zero, and is thus 'well defined', it could be the case that σ is not defined on $m\theta$. If there exists such an m, *the composition $\theta\sigma$ is undefined*. With the above conventions, the composition $\theta\sigma\colon M\to P$ is simply written

$$\forall m\in M\,,\ \forall p\in P \qquad \langle\widehat{\theta\sigma},(m,p)\rangle = \langle(m\theta)\sigma,p\rangle$$

$$= \langle m\theta, p\sigma^{-1}\rangle \tag{2.8}$$

$$= \sum_{n\in N}\langle m\theta,n\rangle\langle n,p\sigma^{-1}\rangle\ , \tag{2.9}$$

and $\theta\sigma$ is defined if and only if the scalar product in (2.8) – that is, the infinite sum in (2.9) – is defined for every pair (m,p) in $M\times P$.

We deduce that if $\mathbb{K}$ is complete, or if we restrict ourselves to certain sub-families of $\mathbb{K}$-relations, for example $\mathbb{K}$-relations of *finite image*, or of *finite co-image*, then the composition is always defined. In the other cases, we must always check that $\theta\sigma$ is defined before stating or attempting to prove properties about it.

Example 2.2 (continued) Let s and t be two series of $\mathbb{K}_{\mathsf{c}}\langle\langle M\rangle\rangle$; we verify that

$$\iota_s\,\iota_t = \iota_t\,\iota_s = \iota_{s\odot t}\ . \qquad \blacksquare$$

We verify the classic expression of the inverse of the composition of two relations:

Property 2.4 *Let $\theta\colon M \to N$ and $\sigma\colon N \to P$ be two $\mathbb{K}$-relations such that $\theta\sigma$ is defined. Then $\sigma^{-1}\theta^{-1}$ is defined and we have*

$$(\theta\sigma)^{-1} = \sigma^{-1}\theta^{-1} \quad .$$

Proof. For every pair (m,p) of $M \times P$, we have

$$\begin{aligned} <\widehat{(\theta\sigma)^{-1}}, (m,p)> &= <\widehat{\theta\sigma}, (m,p)> = <(m\theta)\sigma, p> \\ &= <m\theta, p\sigma^{-1}> \ = <m, (p\sigma^{-1})\theta^{-1}> \qquad \text{by (2.6)} \\ &= <\widehat{\sigma^{-1}\theta^{-1}}, (m,p)> \quad . \end{aligned}$$

■

The following proposition gives a powerful expression for $\widehat{\theta\sigma}$, which we will use later.

Proposition 2.5 *Let $\theta\colon M \to N$ and $\sigma\colon N \to P$ be two $\mathbb{K}$-relations such that $\theta\sigma$ is defined; we then have*

$$\widehat{\theta\sigma} = (\underline{N})[\theta^{-1} \odot \sigma] \quad .$$

Proof. We first make sure that the right hand side of the equation makes sense; the verification of the equality is then a formal computation.

The relation θ^{-1} is an element of $[\mathbb{K}_{\mathsf{c}}\langle\langle M\rangle\rangle]\langle\langle N\rangle\rangle$, and the relation σ is an element of $[\mathbb{K}_{\mathsf{c}}\langle\langle P\rangle\rangle]\langle\langle N\rangle\rangle$; the two canonical inclusions ensure they are both elements of $[\mathbb{K}_{\mathsf{c}}\langle\langle M\times P\rangle\rangle]\langle\langle N\rangle\rangle$, that their Hadamard product $\theta^{-1}\odot\sigma$ is well defined, and that it is also a $\mathbb{K}$-relation from N to $M \times P$.

It remains to verify that $\widehat{\theta\sigma}$ is indeed the image of $\underline{N}$ under this $\mathbb{K}$-relation. It follows, for all (m,p) in $M \times P$, that

$$\begin{aligned} <\widehat{\theta\sigma}, (m,p)> &= \sum_{n\in N} <m\theta, n><n, p\sigma^{-1}> = \sum_{n\in N} <m, n\theta^{-1}><n\sigma, p> \\ &= \sum_{n\in N} <n\theta^{-1}, m><n\sigma, p> \quad . \end{aligned}$$

If we identify θ^{-1} and σ with series of $[\mathbb{K}_{\mathsf{c}}\langle\langle M \times P\rangle\rangle]\langle\langle N\rangle\rangle$, and return to the notation $n\theta^{-1} = <\theta^{-1}, n>$ and $n\sigma = <\sigma, n>$, it follows that

$$\begin{aligned} <\widehat{\theta\sigma}, (m,p)> &= \sum_{n\in N} \left[<<\theta^{-1}, n>, (m, 1_P)><<\sigma, n>, (1_M, p)>\right] \\ &= \sum_{n\in N} <<\theta^{-1}, n><\sigma, n>, (m,p)> \end{aligned}$$

since $\mathsf{supp}<\theta^{-1}, n> \subseteq M \times 1_P$ and $\mathsf{supp}<\sigma, n> \subseteq 1_M \times P$. Hence

$$<\widehat{\theta\sigma}, (m,p)> = \sum_{n\in N} <<\theta^{-1}\odot\sigma, n>, (m,p)> = <(\underline{N})[\theta^{-1}\odot\sigma], (m,p)> \quad ,$$

by definition of the image of a series under a $\mathbb{K}$-relation. ■

A similar computation yields the following more general statement.

Proposition 2.6 *Let $\theta\colon M \to N$ and $\sigma\colon N \to P$ be two $\mathbb{K}$-relations and let s be a series of $\mathbb{K}_{\mathsf{c}}\langle\langle N\rangle\rangle$. If $\theta\,\iota_s\,\sigma$ is defined, we have:*

$$\widehat{\theta\,\iota_s\,\sigma} = (s)\left[\theta^{-1}\odot\sigma\right] \quad . \qquad \blacksquare$$

The following lemma[18] is an extension of Property 1.4 to series. *cf. p. 535*

Lemma 2.7 *Let $\theta\colon M \to N$ be a function and s a $\mathbb{K}$-series on N. We have*

$$\underline{\theta}\,\iota_s = \iota_{s\underline{\theta}^{-1}}\,\underline{\theta} \quad .$$

Proof. For every m in M we have

$$\begin{aligned}(m\iota_{s\underline{\theta}^{-1}})\underline{\theta} &= (s\underline{\theta}^{-1}\odot m)\underline{\theta} = (\langle s\underline{\theta}^{-1}, m\rangle m)\underline{\theta}\\ &= \langle s, m\underline{\theta}\rangle m\underline{\theta} = s\odot m\underline{\theta} = (m\underline{\theta})\iota_s \quad . \qquad \blacksquare\end{aligned}$$

$$\begin{tikzcd} M \arrow[r, "\iota_{(s)\underline{\theta}^{-1}}"] \arrow[d, "\underline{\theta}"'] & M \arrow[d, "\underline{\theta}"] \\ N \arrow[r, "\iota_s"'] & N \end{tikzcd}$$

Exercises

2.4 Consider the $\mathbb{N}$-relations $\underline{\theta}$ and τ of Example 2.1 (where $M = N = \{a\}^*$). Compute the graph of: (a) $\underline{\theta}^{-1}\,\underline{\theta}$; (b) $\underline{\theta}\,\underline{\theta}$; (c) $\tau^{-1}\,\tau$; (d) $\tau\,\tau$.

• 2.5 Give the statement of Proposition 2.5 for the case of classic relations between words (that is, $\mathbb{K} = \mathbb{B}$).

2.6 Verify Proposition 2.6.

• 2.7 Prove:

Lemma 2.8 *Let $\theta\colon M \to N$ be a* function *and T a subset of M in bijection with N under θ. Then for all s in $\mathbb{K}\langle\langle N\rangle\rangle$, we have $s = \left[\left(s\underline{\theta^{-1}}\right)\odot\underline{T}\right]\underline{\theta}$.*

2.3 Multiplicative $\mathbb{K}$-relations

This subsection is preliminary to the study of the image under morphisms of rational $\mathbb{K}$-series, which will be crucial for that of rational $\mathbb{K}$-relations and with which we shall start the next section. We start with a proposition at a more general level than that of morphisms between monoids. A $\mathbb{K}$-relation $\theta\colon M \to N$ is *multiplicative* if its restriction to M is a morphism to $\mathbb{K}\langle\langle N\rangle\rangle$ viewed as a multiplicative monoid. The definition of $\mathbb{K}$-relations implies in fact that θ is a morphism from M to $\mathbb{K}_{\mathsf{c}}\langle\langle N\rangle\rangle$. In particular, the characteristic relation of a morphism from M to N is a multiplicative $\mathbb{K}$-relation.

Proposition 2.9 *If $\theta\colon M \to N$ is a multiplicative $\mathbb{K}$-relation, then θ is a morphism from the $\mathbb{K}$-semialgebra of $\mathbb{K}\langle M\rangle$ to $\mathbb{K}\langle\langle N\rangle\rangle$.*

Proof. By Properties 2.2, θ is a morphism on $\mathbb{K}$-modules. It therefore suffices to verify that, for all P and Q in $\mathbb{K}\langle M\rangle$, we have $(PQ)\theta = P\theta Q\theta$. This is what the following sequence of equalities does, for every n in N. *See p. 551*

[18] Which is found in Eilenberg, [80].

$$
\begin{aligned}
\langle P\theta\, Q\theta, n\rangle &= \sum_{\substack{r,s\in N,\\ rs=n}} \langle P\theta, r\rangle\langle Q\theta, s\rangle \\
&= \sum_{\substack{r,s\in N,\\ rs=n}} \left[\left(\sum_{p\in M}\langle P,p\rangle\langle p\theta, r\rangle\right)\left(\sum_{q\in M}\langle Q,q\rangle\langle q\theta, s\rangle\right)\right] \\
&= \sum_{\substack{r,s\in N,\\ rs=n}} \left[\sum_{p,q\in M}\langle P,p\rangle\langle p\theta, r\rangle\langle Q,q\rangle\langle q\theta, s\rangle\right] \\
&= \sum_{\substack{r,s\in N,\\ rs=n}} \left[\sum_{p,q\in M}\langle P,p\rangle\underbrace{\langle Q,q\rangle\langle p\theta, r\rangle}_{\text{since } \langle p\theta, r\rangle\in\mathbb{K}_{\mathsf{c}}}\langle q\theta, s\rangle\right] \\
&= \sum_{\substack{r,s\in N,\\ rs=n}} \left[\sum_{m\in M}\left(\sum_{\substack{p,q\in M,\\ pq=m}}\langle P,p\rangle\langle Q,q\rangle\langle p\theta, r\rangle\langle q\theta, s\rangle\right)\right] \\
&= \sum_{m\in M}\left[\sum_{\substack{p,q\in M,\\ pq=m}}\langle P,p\rangle\langle Q,q\rangle\left(\sum_{\substack{r,s\in N,\\ rs=n}}\langle p\theta, r\rangle\langle q\theta, s\rangle\right)\right] \\
&= \sum_{m\in M}\left[\sum_{\substack{p,q\in M,\\ pq=m}}\langle P,p\rangle\langle Q,q\rangle\langle p\theta q\theta, n\rangle\right] \\
&= \sum_{m\in M}\left[\sum_{\substack{p,q\in M,\\ pq=m}}\langle P,p\rangle\langle Q,q\rangle\right]\langle \underbrace{m\theta}_{\text{since } \theta \text{ is a multiplicative relation}}, n\rangle \\
&= \sum_{m\in M}\langle P\,Q, m\rangle\langle m\theta, n\rangle = \langle (P\,Q)\theta, n\rangle \quad . \qquad \blacksquare
\end{aligned}
$$

Proposition 2.10 *Let M and N be graded monoids of finite type and $\theta\colon M\to N$ a multiplicative $\mathbb{K}$-relation. If, for all m in M, $m\theta$ is a* proper series *of $\mathbb{K}\langle\langle N\rangle\rangle$, then θ is of finite co-image, hence is defined on all of $\mathbb{K}\langle\langle M\rangle\rangle$ and is continuous.*

Proof. The monoids M and N are graded. If $m\theta$ is a proper series, the valuation[19]

[19]The valuation of a series on a graded monoid is the minimum of the lengths of elements whose coefficient in s is non-zero; it is the 'gap' between s and the zero series (see Rem. III.1.5, p. 387).

of $m\theta$, $\mathbf{v}(m\theta)$, is by definition greater than 1; if m is the product of k elements of M, then

$$\mathbf{v}((m_1 m_2 \cdots m_k)\theta) = \mathbf{v}(m_1\theta\, m_2\theta \cdots m_k\theta) = \sum_{i=1}^{k} \mathbf{v}(m_i\theta) \geqslant k \ .$$

Hence we deduce that if $(m\theta, n)$ is different from $0_{\mathbb{K}}$, we have

$$\mathbf{v}(m\theta) \leqslant |n|$$

and hence m is the product of at most $|n|$ elements of M. Since M is finitely generated, there is only a finite number of such m. The family $\{m\theta\}_{m\in M}$ is locally finite or, which comes to the same thing, each $n\theta^{-1}$ is a polynomial: θ is of finite co-image. Proposition 2.4 then ensures the continuity of θ. ■

If M and N are two graded monoids, we say that a morphism (on monoids) is *continuous* if $m\theta \neq 1_N$ for all m in M (in fact, for all m in a generating set of M). The preceding proposition can be restated for this special case:

Corollary 2.11 *Let M and N be two graded monoids. If $\theta\colon M \to N$ is a continuous morphism of monoids, the $\mathbb{K}$-relation $\underline{\theta}$ is a continuous morphism of $\mathbb{K}$-semialgebras from $\mathbb{K}\langle\langle M\rangle\rangle$ to $\mathbb{K}\langle\langle N\rangle\rangle$.* ■

Let us end this section with the extension of Corollary II.2.2 to series. *cf. p. 237*

Proposition 2.12 *Let $\theta\colon M \to N$ be a morphism of monoids and t a $\mathbb{K}$-recognisable series on N. Then $t\underline{\theta}^{-1}$ is a $\mathbb{K}$-recognisable series on M.*

Proof. By assumption, there exists (λ, μ, ν), a $\mathbb{K}$-representation of N, such that for all n in N, $\langle t, n\rangle = (\lambda \cdot n\mu \cdot \nu)$. Whence, for all m in M,

$$\langle t\underline{\theta}^{-1}, m\rangle = \langle t, m\theta\rangle = \lambda \cdot (m\theta)\mu \cdot \nu \ .$$

Thus, the $\mathbb{K}$-representation of M $(\lambda, \theta\mu, \nu)$ recognises the series $t\underline{\theta}^{-1}$. ■

3 Rational $\mathbb{K}$-relations

We know from the previous section how to define the *graph* of a $\mathbb{K}$-relation, and we even defined $\mathbb{K}$-relations *in terms of* their graph, a series with coefficients in the *centre* of $\mathbb{K}$. The definition of a $\mathbb{K}$-rational relation is thus as logical and natural as in the set-based case treated in Section 1:

Definition 3.1 A $\mathbb{K}$-relation from M to N is *rational* if its graph is a $\mathbb{K}_{\mathsf{c}}$-rational series of $\mathbb{K}\langle\langle M \times N\rangle\rangle$. □

We shall use the terms 'rational $\mathbb{K}$-relation' and '$\mathbb{K}$-rational relation' interchangeably, as we have already done. From this definition follow several properties, as in Section 1.

Property 3.1 *The inverse of a rational $\mathbb{K}$-relation from M to N is a rational $\mathbb{K}$-relation from N to M.* ■

Property 3.2 *If τ is a rational $\mathbb{K}$-relation from M to N, there exists a finitely generated submonoid M' of M and a finitely generated submonoid N' of N such that τ is a rational $\mathbb{K}$-relation from M' to N'.*

In particular, the domain of τ is composed of series on M whose supports are contained in M'; and the image of τ is composed of series on N whose supports are contained in N'. ■

From here on, we therefore suppose that M and N are finitely generated monoids. Furthermore, and as in the previous section, either $\mathbb{K}$ is arbitrary and M and N are *graded*; or K is *continuous* and M and N are arbitrary. Finally, since we are dealing with rational series, we shall assume that $\mathbb{K}$ is a strong semiring.

cf. Def. 2.3, p. 552, Propty III.1.1, p. 382 and Propty III.2.4, p. 397

By definition, a $\mathbb{K}$-relation is *unambiguous* if its graph is a characteristic series.[20] In accordance with custom, we make the following definition.

Definition 3.2 A $\mathbb{B}$-relation from M to N is *unambiguous rational* if its characteristic series is an $\mathbb{N}$-rational series of $\mathbb{N}\langle\langle M \times N\rangle\rangle$. □

cf. Fact II.1.1, p. 231

Contrary to rational languages over A^*, rational relations from A^* to B^* are not all unambiguous.

3.1 Reasonable semirings

via the Prop. II.1.7, p. 223

The closure under morphisms of rational subsets is at the heart of the set theory of rational relations seen in Section 1. It is natural to seek to extend the method to the treatment of rational $\mathbb{K}$-relations. This step uncovers a real difficulty which will lead us to define a special class of semirings.

3.1.1 Image of series under continuous morphisms

From Corollary 2.11 we deduce:

Corollary 3.1 *Let M and N be two graded monoids and $\theta\colon M \to N$ a* continuous *morphism of monoids. If s is a series of $\mathbb{K}\langle\langle M\rangle\rangle$, the star of s is defined if and only if the star of $s\underline{\theta}$ is defined and we have $(s^*)\underline{\theta} = (s\underline{\theta})^*$.*

cf. p. 394

Proof. The forward direction is exactly Property III.2.1: if s^* is defined, it is the limit of the $s^{\leqslant n}$ and the continuity of $\underline{\theta}$ implies

$$(s^*)\underline{\theta} = \left(\lim_{n\to\infty}\sum_{k\leqslant n} s^k\right)\underline{\theta} = \lim_{n\to\infty}\sum_{k\leqslant n}(s^k)\underline{\theta} = \lim_{n\to\infty}\sum_{k\leqslant n}(s\underline{\theta})^k = (s\underline{\theta})^* \ .$$

cf. Prop. III.2.6, p. 396

Conversely, if s^* is undefined, then the star of its constant term $\mathsf{c}(s)$ is undefined.

[20] This really makes sense only if $1_{\mathbb{K}}$ generates a sub-semiring of $\mathbb{K}$ isomorphic to $\mathbb{N}$.

Then, $\mathsf{c}(s\underline{\theta}) = \mathsf{c}(s)$ since θ is continuous and N is graded: the star of $s\underline{\theta}$ is undefined as well. ■

We can now prove the first generalisation to rational series of Proposition II.1.7, with the same proof method. *cf. p. 223*

Theorem 3.1 *Let M and N be graded monoids and $\theta\colon M \to N$ a* continuous *morphism of monoids.*

(i) *If $s \in \mathbb{K}\mathrm{Rat}\, M$, then $s\underline{\theta} \in \mathbb{K}\mathrm{Rat}\, N$.*

(ii) *If θ is surjective and $t \in \mathbb{K}\mathrm{Rat}\, N$, then there exists $s \in \mathbb{K}\mathrm{Rat}\, M$ such that $s\underline{\theta} = t$.*

Proof. (i) Let $\mathcal{D}$ be the family of series in $\mathbb{K}\langle\langle M\rangle\rangle$ whose image under $\underline{\theta}$ is a $\mathbb{K}$-rational series of $\mathbb{K}\langle\langle N\rangle\rangle$ (that is, $s \in \mathcal{D} \Rightarrow s\underline{\theta} \in \mathbb{K}\mathrm{Rat}\, N$). This family contains the polynomials, and is closed under the $\mathbb{K}$-module operations by Properties 2.2, under product by Proposition 2.9 and under star by Corollary 3.1: $\mathcal{D}$ contains $\mathbb{K}\mathrm{Rat}\, M$.

(ii) Let $\mathcal{E}$ be the family of series of $\mathbb{K}\langle\langle N\rangle\rangle$ which are the image of at least one $\mathbb{K}$-rational series on M under $\underline{\theta}$ (that is, $t \in \mathcal{E} \Rightarrow \exists s \in \mathbb{K}\mathrm{Rat}\, M \quad s\underline{\theta} = t$). As above, $\mathcal{E}$ is rationally closed (for the closure under star, we use the converse part of Corollary 3.1). Since θ is surjective, $\mathcal{E}$ contains all the monomials and hence $\mathbb{K}\mathrm{Rat}\, N$. ■

3.1.2 Image of series under projections

The morphisms, and particularly the morphisms which we shall be led to consider in the study of rational $\mathbb{K}$-relations, *are not all continuous*. For a discontinuous morphism $\theta\colon M \to N$, as for every other $\mathbb{K}$-relation from M to N, it can happen that the image $s\theta$ of a series s of $\mathbb{K}\langle\langle M\rangle\rangle$ is undefined; this is not a problem. But *if $s\theta$ is defined*, and if s belongs to $\mathbb{K}\mathrm{Rat}\, M$, are we then sure that $s\theta$ belongs to $\mathbb{K}\mathrm{Rat}\, N$? This is the problem.

I do not know how to answer this question in the affirmative in all generality, even if on the other hand I can produce no counterexamples. As it will play a key role in the extension to $\mathbb{K}$-relations of the Evaluation and Composition Theorems, we shall make the corresponding property an assumption.

Definition 3.3 We call a strong semiring $\mathbb{K}$ *reasonable* if for every *projection* $\pi\colon A^* \to B^*$ and every series s in $\mathbb{K}\mathrm{Rat}\, A^*$, if $s\pi$ is defined then $s\pi$ belongs to $\mathbb{K}\mathrm{Rat}\, B^*$. □

Clearly, a continuous semiring is reasonable. More importantly, this property holds for all the (non-continuous) semirings that we use most often: $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, *etc.*, thanks to the following result. *Def. III.5.1, p. 471*

Proposition 3.2 [**Fliess**] *A sub-semiring of a skew field is a reasonable semiring.*

Proof. Let $A = B \cup C$ be the disjoint union of two alphabets and $\pi\colon A^* \to B^*$

the projection. For all $u = b_1 b_2 \cdots b_n$ in B^*, we have

$$u\underline{\pi}^{-1} = \underline{C^*}\, b_1\, \underline{C^*}\, b_2\, \underline{C^*} \cdots \underline{C^*}\, b_n\, \underline{C^*} \ . \tag{3.1}$$

Let s be an element of $\mathbb{K}\mathrm{Rat}\, A^*$ recognised by a representation (λ, μ, ν); if $s\pi$ is defined then, in particular,

$$\langle s\pi, 1_{B^*}\rangle = \langle s, 1_{B^*}\underline{\pi}^{-1}\rangle = \langle s, \underline{C^*}\rangle = \sum_{f \in C^*} \lambda \cdot f\mu \cdot \nu$$

is defined. *If the family* $\{f\mu \mid f \in C^*\}$ *is summable*, with sum U_π, then

$$\langle s\pi, 1_{B^*}\rangle = \lambda \cdot U_\pi \cdot \nu$$

and, more generally, by setting $\nu' = U_\pi \cdot \nu$ and, for all b in B, $b\mu' = U_\pi \cdot b\mu$, it follows from (3.1) that $s\pi$ is recognised by (λ, μ', ν') and is hence rational.

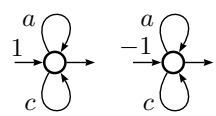
If $\mathbb{K}$ is a sub-semiring of a skew field $\mathbb{F}$, then s belongs to $\mathbb{F}\mathrm{Rat}\, A^*$ and then has a *reduced representation* (λ, μ, ν), of dimension n. By Proposition III.4.13, there exist two sets of n words, $P = \{p_1, p_2, \ldots, p_n\}$ and $Q = \{q_1, q_2, \ldots, q_n\}$, and two $n \times n$ matrices α_P and β_Q such that

cf. p. 459

$$f\mu = \alpha_P \cdot (\langle s, p_i\, f\, q_j\rangle) \cdot \beta_Q \ ,$$

for all f in A^*. For each (i, j), the family $\{\langle s, g\rangle \mid g \in (p_i\, q_j)\pi\, \pi^{-1}\}$ is summable and hence, *because* $\mathbb{F}$ *is a field*, the family $\{\langle s, p_i\, f\, q_j\rangle \mid f \in C^*\}$ is summable, hence so is the family $\{f\mu \mid f \in C^*\}$. ■

A rather simplistic example will illustrate the problem in the general case. The $\mathbb{Z}$-automaton in the adjacent figure recognises the zero series, whose image under projection $\pi\colon \{a, c\}^* \to \{a\}^*$ is clearly defined. On the other hand, the image under π of this automaton gives a $\mathbb{Z}$-automaton whose behaviour is undefined. In the case of a sub-semiring of a skew field, the reduced representation allows us, in a way, to construct an automaton from the series, an automaton in which no useless or harmful transitions are hidden as in the example above. The problem is thus to determine in which cases such a back-formation of a sufficiently 'reduced' automaton from a series is possible.



Exercises

3.1 Explain why it is sufficient to consider projections between free monoids in order to treat all discontinuous morphisms between graded morphisms.

3.2 Prove:

Proposition 3.3 *If* $\mathbb{K}$ *is a reasonable semiring, then so is* $\mathbb{K}\mathrm{Rat}\, A^*$.

3.1.3 $\mathbb{K}$-intersections

Before moving on to the recognition of rational $\mathbb{K}$-relations, we can, thanks to the theorem above, first prove a result postponed from the last chapter which generalises to series the property that *the intersection of a recognisable subset and a rational subset is rational*: *Prop. II.2.6, p. 239*

Proposition III.3.12 *The Hadamard product of two commutable $\mathbb{K}$-series on any monoid N, one $\mathbb{K}$-rational and the other $\mathbb{K}$-recognisable, is a $\mathbb{K}$-rational series on N.* *cf. p. 436*

Proof. Let t in $\mathbb{K}\mathrm{Rat}\,N$ and s in $\mathbb{K}\mathrm{Rec}\,N$ be commutable; that is, $t \in \mathbb{L}\mathrm{Rat}\,N$ and $s \in \mathbb{I}\mathrm{Rec}\,N$ where $\mathbb{I}$ and $\mathbb{L}$ are commutable sub-semirings of $\mathbb{K}$. Without loss of generality we can suppose that N is finitely generated and hence that there exists a finite alphabet A and a surjective continuous morphism $\theta\colon A^* \to N$. By Theorem 3.1(ii) there exists a series u in $\mathbb{L}\mathrm{Rat}\,A^*$ such that *cf. Exer. III.3.2, p. 426*

$$u\underline{\theta} = t \ .$$

The coefficients of u commute with those of s, and hence with those of $s\underline{\theta^{-1}}$. Lemma 2.7 allows us to rewrite the equality $s \odot t = s \odot u\underline{\theta}$ as *cf. p. 555*

$$s \odot t = \left[s\underline{\theta^{-1}} \odot u\right] \underline{\theta} \ . \tag{3.2}$$

Proposition 2.12 ensures that $s\underline{\theta^{-1}}$ is $\mathbb{I}$-recognisable (on A^*), Theorem III.3.2 that $s\underline{\theta^{-1}} \odot u$ is $\mathbb{K}$-recognisable, hence $\mathbb{K}$-rational, and finally (3.2) and Theorem 3.1 (i) that $s \odot t$ is $\mathbb{K}$-rational on N. ■ *cf. p. 428*

Recall that Proposition III.3.12 implies directly that if M is *rationally enumerable*, then $\mathbb{K}\mathrm{Rec}\,M \subseteq \mathbb{K}\mathrm{Rat}\,M$. *cf. Prop. III.3.13, p. 436; cf. also Exer. 3.3*

Exercises

- 3.3 Give a direct proof (that is, without using Proposition III.3.12) of Proposition III.3.13: *If M is rationally enumerable,* $\mathbb{K}\mathrm{Rec}\,M \subseteq \mathbb{K}\mathrm{Rat}\,M$.

- 3.4 **Support of $\mathbb{R}$-rational series on A^*.** Prove:

Proposition 3.4 *Let $\mathbb{K}$ be a sub-semiring of $\mathbb{R}$. The set of supports of the $\mathbb{K}$-rational series on A^* is closed under morphism and inverse morphism.* *cf. Prop. III.4.23*

3.2 Realisation of rational $\mathbb{K}$-relations

As in the case of rational relations between words, we shall give three types of realisation of rational $\mathbb{K}$-relations, but in a more natural order: realisation by ($\mathbb{K}$-)automata, realisation by ($\mathbb{K}$-)representations, and realisation by morphisms.

3.2.1 Realisation by $\mathbb{K}$-automaton

cf. Th. III.2.1, p. 409 and Cor. III.2.15, p. 414

The Fundamental Theorem of ($\mathbb{K}$-)automata for $\mathbb{K}$-relations gives:

Theorem 3.2 *Let E and F be generating sets of M and N respectively. A $\mathbb{K}$-relation $\tau\colon M \to N$ is rational if and only if its graph is the result of a proper finite $\mathbb{K}$-automaton whose coefficients are $\mathbb{K}_{\mathsf{c}}$-linear combinations of elements of $(E\times 1_N)\cup(1_M\times F)$.* ■

In what follows, and particularly when M and N are free monoids, we shall call a $\mathbb{K}$-automaton over the direct product $M\times N$ a $\mathbb{K}$-*transducer*. The transducers of Section 1 are thus $\mathbb{B}$-transducers.

3.2.2 Realisation by $\mathbb{K}$-representation

Proposition 3.5 *The canonical isomorphism from $\mathbb{K}_{\mathsf{c}}\langle\langle M\times N\rangle\rangle$ to $[\mathbb{K}_{\mathsf{c}}\langle\langle N\rangle\rangle]\langle\langle M\rangle\rangle$ sends $\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,(M\times N)$ to $[\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,N]\mathrm{Rat}\,M$.*

Proof. From the inclusion

$$\mathbb{K}_{\mathsf{c}}\langle N\rangle \quad\subseteq\quad \mathbb{K}_{\mathsf{c}}\langle M\times N\rangle \quad\subseteq\quad \mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,(M\times N)\ ,$$

we deduce successively, by liberal use of the canonical identities,

$$\begin{aligned}\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,N &\subseteq \mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,(M\times N)\ ,\\ [\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,N]\langle M\rangle &\subseteq \mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,(M\times N)\ ,\\ [\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,N]\mathrm{Rat}\,M &\subseteq \mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,(M\times N)\ .\end{aligned}$$

Conversely, let r be in $\mathbb{K}\mathrm{Rat}\,(M\times N)$. There exists a proper $\mathbb{K}_{\mathsf{c}}$-automaton $\langle I,X,T\rangle$ such that $r = I\cdot X^*\cdot T$ and such that the coefficients of X are finite $\mathbb{K}_{\mathsf{c}}$-linear combinations of elements of $(M\times 1)\cup(1\times N)$. We write

$$X = Y + Z\ ,$$

in such a way that the coefficients of Y are linear combinations of elements of $M\times 1$ and those of Z are linear combinations of elements of $1\times N$ (with coefficients in $\mathbb{K}_{\mathsf{c}}$). The series r is the result of the automaton $\langle I, Z^*\cdot Y, Z^*\cdot T\rangle$ (and also of the automaton $\langle I\cdot Z^*, Y\cdot Z^*, T\rangle$) whose coefficients are linear combinations of elements of $M\times 1$, with coefficients in $1\times\mathbb{K}\mathrm{Rat}\,N$. ■

cf. Prop. III.2.5, p. 396 and Prop. III.5.3, p. 473

We then deduce:

Theorem 3.3 *A $\mathbb{K}$-relation $\tau\colon A^*\to N$ is rational if and only if there exists a $(\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,N)$-representation of A^*, namely (λ,μ,ν), such that, for all f in A^*,*

$$f\tau = \lambda\cdot f\mu\cdot\nu\ .$$

Proof. By the above proposition, τ is rational (that is, $\widehat{\tau}\in\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,(A^*\times N)$) if and only if it is an element of $[\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,N]\mathrm{Rat}\,A^*$ which means, by the Kleene–Schützenberger Theorem, if and only if $\widehat{\tau}$ is an element of $[\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,N]\mathrm{Rec}\,A^*$, which is exactly what the theorem says. ■

cf. Th. III.3.5, p. 433

Note that if the coefficients of $\widehat{\tau}$, viewed as series on $A^* \times N$, belong to the centre of $\mathbb{K}$, it is not true that the coefficients of the same $\widehat{\tau}$, viewed as a series on A^*, belong to the centre of $\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\, N$. We also note, in the case of a relation on a direct product of more than two free monoids, that Theorem 3.3 enables us to *isolate one of the factors* of the product, although not to make an arbitrary partition of the factors. For example, a $\mathbb{K}$-transducer over $A^* \times B^* \times C^*$ is equivalent to a $\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,(A^* \times B^*)$-representation over C^* but not to a $\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\, A^*$-representation over $B^* \times C^*$.

3.2.3 Realisation by morphisms

We start with a very general construction.

Lemma 3.6 *Let A and B be disjoint finite alphabets and let $\varphi\colon A^* \to M$ and $\psi\colon B^* \to N$ be two surjective morphisms. We set $Z = A \cup B$ and extend φ and ψ to all of Z^* by setting $b\varphi = 1_M$ for all b in B and $a\psi = 1_N$ for all a in A. Then the morphism $\pi\colon Z^* \to M \times N$ defined by* *cf. Sec. 1.3, p. 531*

$$\forall f \in Z^* \qquad f\pi = (f\varphi, f\psi)$$

is surjective and can be written in the form

$$\underline{\pi} = \underline{\varphi} \odot \underline{\psi} \ . \tag{3.3}$$

Proof. The morphism π is surjective since A and B are disjoint. Since M and N are, canonically, two submonoids of $M \times N$ we can consider φ and ψ to be two morphisms from Z^* to $M \times N$ and write, for all f in Z^* $f\underline{\varphi} = \langle \underline{\varphi}, f \rangle = (f\underline{\varphi}, 1_N)$ and $f\underline{\psi} = \langle \underline{\psi}, f \rangle = (1_M, f\underline{\psi})$ from which

$$f\underline{\pi} = \langle \underline{\pi}, f \rangle = (f\underline{\varphi}, 1_N)\,(1_M, f\underline{\psi}) = \langle \underline{\varphi}, f \rangle \langle \underline{\psi}, f \rangle \ . \qquad \blacksquare$$

Theorem 3.4 *A $\mathbb{K}$-relation τ from M to N is rational if and only if there exists a finite alphabet Z, two morphisms $\varphi\colon Z^* \to M$, and $\psi\colon Z^* \to N$, and a $\mathbb{K}_{\mathsf{c}}$-rational series s on Z^*, such that*[21]

$$\tau = \underline{\varphi}^{-1} \circ \iota_s \circ \underline{\psi} \ .$$

Proof. By Property 3.2, we can suppose without loss of generality that M and N are finitely generated. We use the notation of Lemma 3.6: $\varphi\colon A^* \to M$ and $\psi\colon B^* \to N$ are surjective morphisms, and $Z = A \cup B$. By Theorem 3.1, there exists a series s, $\mathbb{K}_{\mathsf{c}}$-rational on Z^* and such that

$$\widehat{\tau} = s\underline{\pi} = (s)\underline{\varphi} \odot \underline{\psi} \ ;$$

$$\begin{tikzcd} Z^* \arrow[r, "\iota_s"] \arrow[d, "\underline{\varphi}"'] & Z^* \arrow[d, "\underline{\psi}"] \\ M \arrow[r, "\tau"'] & N \end{tikzcd}$$

and by Proposition 2.6 we deduce

$$\tau = \underline{\varphi}^{-1} \circ \iota_s \circ \underline{\psi} \ .$$

Conversely, if $\tau = \underline{\varphi}^{-1} \circ \iota_s \circ \underline{\psi}$, we have $\widehat{\tau} = (s)\underline{\varphi} \odot \underline{\psi}$. Now $\underline{\varphi} \odot \underline{\psi}\colon Z^* \to M \times N$ is a morphism, and hence, still by Theorem 3.1, τ is rational. $\blacksquare$ *cf. Prop. 2.6, p. 555*

[21] As in Section 1, $\circ$ symbols are added for legibility.

The case of $\mathbb{K}$-relations between free monoids makes it worth giving a special version of Theorem 3.4.

Corollary 3.7 *Let A and B be disjoint alphabets; we write Z for their union and π_A and π_B for the projections from Z onto A and B respectively. A $\mathbb{K}$-relation τ from A^* to B^* is rational if and only if there exists a $\mathbb{K}_{\mathsf{c}}$-rational series s on Z^* such that $\tau = \underline{\pi_A}^{-1} \circ \iota_s \circ \underline{\pi_B}$.* ■

3.3 Evaluation and Composition Theorems

The realisation of $\mathbb{K}$-relations by representation and by morphism give two slightly different versions of the pivotal theorems.

3.3.1 Using realisation by morphisms

Proposition 3.8 *Let $\mathbb{K}$ be a reasonable semiring. Let τ be a $\mathbb{K}$-rational relation from M to N. The image under τ of a $\mathbb{K}$-recognisable series of M, if it is defined, is a $\mathbb{K}$-rational series of N.*

Proof. Let $\varphi\colon Z^* \to M$, $\psi\colon Z^* \to N$ and $s \in \mathbb{K}_{\mathsf{c}}\mathrm{Rat}\, Z^*$ be a realisation of τ by morphisms. Let r be a series of $\mathbb{K}\langle\langle M\rangle\rangle$; there follows

$$r\tau = (r)\left[\underline{\varphi}^{-1} \circ \iota_s \circ \underline{\psi}\right]$$

and, since $t = (r)\left[\underline{\varphi}^{-1} \circ \iota_s\right]$ is defined for every series r, $r\tau$ is defined if and only if $t\underline{\psi}$ is. We then note that ψ factorises obviously into the product of a projection by a continuous morphism.

cf. Prop. 2.12, p. 557 If r is $\mathbb{K}$-recognisable on M, then so is $r\underline{\varphi}^{-1}$ on A^*, and hence also

$$t = (r)\left[\underline{\varphi}^{-1} \circ \iota_s\right] = (r\underline{\varphi}^{-1}) \odot s \ ,$$

cf. Th. III.3.2, p. 428 since s is commutable with every $\mathbb{K}$-series. Hence t is $\mathbb{K}$-rational on Z^* and $r\tau = t\underline{\psi}$ is $\mathbb{K}$-rational on N if it is defined. ■

The Kleene–Schützenberger Theorem enables us to specialise Proposition 3.8 to free monoids.

Corollary 3.9 *Let $\mathbb{K}$ be a reasonable semiring. The image under a $\mathbb{K}$-rational relation from A^* to B^* of a $\mathbb{K}$-rational series on A^*, if it is defined, is a $\mathbb{K}$-rational series on B^*.* ■

The following result extends the Composition Theorem to $\mathbb{K}$-relations. The reader will appreciate that while the new statement is a simple copy of the old, the proof given, which is based on the results of Chapter III, is incomparably more elegant and concise than that of its model.

Theorem 3.5 *Let $\mathbb{K}$ be a reasonable semiring. Let $\theta\colon M \to B^*$ and $\sigma\colon B^* \to P$ be two rational $\mathbb{K}$-relations. Then the composition $\theta\sigma\colon M \to P$, if it is defined, is a rational $\mathbb{K}$-relation.*

Proof. By Proposition 2.5, we have

$$\widehat{\theta\sigma} = (\underline{B^*})\theta^{-1} \odot \sigma \ .$$

We can suppose without loss of generality that B is finite since θ and σ are rational. By Proposition 3.5, it follows that

$$\theta^{-1} \in (\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\, M)\mathrm{Rat}\, B^* \qquad \text{and} \qquad \sigma \in (\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\, P)\mathrm{Rat}\, B^* \ ,$$

and by Theorem 3.3

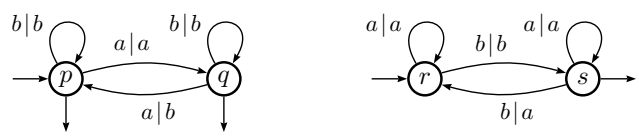
$$\theta^{-1} \in (\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\, M)\mathrm{Rec}\, B^* \qquad \text{and} \qquad \sigma \in (\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\, P)\mathrm{Rec}\, B^* \ .$$

Since $\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\, M$ and $\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\, P$ are (identified with) two *commutable* sub-semirings of $\mathbb{K}\mathrm{Rat}\,(M \times P)$, we have *cf. Th. III.3.2, p. 428*

$$\theta^{-1} \odot \sigma \in (\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,(M \times P))\mathrm{Rec}\, B^* \ .$$

Applying Theorem 3.3 once more, we obtain that $\theta^{-1} \odot \sigma$ is a rational $\mathbb{K}$-relation from B^* to $M \times P$. The image of the $\mathbb{K}$-rational series $\underline{B^*}$ under this rational $\mathbb{K}$-relation, the graph of $\theta\sigma$, if it is defined, is a $\mathbb{K}$-rational series on $M \times P$. ■

Example 3.1 A special case where the proof of Theorem 3.5 translates directly into a construction of a ($\mathbb{K}$-)transducer is that of the composition of letter-to-letter transducers. Let us observe it in the composition of α_1, a relation which replaces every second a by a b, and β_1, which replaces every other b by an a and which accepts only words with an odd number of b's, realised respectively by the transducers $\mathcal{T}_1$ and $\mathcal{S}_1$ below: *cf. Exer. 1.26*



We take $\mathcal{T}_1'$ to be the transducer which realises the inverse of α_1, by swapping the components of the labels in $\mathcal{T}_1$, and rewrite the transducers $\mathcal{T}_1'$ and $\mathcal{S}_1$ from A^* to B^* as $(\mathbb{B}\langle\langle A^* \times A^* \rangle\rangle)$-automata on A^*, on the left and at the top of Figure 3.1. The product of these two automata, at the centre of the figure, realises the Hadamard product of the two series and if we keep only the coefficients of the transitions we obtain the set of coefficients of the series recognised and hence, by Proposition 2.5, the graph of $\alpha_1\beta_1$. □ *cf. Prop. III.3.11* *cf. also Sec. V.1.1.*

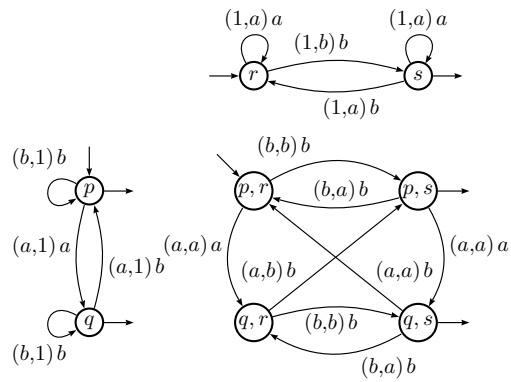


Figure 3.1: Graph of $\alpha_1 \beta_1$

3.3.2 Using realisation by representation

We shall give a more precise statement of Theorem 3.5 in the case where (with the notations of the previous subsubsection) the monoid M is a free monoid. The Kleene–Schützenberger Theorem then enables us to give a realisation by representation of each of the relations θ and σ; we shall describe the construction, from these two representations, of the realisation by representation of the product $\theta\sigma$.

Given μ, a $\mathbb{K}\langle\langle N\rangle\rangle$-representation of M, and κ, a $\mathbb{K}_{\mathsf{c}}\langle\langle P\rangle\rangle$-representation of N, we want to 'compose' these two representations, to give a meaning to the writing $(m\mu)\kappa$, for all m in M. The simplest solution is not just the best: it is *the* solution. We take the image under κ of each coefficient of $m\mu$ and by juxtaposing these blocks form the matrix $(m\mu)\kappa$. It is the aim of the following lemma to prove that we have indeed thus constructed a $\mathbb{K}\langle\langle P\rangle\rangle$-representation of M (and that we can define this construction formally).

Lemma 3.10 *Let Q and R be finite sets and let*

$$\mu\colon M \to \mathbb{K}\langle\langle N\rangle\rangle^{Q\times Q} \quad \textit{and} \quad \kappa\colon N \to \mathbb{K}_{\mathsf{c}}\langle\langle P\rangle\rangle^{R\times R}$$

be two morphisms (of monoids). If $(m\mu_{q,s})\,\kappa$ is defined for all m in M and all (q,s) in $Q\times Q$, then the map

$$\pi\colon M \to \mathbb{K}\langle\langle P\rangle\rangle^{(Q\times R)\times(Q\times R)} \quad ,$$

defined by

$$m\pi_{(q,r),(s,t)} = (m\mu_{q,s})\kappa_{r,t} \quad , \tag{3.4}$$

is a morphism (on monoids).

Proof. Let us rewrite (3.4) in order to make the block decomposition of the matrix $m\pi$ apparent,

$$m\pi_{(q,R),(s,R)} = (m\mu_{q,s})\kappa \tag{3.4'}$$

and use this block decomposition to prove the required property:

$$\begin{aligned}\forall m,n\in M \qquad (m\,n)\pi_{(q,R),(s,R)} &= ((m\,n)\mu_{q,s})\kappa = \left(\sum_{u\in Q} m\mu_{q,u}\,n\mu_{u,s}\right)\kappa \\ &= \sum_{u\in Q}(m\mu_{q,u})\kappa\,(n\mu_{u,s})\kappa \\ &= \sum_{u\in Q} m\pi_{(q,R),(u,R)}\,n\pi_{(u,R),(s,R)} \\ &= (m\pi\,n\pi)_{(q,R),(s,R)} \quad . \end{aligned}$$

■

We note that only the map κ needs to be multiplicative in this lemma. The morphism (or representation) π thus defined from the representations μ and κ is called the *composition product of* μ *by* κ (beware! the order is important and this product is rabidly non-commutative) and is written

$$\pi = \mu\circ\kappa \quad .$$

The composition product of two representations (λ,μ,ν) and (η,κ,χ) is then defined most naturally. Let (λ,μ,ν) be a $\mathbb{K}\langle\langle N\rangle\rangle$-representation (of dimension Q) of M and (η,κ,χ) be a $\mathbb{K}_{\mathsf{c}}\langle\langle P\rangle\rangle$-representation (of dimension R) of N. *The composition product of* (λ,μ,ν) *by* (η,κ,χ),

$$(\zeta,\pi,\omega) = (\lambda,\mu,\nu)\circ(\eta,\kappa,\chi) \quad ,$$

is the $\mathbb{K}\langle\langle P\rangle\rangle$-representation (of dimension $Q\times R$) of M defined by

$$\begin{gathered}\pi = \mu\circ\kappa \quad , \\ \forall q\in Q \qquad \zeta_{(q,R)} = \eta\cdot(\lambda_q)\kappa \qquad \text{and} \qquad \omega_{(q,R)} = (\nu_q)\kappa\cdot\chi \quad ,\end{gathered} \tag{3.5}$$

if $\pi = \mu\circ\kappa$ is defined and if $(\lambda_q)\kappa$ and $(\nu_q)\kappa$ are defined for all q in Q. Before moving on to the Composition Theorem, we must *use* the Evaluation Theorem seen in the previous subsubsection.

Lemma 3.11 *Let $\mathbb{K}$ be a reasonable semiring. Let s be in $\mathbb{K}\mathrm{Rat}\,A^*$ and $\kappa\colon A^*\to(\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,P)^{R\times R}$ be a morphism. If $s\kappa$ is defined, then $s\kappa$ belongs to $(\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,P)^{R\times R}$.*

Proof. For each (r,t) in $R\times R$, the $\mathbb{K}$-relation $\kappa_{r,t}$ which relates each f in A^* to the coefficient (r,t) of $f\kappa$ is a rational $\mathbb{K}$-relation from A^* to P and $s\kappa$ is defined if and only if all the $s\kappa_{r,t}$ are defined. In this case, and since $\mathbb{K}$ is reasonable, each $s\kappa_{r,t}$ belongs to $\mathbb{K}\mathrm{Rat}\,P$. ■

We can thus state:

Theorem 3.6 *Let $\mathbb{K}$ be a reasonable semiring. Let $\theta\colon A^*\to B^*$ and $\sigma\colon B^*\to P$ be two rational $\mathbb{K}$-relations respectively realised by the $\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,B^*$-representation (λ,μ,ν) and the $\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,P$-representation (η,κ,χ). The composition product $\theta\sigma$ is defined if and only if the composition product $(\zeta,\pi,\omega) = (\lambda,\mu,\nu)\circ(\eta,\kappa,\chi)$ is defined, and in this case $\theta\sigma$ is realised by (ζ,π,ω), which is a representation from A^* to $\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\,P$.*

Proof. A formal computation, since everything has been arranged in advance. By definition

$$\forall m \in M \qquad m\theta = \sum_{q,s\in Q} \lambda_q\, m\mu_{q,s}\, \nu_s \quad ,$$

from which

$$\begin{aligned}\forall m \in M \qquad m\theta\,\sigma &= \left(\sum_{q,s\in Q} \lambda_q\, m\mu_{q,s}\, \nu_s\right)\sigma = \sum_{q,s\in Q} (\lambda_q\, m\mu_{q,s}\, \nu_s)\sigma \\ &= \sum_{q,s\in Q} \eta\cdot(\lambda_q\, m\mu_{q,s}\, \nu_s)\kappa\cdot\chi \\ &= \sum_{q,s\in Q} \eta\cdot(\lambda_q)\kappa\cdot(m\mu_{q,s})\kappa\cdot(\nu_s)\kappa\cdot\chi \\ &= \sum_{q,s\in Q} \zeta_{(q,R)}\cdot m\pi_{(q,R),(s,R)}\cdot\omega_{(s,R)} = \zeta\cdot m\pi\cdot\omega \quad . \end{aligned}$$

■

Theorem 3.6 enables us to return to the Evaluation Theorem, not to prove it, since we used it in Lemma 3.11, but to give a method of effective computation.

Let s be in $\mathbb{K}\mathrm{Rat}\, A^*$, realised by the $\mathbb{K}$-representation (λ,μ,ν), and $\theta\colon A^* \to P$ a $\mathbb{K}$-relation realised by the $\mathbb{K}_{\mathsf{c}}\mathrm{Rat}\, P$-representation (η,κ,χ). Then

$$s = \sum_{f\in A^*} (\lambda\cdot f\mu\cdot\nu)\, f = \sum_{f\in A^*} (\lambda\cdot f\mu'\cdot\nu) \quad ,$$

by which, if $s\theta$ is defined,

$$s\theta = \eta\cdot\left[\sum_{f\in A^*} (\lambda\cdot f\mu'\cdot\nu)\right]\kappa\cdot\chi = \sum_{f\in A^*} (\zeta\cdot f\pi'\cdot\omega)$$

with $a\mu' = a\mu\, a$ for all a in A and by setting $(\zeta,\pi',\omega) = (\lambda,\mu',\nu)\circ(\eta,\kappa,\chi)$. That is to say, $s\theta$ is equal to the behaviour of the $\mathbb{K}$-automaton over P obtained by constructing the automaton corresponding to the representation (ζ,π',ω) and by keeping as the label of each transition only the coefficients in $\mathbb{K}\mathrm{Rat}\, P$ (that is, by deleting the letters of A).

Exercise

• 3.5 Let $(\zeta,\pi,\omega) = (\lambda,\mu,\nu)\circ(\eta,\kappa,\chi)$, a composition product of representations. Verify that if all the coordinates of ν and χ (resp. of λ and η) are equal to 1, then so are the coordinates of ω (resp. of ζ).

4 Equivalence of finite $\mathbb{K}$-transducers

Given two rational relations, specified by rational expressions or finite transducers, which is the same thing by the Fundamental Theorem, can we decide if they are equivalent? In the first two subsections, we shall show that this problem is undecidable for rational relations on $A^*\times B^*$ (and hence over $A_1^*\times A_2^*\times\cdots\times A_n^*$) as well as several

other related undecidability results. In the following subsection, we shall prove that this problem is decidable for rational relations on $A_1^* \times A_2^* \times \cdots \times A_n^*$, with weights in $\mathbb{N}$ (in fact, in every sub-semiring of a skew field). The proof of the results in algebra on which this demonstration is based is given in the supplementary Section 7.

Recall that we are going to reduce the problems that we want to show to be undecidable to the Post Correspondence Problem, which is naturally expressed in terms of morphisms: *cf. Sec. 0.8, p. 34*

Theorem 0.8.2 *Let θ and μ be morphisms from A^* to B^*. It is undecidable whether there exists a word f in A^* such that $f\theta = f\mu$.* ■

For the rest of this section we shall use the following notation for the graph of a morphism (or of a function) $\theta\colon A^* \to B^*$:

$$U_\theta = \widehat{\theta} = \{(f, f\theta) \mid f \in A^*\} \ .$$

Recall also that if E is a set and $\mathcal{F}$ is a family of subsets, *with an effective description*, of E (and to which E belongs), *intersection*, *universality*, *equivalence* and *inclusion* are said to be decidable (or undecidable) in $\mathcal{F}$ if it is decidable (or undecidable) respectively whether the intersection of two elements of $\mathcal{F}$ is empty, whether an element of $\mathcal{F}$ is equal to E, whether two elements of $\mathcal{F}$ are equal, or whether one element of $\mathcal{F}$ is included in another.

If inclusion is decidable then so is equivalence and hence universality. On the other hand, if universality is undecidable, so are equivalence and inclusion. However, it is possible for inclusion to be undecidable and equivalence to be decidable. *cf. Sec. 5.4, p. 588*

If $\mathcal{F}$ and $\mathcal{G}$ are effectively described families of subsets of a set E such that $\mathcal{G} \subseteq \mathcal{F}$, we say that $\mathcal{G}$ *is decidable in* $\mathcal{F}$ if it is decidable whether an effectively given element of $\mathcal{F}$ belongs to $\mathcal{G}$.

In the rest of this section (and of this chapter), E will be equal to $A^* \times B^*$, $\mathcal{F}$ to $\mathrm{Rat}\, A^* \times B^*$, and $\mathcal{G}$ to a family of rational relations (recognisable, synchronous, deterministic).

4.1 Equivalence of $\mathbb{B}$-transducers, general case

In this subsection we are interested in automata over $A^* \times B^*$, or, more generally, over $A_1^* \times A_2^* \times \cdots \times A_n^*$, which hence accept *subsets* of $A^* \times B^*$ (resp. of $A_1^* \times A_2^* \times \cdots \times A_n^*$): that is, transducers which one could say are *unweighted*. Once the Post Correspondence Problem has been formulated in terms of the intersection of two morphisms, Post's Theorem implies:

Theorem 4.1 [**Rabin–Scott**] *If A and B are two alphabets of at least two letters, intersection is undecidable in* $\mathrm{Rat}\, A^* \times B^*$ *(and hence in* $\mathrm{Rat}\, A_1^* \times A_2^* \times \cdots \times A_n^*$*).*

Proof. Let k be an integer such that (PCP_k) is undecidable. If A has k elements or more, the graphs U_θ and U_μ of two morphisms $\theta\colon A^* \to B^*$ and $\mu\colon A^* \to B^*$ are

rational subsets of $A^* \times B^*$. To determine whether their intersection is empty or not is an instance of the Post Correspondence Problem, hence undecidable. It remains to free ourselves from the constraint on the cardinal of the alphabet A.

Let $\mathbf{k} = \{0, 1, \ldots, k-1\}$ be a canonical alphabet with k letters and $\kappa \colon \mathbf{k}^* \to A^*$ be an injective rational function (we can take for example the morphism defined by $i\kappa = a^i b$ for all i in $\mathbf{k}$).

Let, as above, θ and μ be two morphisms from $\mathbf{k}^*$ to B^*. We write V_θ (resp. V_μ) for the graph of $\kappa^{-1}\theta$ (resp. the graph of $\kappa^{-1}\mu$). Since κ is *injective*, $\kappa^{-1}\theta$ (and $\kappa^{-1}\mu$) are two functions (partial, of course) from A^* to B^* and we have

$$V_\theta \cap V_\mu \neq \emptyset \quad \Longleftrightarrow \quad U_\theta \cap U_\mu \neq \emptyset \ .$$

$$\begin{tikzcd} A^* \arrow[r, dashed] \arrow[d, "\kappa^{-1}"'] & B^* \\ \mathbf{k}^* \arrow[ur, shift left, "\theta"] \arrow[ur, shift right, "\mu"'] & \end{tikzcd}$$

Indeed

$$\begin{aligned} &(f,g) \in V_\theta \cap V_\mu \quad \Longrightarrow \\ &\quad \exists h, h' \in \mathbf{k}^* \qquad h\kappa = f \quad \text{and} \quad h\theta = g \ , \quad h'\kappa = f \quad \text{and} \quad h'\mu = g \quad \Longrightarrow \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad h = h' \quad \text{and} \quad (h,g) \in U_\theta \cap U_\mu \ . \end{aligned}$$

and the opposite implication is obvious. ■

Theorem 4.2 *If A and B are two alphabets of at least two letters, universality is undecidable in* $\mathrm{Rat}\, A^* \times B^*$ *(and hence in* $\mathrm{Rat}\, A_1^* \times A_2^* \times \cdots \times A_n^*$ *).*

Proof. We reuse the notation of the previous proof: V_θ and V_μ are the graphs of two *rational functions* from A^* to B^*. Their complements $\complement V_\theta$ and $\complement V_\mu$ are also rational subsets of $A^* \times B^*$, hence also $\complement V_\theta \cup \complement V_\mu$. Then

cf. Cor. 1.5, p. 536

$$\complement V_\theta \cup \complement V_\mu = \complement\,(V_\theta \cap V_\mu) = A^* \times B^* \quad \Longleftrightarrow \quad V_\theta \cap V_\mu = \emptyset \quad \Longleftrightarrow \quad U_\theta \cap U_\mu = \emptyset \ .$$

Hence the universality problem in $\mathrm{Rat}\, A^* \times B^*$ is equivalent to the Post Correspondence Problem. ■

Corollary 4.1 *If A and B are alphabets with at least two letters, equivalence is undecidable in* $\mathrm{Rat}\, A^* \times B^*$ *(and hence in* $\mathrm{Rat}\, A_1^* \times A_2^* \times \cdots \times A_n^*$ *).* ■

We also reduce the problem of the decidability of $\mathrm{Rec}\, A^* \times B^*$ in $\mathrm{Rat}\, A^* \times B^*$ to the Post Problem.

Corollary 4.2 *If A and B are alphabets with at least two letters,* $\mathrm{Rec}\, A^* \times B^*$ *is undecidable in* $\mathrm{Rat}\, A^* \times B^*$ *.*

Proof. If $\complement V_\theta \cup \complement V_\mu$ is recognisable, its complement $V_\theta \cap V_\mu$ is also. However, $V_\theta \cap V_\mu$ is the graph of a function, as the intersection of the graph of two functions. On the other hand, if it is not empty, it is of infinite image, since if $(f, f\theta)$ belongs to $V_\theta \cap V_\mu$, then so does $(f^n, (f^n)\theta)$. This contradicts Property 1.9.

cf. p. 540

Thus, $\complement V_\theta \cup \complement V_\mu$ is recognisable if and only if $V_\theta \cap V_\mu$ is empty, which we know is equivalent to the Post Correspondence Problem. ■

Exercises

4.1 Show that it is undecidable whether the complement of a rational relation is finite.

4.2 Show that it is undecidable whether a rational relation is a (rational) substitution.

4.3 Show that it is undecidable whether the intersection of two rational relations is a rational relation.

4.2 Equivalence of $\mathbb{B}$-transducers, case of small alphabets

All the above only holds if the two alphabets A and B are big enough to ensure that it is possible to construct undecidable Post Correspondence Problems with rational functions from A^* to B^*. It therefore remains to examine the case in which this condition is not satisfied.

First of all, if $\|A\| = \|B\| = 1$, then $A^* \times B^*$ is isomorphic to the commutative monoid $\mathbb{N}^2$ and the equivalence and decidability problems were resolved positively in Chapter II: $\mathrm{Rat}\,\mathbb{N}^k$ *is an effective Boolean algebra* and $\mathrm{Rec}\,\mathbb{N}^k$ *is decidable in* $\mathrm{Rat}\,\mathbb{N}^k$.

cf. Sec. II.7, p. 323, Th. II.7.3 and Prop. II.7.23

Hence the only case left is $\|A\| \geqslant 2$, $\|B\| = 1$ (or the obviously equivalent case $\|A\| = 1$ and $\|B\| \geqslant 2$). The preceding *constructions* are no longer possible since all Post Correspondence Problems on an alphabet of one letter are decidable. The preceding *results* are not all true either since (a result which we shall prove later):

Proposition 4.3 *For all A, intersection is decidable in* $\mathrm{Rat}\,(A^* \times c^*)$.

However, the principal result, namely the undecidability of universality, is still true in this case. We shall once more reduce universality to the Post Correspondence Problem, but by means of a more sophisticated construction than that of the previous subsection.

Let $\theta\colon X^* \to Y^*$ be a morphism. We set

$$W_\theta = \{(f\,(f\theta), c^{|f\theta|}) \mid f \in X^*\} \ .$$

Thus, with $A = X \cup Y$, we have $W_\theta \subseteq X^*Y^* \times c^* \subseteq A^* \times c^*$. Obviously, W_θ is not a rational set. All the subtlety of the construction consists of the proof that its complement is:

Lemma 4.4 [**Lisovik**] $\complement W_\theta \in \mathrm{Rat}\,(A^* \times c^*)$.

Proof. First we write

$$R = X^*Y^* \times c^* = \{(f\,g, c^n) \mid f \in X^*,\ g \in Y^* \quad n \in \mathbb{N}\} \ .$$

This set is recognisable and, by definition, $A^* \times c^* = R \cup \complement R$. We define two partitions of R: S_θ, L_θ and E_θ on one hand:

$$S_\theta = \{(f\,g, c^n) \mid f \in X^*,\ g \in Y^* \quad n < |f\theta|\} \ ,$$
$$L_\theta = \{(f\,g, c^n) \mid f \in X^*,\ g \in Y^* \quad n > |f\theta|\} \ ,$$
$$E_\theta = \{(f\,g, c^n) \mid f \in X^*,\ g \in Y^* \quad n = |f\theta|\} \ ,$$

and S'_θ, L'_θ and E'_θ on the other:

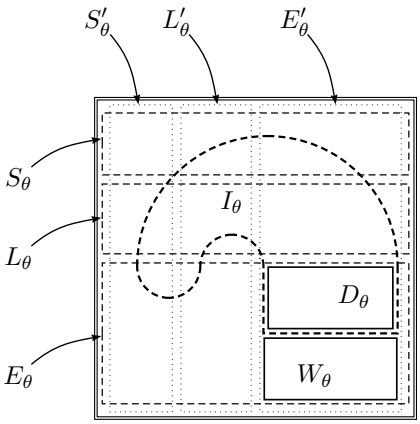


Fig. 4.1: Lisovik's construction

$$S'_\theta = \{(f\,g, c^n) \mid f \in X^*\,,\ g \in Y^* \quad n < |g|\}\ ,$$
$$L'_\theta = \{(f\,g, c^n) \mid f \in X^*\,,\ g \in Y^* \quad n > |g|\}\ ,$$
$$E'_\theta = \{(f\,g, c^n) \mid f \in X^*\,,\ g \in Y^* \quad n = |g|\}\ .$$

Obviously these six sets are rational and

$$A^* \times c^* = \complement R \cup S_\theta \cup L_\theta \cup E_\theta = \complement R \cup S'_\theta \cup L'_\theta \cup E'_\theta\ .$$

Let us rewrite W_θ as

$$W_\theta = \{(f\,g, c^n) \mid f \in X^*\,,\ g \in Y^* \quad g = f\theta \quad n = |f\theta| = |g|\}$$

and write

$$D_\theta = \{(f\,g, c^n) \mid f \in X^*\,,\ g \in Y^* \quad g \neq f\theta \quad n = |f\theta| = |g|\}\ .$$

We have $\quad W_\theta \cap D_\theta = \emptyset\ ,\quad W_\theta \cup D_\theta = E_\theta \cap E'_\theta\,,\quad$ and hence

$$\complement W_\theta = \complement R \cup S_\theta \cup L_\theta \cup S'_\theta \cup L'_\theta \cup D_\theta\ .$$

The set D_θ is clearly non-rational. The crux of the proof is in the definition of a new set, I_θ, which is rational, and which can replace D_θ in the preceding equality; that is,

$$\complement W_\theta = \complement R \cup S_\theta \cup L_\theta \cup S'_\theta \cup L'_\theta \cup I_\theta\ . \tag{4.1}$$

We first set, for all x in X,

$$C_{(x,\theta)} = [Y^{|x\theta|} \setminus \{x\theta\}]\ ;$$

that is, $C_{(x,\theta)}$ is the set of words in Y^* of the same length as $x\theta$ but different from $x\theta$. We shall show that

$$I_\theta = \left[\sum_{x\in X}(x, c^{|x\theta|})\right]^* \sum_{x\in X}\left[(x,1)(X^*,1)(Y^*,1)(C_{(x,\theta)}, c^{|x\theta|})\right]\left[\sum_{y\in Y}(y,c)\right]^*$$

satisfies (4.1), or rather, equivalently,

$$D_\theta = E_\theta \cap E'_\theta \cap I_\theta\ . \tag{4.2}$$

First we show that $D_\theta \subseteq I_\theta$. Let $(f\,g, c^n)$ be in D_θ; f is different from 1_{X^*} and $f = f_1 f_2 \cdots f_k$ with f_i in X for all i. By assumption (definition of D_θ), we can factorise g as

$$g = g_1 g_2 \cdots g_k \qquad \text{with} \qquad g_i \in Y^* \quad \text{and} \quad |g_i| = |f_i\theta|\ .$$

and *there exists an i such that* $g_i \neq f_i\theta$. We can then write

$$(f\,g, c^n) = (f_1 \cdots f_{i-1}, c^{|(f_1\cdots f_{i-1})\theta|})$$
$$\left[(f_i,1)(f_{i+1}\cdots f_k,1)(g_1\cdots g_{i-1},1)(g_i, c^{|f_i\theta|})\right](g_{i+1}\cdots g_k, c^{|g_{i+1}\cdots g_k|})$$

which shows that $(f\,g, c^n)$ is in I_θ.

Conversely, let $(f\,g, c^n)$ be in $E_\theta \cap E'_\theta \cap I_\theta$. Since $(f\,g, c^n)$ is in I_θ, we can write

$$(f\,g, c^n) = (u\,x\,v\,w\,t\,z, c^{i+j+k}) \ ,$$

with $f = u\,x\,v$, $x \in X$, $g = w\,t\,z$, $i = |u\theta|$, $j = |x\theta| = |t|$, $t \neq x\theta$, and $k = |z|$.

Since $(f\,g, c^n)$ is in E_θ, $|u\theta\,x\theta\,v\theta| = i + j + k$ and thus $|v\theta| = k$. Since $(f\,g, c^n)$ is in E'_θ, $|w\,t\,z| = i + j + k$ and thus $|w| = |u\theta|$.

Suppose that $(f\,g, c^n)$ is in W_θ, i.e. that $g = f\theta$, so $w\,t\,z = u\theta\,x\theta\,v\theta$. From $|w| = |u\theta|$ we obtain $w = u\theta$, which means that $|t| = |x\theta|$ implies $t = x\theta$, which is a contradiction; $(f\,g, c^n)$ belongs to D_θ. ■

We then deduce:

Theorem 4.3 *If A is an alphabet with at least two letters, universality is undecidable in* $\mathrm{Rat}\,(A^* \times c^*)$.

Proof. We can reuse the scheme of the proof of Theorem 4.2. Starting with two morphisms $\theta\colon X^* \to Y^*$ and $\mu\colon X^* \to Y^*$, we set $Z = X \cup Y$ and we construct the subsets W_θ and W_μ of $Z^* \times c^*$. The relation (whose graph is) $\complement W_\theta \cup \complement W_\mu$ is rational by Lemma 4.4. The following equivalences – the second, which is straightforward, being left to the reader – prove the result for alphabets with at least 8 letters (6 for X and 2 for Y):

$$\complement W_\theta \cup \complement W_\mu = Z^* \times c^* \quad \Longleftrightarrow \quad W_\theta \cap W_\mu = \emptyset \quad \Longleftrightarrow \quad U_\theta \cap U_\mu = \emptyset \ .$$

We need then to consider a rational injective map from $\{a, b\}^*$ to Z^*, as well as to verify that the construction of the proof of Lemma 4.4 works over that coding in order to conclude in full generality. ■

Hence:

Corollary 4.5 *If A is an alphabet with at least two letters, equivalence is undecidable in* $\mathrm{Rat}\,(A^* \times c^*)$. ■

By repeating the proof of Corollary 4.2, we prove:

Corollary 4.6 *If A is an alphabet with at least two letters,* $\mathrm{Rec}\,(A^* \times c^*)$ *is undecidable in* $\mathrm{Rat}\,(A^* \times c^*)$. ■

We now return to the intersection problem.

Proof of Proposition 4.3. Let K and L be two rationals of $A^* \times c^*$ and $\mathcal{T} = \langle\, Q, A, c, E, I, T \,\rangle$ and $\mathcal{S} = \langle\, R, A, c, F, J, U \,\rangle$ two real-time transducers which realise them. We construct the *product* of these two transducers,

$$\mathcal{T} \times \mathcal{S} = \langle\, Q \times R, A, c^* \times c^*, H, I \times J, T \times U \,\rangle \ ,$$

whose transition set H is defined by

$$H = \{\Big((p, r), (a, (c^k, c^l)), (q, s)\Big) \mid \big(p, (a, c^k), q\big) \in E \quad \text{and} \quad \big(r, (a, c^l), s\big) \in F\} \ .$$

We then project this transducer onto $\mathcal{V}$, an automaton *over* $\mathbb{Z}$ (viewed as additive monoid)

$$\mathcal{V} = \langle\, Q\times R, \mathbb{Z}, G, I\times J, T\times U \,\rangle$$

whose transition set G is defined by

$$G = \{\big((p,r), z, (q,s)\big) \mid \exists a \in A \quad \big(p,(a,c^k),q\big) \in E \ , \quad \big(r,(a,c^l),s\big) \in F \quad \text{and} \quad z = k-l\} \ .$$

An element (f, c^n) of $A^*\times c^*$ belongs to the intersection $K\cap L$ if and only if there exists in $\mathcal{T}\times\mathcal{S}$ a successful computation whose label is $(f,(c^n,c^n))$. From this we deduce that there exists such an (f,c^n) in $K\cap L$ if and only if there exists a successful computation in $\mathcal{V}$ whose label is 0; that is, if and only if $|\mathcal{V}|$ contains the identity of $\mathbb{Z}$, which is decidable. ■

cf. Cor. II.6.21, p. 317

Exercises

- 4.4 Prove Corollary 4.6.
- 4.5 Explain why the fact that we can decide if the label of a computation of $\mathcal{V}$ is equal to 0 (in the proof of Proposition 4.3) does not contradict the fact that it is undecidable whether a $\mathbb{Z}$-rational series on A^* has a coefficient equal to 0.

4.6 Show that equivalence in rational series on A^* with coefficients in $\mathfrak{P}_{\mathsf{fin}}(\mathbb{N})$ (the semiring of finite subsets of $\mathbb{N}$) is undecidable.

4.3 Equivalence of $\mathbb{N}$-transducers

We shall now show the *decidability* of the equivalence of automata over $A^*\times B^*$, or more generally of automata over $A_1^*\times A_2^*\times\cdots\times A_k^*$, *with weights* (in $\mathbb{N}$); that is, of $\mathbb{N}$-automata over $A_1^*\times A_2^*\times\cdots\times A_k^*$, or also of what could be called 'k-tape $\mathbb{N}$-automata':

cf. Sec. 1.6, p. 545

Theorem 4.4 [**Harju and Karhumäki**] *Equivalence in* $\mathbb{N}\mathrm{Rat}\,(A_1^*\times\cdots\times A_k^*)$ *is decidable.*

We know (Theorem 3.3) that an $\mathbb{N}$-automaton over $A^*\times B^*$ (resp. an $\mathbb{N}$-automaton over $A_1^*\times A_2^*\times\cdots\times A_k^*$) can be viewed as an automaton over A^* with weights in $\mathbb{N}\mathrm{Rat}\,B^*$ (resp. an automaton over A_1^* with weights in $\mathbb{N}\mathrm{Rat}\,(A_2^*\times\cdots\times A_k^*)$). If we now recall that *the equivalence of two automata over A^* with weights in a sub-semiring of a skew field is decidable*, then Theorem 4.4 follows immediately from:

cf. Cor. III.4.18, p. 461

Theorem 4.5 *The semiring* $\mathbb{N}\mathrm{Rat}\,B^*$ *is a sub-semiring of a skew field and so is, more generally, the semiring* $\mathbb{N}\mathrm{Rat}\,(A_2^*\times\cdots\times A_k^*)$.

This last result is itself the consequence of a theorem due independently to B. H. Neumann and to A. I. Malcev, whose proof will occupy Section 7. Before stating it in full generality, we can easily present a classic special case. A *Laurent series* is a series in one variable x of the form

cf. p. 616

$$s = \sum_{k=z}^{k=+\infty} c_k\, x^k \ , \qquad \text{for some } z \text{ in } \mathbb{Z},$$

where the coefficients c_k are, for example, in $\mathbb{Q}$. That is, s is an element of $\mathbb{Q}\langle\langle\mathbb{Z}\rangle\rangle$ whose support is contained in an *initial section of* $\mathbb{Z}$. It is not hard to verify that the Laurent series form a ring (in fact, a field) and that closure under multiplication is ensured first, by the fact that $\mathbb{Z}$ is *totally ordered*, and second by the fact that we only consider series whose support has a minimum element, which we write $\mathbb{Q}_{\mathsf{wo}}\langle\langle\mathbb{Z}\rangle\rangle$. It is these two points that we shall successively generalise. *cf. Exer. III.1.3*

Definition 4.1 A total order relation $\leqslant$ on a group G makes G an *ordered group* if it is compatible with left and right multiplication in G; that is, if

$$\forall a, b, c \in G \qquad a \leqslant b \quad \Longrightarrow \quad a\,c \leqslant b\,c \quad \text{and} \quad c\,a \leqslant c\,b \ . \qquad \square$$

Slightly inaccurately, when we say that a group is ordered we will mean that there exists an order which makes it an ordered group.

Theorem 4.6 [**Birkhoff–Tarski–Neumann–Iwazawa**[22]] *A finite direct product of free groups is ordered.*

Recall that a *well ordered* set is a totally ordered set such that every non-empty subset has a smallest element. *cf. Th. II.5.1, p. 294*

Definition 4.2 [**Hahn–Malcev–Neumann**] Let $\mathbb{K}$ be a semiring and G an ordered group. We write $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ to denote the set of series on G with coefficients in $\mathbb{K}$ whose support is a well ordered subset of G. $\square$

Theorem 4.7 [**Malcev–Neumann**] *If $\mathbb{K}$ is a skew field and G an ordered group, then $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ is a skew field.*

Theorems 4.6 and 4.7 are proved in Section 7. They imply that $\mathbb{K}_{\mathsf{wo}}\langle\langle F(B)\rangle\rangle$ (resp. $\mathbb{K}_{\mathsf{wo}}\langle\langle F(A_2)\times\cdots\times F(A_k)\rangle\rangle$) is a skew field (here $F(B)$ is the free group generated by B). To deduce Theorem 4.5, we must also ensure that $\mathbb{K}\mathrm{Rat}\,B^*$ – in fact $\mathbb{K}\langle\langle B^*\rangle\rangle$ – is included in $\mathbb{K}_{\mathsf{wo}}\langle\langle F(B)\rangle\rangle$, respectively that

$$\mathbb{K}\langle\langle A_2^*\times\cdots\times A_k^*\rangle\rangle \subseteq \mathbb{K}_{\mathsf{wo}}\langle\langle F(A_2)\times\cdots\times F(A_k)\rangle\rangle \ ;$$

that is, to be more precise, that we can order $F(A_2)\times\cdots\times F(A_k)$ in such a way that the above inclusion is true. We can also prove Theorem 4.4 from Theorems 4.6 and 4.7, directly, without using Theorem 4.5 (and hence without using realisation by representation) by means of another construction on transducers.[23]

A *strictly graded* monoid is a graded monoid such that every element of the minimum generating set D of M is of length 1. The definition implies that if $d_1\,d_2\cdots d_k = d'_1\,d'_2\cdots d'_l$ with the d_i and the d'_j in D, then $k = l$. A free monoid and a finite direct product of free monoids are both strictly graded monoids.

[22] And possibly others.

[23] This is indeed Harju and Karhumäki's original proof [111].

Let $\{x\}$ be a one letter alphabet. We define a map from $\mathbb{K}\langle\langle M\rangle\rangle$ to $[\mathbb{K}\langle M\rangle]\langle\langle x^*\rangle\rangle$ by grouping the terms of a series into homogeneous polynomials: if s is in $\mathbb{K}\langle\langle M\rangle\rangle$, then the series $s^{(\times)}$ of $[\mathbb{K}\langle M\rangle]\langle\langle x^*\rangle\rangle$ is defined by

$$<s^{(\times)}, x^l> = \sum_{m\in M,\ \varphi(m)=l} <s, m> m \ .$$

We immediately verify:

Proposition 4.7 *The map $s \mapsto s^{(\times)}$ is an injective morphism (of $\mathbb{K}$-semialgebras) from $\mathbb{K}\langle\langle M\rangle\rangle$ to $[\mathbb{K}\langle M\rangle]\langle\langle x^*\rangle\rangle$.* ■

From which we deduce that, for all s and t in $\mathbb{K}\langle\langle M\rangle\rangle$,

$$s = t \iff s^{(\times)} = t^{(\times)} \ . \tag{4.3}$$

A $\mathbb{K}$-rational subset of M is accepted by a $\mathbb{K}$-automaton $\mathcal{A}$ over M, which we can suppose to be in normal form; that is, one whose transitions are labelled by $\mathbb{K}$-linear combinations of elements of D. If $\mathcal{A} = \langle I, E, T\rangle$ is such an automaton of dimension Q, we write $\mathcal{A}^{(\times)} = \langle I, E^{(\times)}, T\rangle$ for the $\mathbb{K}\langle M\rangle$-automaton whose transitions are labelled by the letter x with a coefficient equal to the label of the corresponding transition in $\mathcal{A}$:

$$\forall p, q \in Q \qquad \left(E^{(\times)}\right)_{p,q} = (E_{p,q})\ x \ .$$

We have:

Proposition 4.8 $|\mathcal{A}^{(\times)}| = |\mathcal{A}|^{(\times)}$.

Proof. Since $\mathcal{A}$ is in normal form, a computation in $\mathcal{A}$ labelled m is of length $|m|$, and hence the matrix of labels of computations of $\mathcal{A}$ of length k is E^k: it is the matrix of coefficients of computations labelled x^k in $\mathcal{A}^{(\times)}$. ■

Proof of Theorem 4.4. Let $\mathbb{K}$ be a skew field. The equivalence of $\mathbb{K}$-automata over $A_1^*\times A_2^*\times\cdots\times A_k^*$ reduces, by (4.3) and Proposition 4.8, to that of $\mathbb{K}\langle A_1^*\times A_2^*\times\cdots\times A_k^*\rangle$-automata over x^*. Since $\mathbb{K}\langle A_1^*\times A_2^*\times\cdots\times A_k^*\rangle$ is a sub-semiring of the skew field $\mathbb{K}_{\mathsf{wo}}\langle\langle F(A_1)\times\cdots\times F(A_k)\rangle\rangle$, this is decidable by Corollary III.4.18, quoted above. ■

Theorem 4.4 is not only a good example of the algebraic nature of automata theory; it is also a remarkable illustration of 'the pertinence of enumeration'; that is, of the improvement in understanding of the automaton model made by taking *weights* of computations into account and not just their success. The problem initially posed was that of the equivalence of k-tape automata (in the Rabin–Scott model). The undecidability of this equivalence, in the general case, followed easily, as we saw above, from Post's Theorem. However, this does not settle the question for the equivalence of *deterministic* k-tape automata. Some partial and complex solutions were proposed but the problem had remained unsolved for thirty years. Without anticipating the next section too much (where we shall study these deterministic k-tape automata) let us hint that in such an automaton, each accepted k-tuple is accepted with weight 1 and hence two such automata are equivalent if and only if they are equivalent *as weighted automata*, and the result that was sought for such a long time is a direct consequence of Theorem 4.4.

see the bibliographic notes

cf. Cor. 5.19, p. 589

DEEPER & BROADER

5 Deterministic rational relations

The original model of finite k-tape automata used by Rabin and Scott had two characteristics that we have so far *omitted* from our transcription into the labelled graph model. The first is that the words written on each of the k tapes end in a special symbol, usually called the *endmarker* and always written \$. The second is that the k-tape automata are *deterministic*. *cf. Sec. I.7, p. 171*

There is no point defining an endmarker in the labelled graph model in general because it does not change the power of the model; that is, the sets of words accepted, or the series realised, are the same. In the case of deterministic automata on the other hand, the endmarker becomes useful, and hence necessary.

We start by describing the behaviour of an automaton with an endmarker, then a deterministic automaton over $A^* \times B^*$. We shall then define *deterministic relations* and describe their properties.

5.1 Transducers with an endmarker

Let \$ be a symbol which, by convention, does not belong to any alphabet, which we shall call an *endmarker*. For each alphabet A, we write $A_\$$ for the set $A \cup \{\$\}$.

Definition 5.1 A *transducer with an endmarker over* $A^* \times B^*$, is a transducer $\mathcal{T}$ over $A_\$^* \times B_\$^*$. The *pure behaviour* of a transducer $\mathcal{T}$ over $A^* \times B^*$ with an endmarker is the set of pairs in $A^* \times B^*$ which, when each component is multiplied by \$, are accepted by $\mathcal{T}$; we write it as

$$|\mathcal{T}|_\$ = \{(u, v) \in A^* \times B^* \mid (u\$, v\$) \in |\mathcal{T}|\} \ . \qquad \square$$

Example 5.1 Let $A = \{a, b\}$ and $\mathcal{T}_1$ be the transducer over $A^* \times A^*$ with an endmarker shown in Figure 5.1. We verify that its pure behaviour is the complement of the diagonal of $A^* \times A^*$, i.e. $|\mathcal{T}_1|_\$ = \{(f, g) \in A^* \times A^* \mid f \neq g\}$. $\square$

The following proposition states that the use of an endmarker and the definition of the pure behaviour changes nothing about the relations that we can define on $A^* \times B^*$.

Proposition 5.1 *A subset P of $A^* \times B^*$ is rational if and only if it is the pure behaviour of some transducer with an endmarker over $A^* \times B^*$.*

Proof. Since $|\mathcal{T}|_\$ = (|\mathcal{T}| \cap (A^*\$ \times B^*\$))\,\pi$, where π is the projection of $A_\$^* \times B_\$^*$ onto $A^* \times B^*$, we know $|\mathcal{T}|_\$$ is rational: it is the image under a morphism of the intersection of a rational set with a recognisable set. *cf. Propty 1.8, p. 540*

Conversely, every rational relation is the pure behaviour of a transducer with an endmarker since, for all transducers $\mathcal{T}$, $|\mathcal{T}|(\$,\$)$ is a rational set. ■

These two notions of endmarker and pure behaviour will find their purpose when we use them on the deterministic transducers that we shall now define.

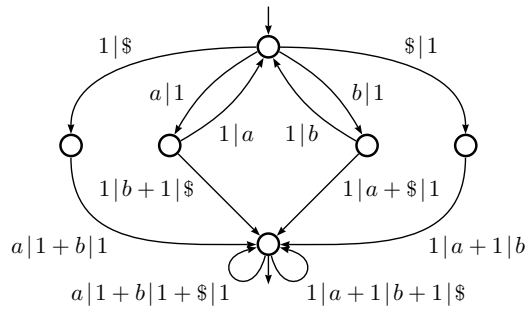


Figure 5.1: $\mathcal{T}_1$, transducer with an endmarker

5.2 Deterministic transducers

A transducer is *deterministic* if each state determines unambiguously from which tape the automaton reads a letter and if that letter defines a unique transition. We deduce that in such a transducer each label of a successful computation is accepted by a *unique* successful computation.

5.2.1 Definition

Definition 5.2 A transducer *in normal form* $\mathcal{T} = \langle Q, A, B, E, I, T \rangle$ is *deterministic* if the following conditions are satisfied:

(i) there exists a partition $Q = Q_A \cup Q_B$ such that the label of a transition whose source is in Q_A (resp. in Q_B) is in $(A \times \{1_{B^*}\})$ (resp. in $(\{1_{A^*}\} \times B)$);

(ii) for all p in Q and every label x in $(A \times \{1_{B^*}\}) \cup (\{1_{A^*}\} \times B)$, there exists at most one q in Q such that (p, x, q) is in E;

(iii) there exists a unique initial state (written i from now on); that is, $\|I\| = 1$. □

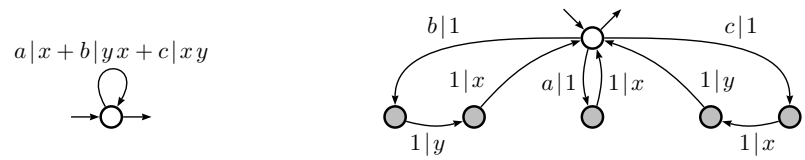
The definition is applied to transducers *in normal form* so that condition (i) makes sense. When we want to be precise, we shall write a deterministic transducer $\mathcal{T} = \langle Q_A, Q_B, A, B, E, i, T \rangle$. The transducer $\mathcal{T}_1$ of Example 5.1 is in normal form but is not deterministic; it does not satisfy condition (i) since, for example, the final state is the source of transitions labelled in $A_\$ \times 1_{A^*}$ *and* in $1_{A^*} \times A_\$$. Later we shall give a deterministic transducer with an endmarker that recognises the complement of the identity.

cf. Fig. 5.1
cf. Fig. 5.7, p. 586

cf. p. 544
Example 5.2 Let us return to the morphism $\varphi_1 \colon \{a, b, c\}^* \to \{x, y\}^*$ of Example 1.12: $a\varphi_1 = x$, $b\varphi_1 = yx$, and $c\varphi_1 = xy$. A trivial transducer, with a single state, recognises φ_1; putting this transducer into normal form; i.e., replacing each

transition by a sequence of transitions each labelled first with an element of $(A\times\{1_{B^*}\})$ and then by elements of $(\{1_{A^*}\}\times B)$, gives a deterministic transducer, as we can see in Figure 5.2. □

cf. Exer. 1.8, p. 531



(a) by a transducer with one state (b) by a deterministic transducer

Figure 5.2: Automata realisations of the morphism φ_1

This construction performed for φ_1 generalises easily, and we have:

Property 5.1 *Every morphism between free monoids (and hence every inverse morphism) is realised by a deterministic transducer.* ■

It is important to note that the conditions of Definition 1.13 apply to the *outgoing* transitions from the states and the *initial states* of the transducers. It thus follows, *ipso facto*, that:

Proposition 5.2 *A covering (or, more generally, an immersion) of a deterministic transducer is a deterministic transducer.* ■

Corollary 5.3 *We can always assume that the behaviour of a deterministic transducer with an endmarker over $A^*\times B^*$ is included in $A^*\$\times B^*\$$ without changing its pure behaviour.*

Proof. We have seen that the pure behaviour of a transducer $\mathcal{T}$ is given by

$$|\mathcal{T}|_{\$} = (|\mathcal{T}\,| \cap (A^*\$\times B^*\$))\,\pi \quad .$$

There thus exists an immersion in $\mathcal{T}$, and hence a transducer, deterministic if $\mathcal{T}$ is, whose behaviour is $|\mathcal{T}\,| \cap (A^*\$\times B^*\$)$ since $A^*\$\times B^*\$$ is a *recognisable* subset of $A^*_{\$}\times B^*_{\$}$. ■

cf. Prop. II.3.10, p. 268

Remark 5.1 The transducer with an endmarker whose behaviour is

$$|\mathcal{T}\,| \cap (A^*\$\times B^*\$)$$

is obtained by forming the product of $\mathcal{T}$ with (the automaton of) the *action* δ_3 of $A^*_{\$}\times B^*_{\$}$ which recognises the subset $A^*\$\times B^*\$$. We show this last again, trimmed, in Figure 5.3. □

cf. Exer. II.2.30, p. 255 and p. 356

It must be noted that we have here an excellent application of the notion of covering: it makes the explicit construction of the transducer mentioned in Corollary 5.3 unnecessary.

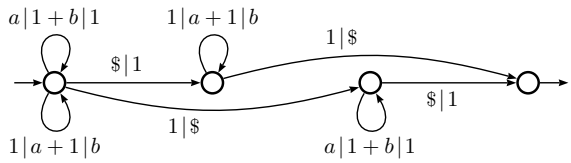


Figure 5.3: The automaton of the action which recognises $A^*\$ \times B^*\$$

5.2.2 Uniqueness of computations

The properties of the computations of a deterministic transducer $\mathcal{T}$ are almost the same as those of a deterministic automaton (over a free monoid), even if $\mathcal{T}$ is over a *non-free* monoid. The precise description of what is similar and what is different between these two types of computation requires a supplementary definition. Let $\mathcal{A}$ be an automaton over a monoid M and c a computation of $\mathcal{A}$:

$$c \ := \ p_0 \xrightarrow[\mathcal{A}]{m_1} p_1 \xrightarrow[\mathcal{A}]{m_2} p_2 \cdots \xrightarrow[\mathcal{A}]{m_n} p_n \ .$$

A *prefix* (resp. *the* prefix of length k) of c is a (resp. *the*) computation d:

$$d \ := \ p_0 \xrightarrow[\mathcal{A}]{m_1} p_1 \xrightarrow[\mathcal{A}]{m_2} p_2 \cdots \xrightarrow[\mathcal{A}]{m_k} p_k \ ,$$

with k less than or equal to n.

By the definition of a free monoid A^*, every word f of A^* has a unique factorisation over A: $f = a_1 a_2 \cdots a_n$. Let $\mathcal{A}$ be a deterministic automaton over A^*. A computation $c \ := \ p \xrightarrow[\mathcal{A}]{f} q$, with source p and label f, *if it exists*, is unique and entirely determined by p and f. The determinism of $\mathcal{A}$ also implies that if g is *the* prefix of f of length k ($k \leqslant n$), then the computation with source p and label g is the prefix of length k of the computation c. Finally, we saw that every word f of A^* defines a (partial) function from Q to itself, and that this representation is an *action* of A^* on Q, which we write $p \cdot f = q$.

This situation generalises in a way to deterministic transducers with the stronger property that the factorisation of the label of a computation does not exist *a priori* but is determined by the transducer itself. Let us start by stating that the computations are uniquely determined *once the factorisations are given*, which is a simple extension of the case of computations in an automaton (over a free monoid), with the same obvious proof, by induction on the length of computations.

Lemma 5.4 *Let $\mathcal{T}$ be a deterministic transducer on $A^* \times B^*$ and let*

$$c \ := \ p = p_0 \xrightarrow[\mathcal{T}]{x_1} p_1 \xrightarrow[\mathcal{T}]{x_2} p_2 \cdots \xrightarrow[\mathcal{T}]{x_{n-1}} p_{n-1} \xrightarrow[\mathcal{T}]{x_n} p_n \qquad \textit{and}$$

$$c' \ := \ p = p_0 \xrightarrow[\mathcal{T}]{x_1} p'_1 \xrightarrow[\mathcal{T}]{x_2} p'_2 \cdots \xrightarrow[\mathcal{T}]{x_{n-1}} p'_{n-1} \xrightarrow[\mathcal{T}]{x_n} p'_n$$

be two computations with the same source p, the same label (f, g) and that determine the same factorisation (f, g) of

$$(f, g) = x_1 x_2 \cdots x_n \ .$$

Then $c = c'$; that is, for each i, $1 \leqslant i \leqslant n$, we have $p_i = p'_i$. ■

It is consistent, and useful for proofs by induction, to say that in each state p of a transducer $\mathcal{T}$ there exists a computation of length 0 with source p labelled $(1_{A^*}, 1_{B^*})$. The fact that the transducer $\mathcal{T}$ determines a unique factorisation of each element of $A^* \times B^*$ is expressed in the following lemma, which we could call the 'Corrugated Cardboard Lemma'.

Lemma 5.5 *Let $\mathcal{T}$ be a deterministic transducer on $A^* \times B^*$. Let p be a state of $\mathcal{T}$ and let (f, g) be in $A^* \times B^*$. The set of prefixes (u, v) of (f, g) such that there exists a computation with source p labelled (u, v) is a* chain *for the prefix order, maximal between $(1_{A^*}, 1_{B^*})$ and its largest element.*

We first prove a lemma on the factorisations of elements of $A^* \times B^*$. The monoid $A^* \times B^*$ is not free, but in a way it is not far from being so; it is this that we want to express formally. The set $X = (A \times \{1_{B^*}\}) \cup (\{1_{A^*}\} \times B)$ is the minimal generating set of $A^* \times B^*$; that is, each element (u, v) of $A^* \times B^*$ factorises into a product of elements of X: $(u, v) = x_1 x_2 \cdots x_n$, $x_i \in X$. Obviously, this factorisation is not unique; for example, $(a, b) = (a, 1)(1, b) = (1, b)(a, 1)$. The *length* of (u, v) is the quantity $|u| + |v|$, written $|(u, v)|$. This is the common length of all the factorisations of (u, v) over X. The terminology is appropriate since $|(u, v)(u', v')| = |(u, v)| + |(u', v')|$.

An element (u, v) of $A^* \times B^*$ is a *prefix* of (f, g) if there exists (h, k) such that $(f, g) = (u, v)(h, k)$. In this case, u is a prefix of f, $f = u\,h$, and v is a prefix of g, $g = v\,k$. The relation 'prefix of' is a (partial) order on $A^* \times B^*$ called the *prefix order*. The set of all the prefixes of an arbitrary element (f, g) is a lattice for the prefix order. A maximal chain of prefixes of (f, g) uniquely determines a factorisation of (f, g) over X and *vice versa*.

Lemma 5.6 *Let (f, g) be in $A^* \times B^*$ and let (u, v) and (u', v') be two arbitrary prefixes of (f, g) such that neither of them is a prefix of the other. Let*

$$(u, v) = x_1 x_2 \cdots x_n \qquad \textit{and} \qquad (u', v') = y_1 y_2 \cdots y_m \ ,$$

be two distinct *factorisations of these elements (a necessary hypothesis if $(u, v) = (u', v')$) and let*

$$x_1 x_2 \cdots x_{i-1} = y_1 y_2 \cdots y_{i-1}$$

be the longest common prefix of these two factorisations (as words of X^*; that is, $x_j = y_j$ for all j such that $0 \leqslant j \leqslant i-1$ and $x_i \neq y_i$). Then x_i belongs to $(A \times \{1_{B^*}\})$ and y_i belongs to $(\{1_{A^*}\} \times B)$, or vice versa.

Proof. Let $(h, k) = x_1 x_2 \cdots x_{i-1}$. Suppose that x_i and y_i both belong to $(A \times \{1_{B^*}\})$: $x_i = (a', 1)$ and $y_i = (a'', 1)$. Then

$$(h, k)\, x_i = (h\, a', k) \qquad \text{and} \qquad (h, k)\, y_i = (h\, a'', k) \ .$$

The words $h\,a'$ and $h\,a''$ are both *the* prefix of length $|h| + 1$ of f and hence $a' = a''$, which contradicts $x_i \neq y_i$. We use identical reasoning if x_i and y_i are both in $(\{1_{A^*}\} \times B)$. ∎

Proof of Lemma 5.5. Thanks to the convention that we have adopted, the set of prefixes of (f,g) which are the label of a computation with source p is non-empty. Let

$$\begin{aligned} c &:= \quad p = p_0 \xrightarrow[\mathcal{T}]{x_1} p_1 \xrightarrow[\mathcal{T}]{x_2} p_2 \cdots \xrightarrow[\mathcal{T}]{x_n} p_n = q \qquad \text{and} \\ d &:= \quad p = p_0 \xrightarrow[\mathcal{T}]{y_1} p'_1 \xrightarrow[\mathcal{T}]{y_2} p'_2 \cdots \xrightarrow[\mathcal{T}]{y_m} p'_m = r \end{aligned}$$

be two computations of $\mathcal{T}$ labelled (u,v) and (u',v') respectively, such that (u,v) and (u',v') are two prefixes of (f,g). Suppose that neither of the two is a prefix of the other and let

$$x_1\, x_2 \cdots x_{i-1} = y_1\, y_2 \cdots y_{i-1} \quad ,$$

be the longest common prefix of the two factorisations (still taken as words of X^*). As we noted in Lemma 5.4, the prefixes of length $i-1$ of c and d coincide:

$$\begin{aligned} c &:= \quad p = p_0 \xrightarrow[\mathcal{T}]{x_1} p_1 \xrightarrow[\mathcal{T}]{x_2} p_2 \cdots \xrightarrow[\mathcal{T}]{x_{i-1}} p_{i-1} \xrightarrow[\mathcal{T}]{x_i} p_i \cdots \xrightarrow[\mathcal{T}]{x_n} p_n = q \qquad \text{and} \\ d &:= \quad p = p_0 \xrightarrow[\mathcal{T}]{x_1} p_1 \xrightarrow[\mathcal{T}]{x_2} p_2 \cdots \xrightarrow[\mathcal{T}]{x_{i-1}} p_{i-1} \xrightarrow[\mathcal{T}]{y_i} p'_i \cdots \xrightarrow[\mathcal{T}]{y_m} p'_m = r \quad . \end{aligned}$$

We hence deduce from Lemma 5.6 that $x_i = (a,1)$ and $y_i = (1,b)$ (or *vice versa*). This is a contradiction since p_{i-1} cannot be the source of two arcs, one with a label in $(A \times \{1_{B^*}\})$, and the other with a label in $(\{1_{A^*}\} \times B)$. Thus, one of (u,v) and (u',v') must be a prefix of the other, and the prefixes of (f,g) which are labels of computations with source p form a chain.

Let (h,k) be the largest element of this chain. The computation

$$d \;:=\; p \xrightarrow[\mathcal{T}]{(h,k)} q$$

determines a factorisation of (h,k) over $(A \times \{1_{B^*}\}) \cup (\{1_{A^*}\} \times B)$, which in its turn defines a *maximal* chain between $(1_{A^*}, 1_{B^*})$ and (h,k). ■

As an immediate consequence of Lemma 5.4 and Lemma 5.5, we have:

Corollary 5.7 *Let $\mathcal{T}$ be a deterministic transducer on $A^* \times B^*$. Let p be a state of $\mathcal{T}$ and (f,g) be in $A^* \times B^*$. If it exists, a computation with source p labelled (f,g) is unique.* ■

In other words:

Corollary 5.8 *The behaviour of a deterministic transducer is an unambiguous rational subset of $A^* \times B^*$,* which can also be stated: *the behaviour of a deterministic transducer, seen as an $\mathbb{N}$-automaton over $A^* \times B^*$, is a characteristic series in $\mathbb{N}\mathrm{Rat}\, A^* \times B^*$.* ■

cf. Def. 3.2, p. 558

5.2.3 Almost an action

As in the case of deterministic automata over a free monoid, given a transducer $\mathcal{T}$ on $A^* \times B^*$, each (f,g) in $A^* \times B^*$ defines a *partial function* from Q to itself, and we write

$$p \circ (f,g) = q \qquad \text{if} \qquad c \;:=\; p \xrightarrow[\mathcal{T}]{(f,g)} q \quad .$$

But, contrary to the case of deterministic automata, this, in general, *does not define an action* from $A^* \times B^*$ to Q (and that is the reason that we do not use the notation $p \cdot (f,g)$). In fact if, with the notation of Definition 5.2, $p \circ (f,g) = q$ belongs to Q_A and $q \circ (a,1) = r$ belongs to Q_B, then

$$p \circ (f\,a, g\,b) = ((p \circ (f,g)) \circ (a,1)) \circ (1,b)$$

is defined, although $((p \circ (f,g)) \circ (1,b)$ is not and yet

$$(f\,a, g\,b) = (f,g)\,(1,b)\,(a,1) \quad .$$

A transducer $\mathcal{T} = \langle\, Q, A, B, E, I, T \,\rangle$ determines, for each pair (f,g) in $A^* \times B^*$ for which $i \circ (f,g)$ is defined, a *unique* factorisation; we shall call it *the* factorisation of (f,g), if there is no ambiguity.

Lemma 5.5 ensures, given $\mathcal{T}$, p and (f,g), the existence and uniqueness of a computation of $\mathcal{T}$, which we can view as a traversal of the words f and g which goes alternately from one to the other and which can thus naturally be represented as in Figure 5.4, hence the name suggested for this lemma.

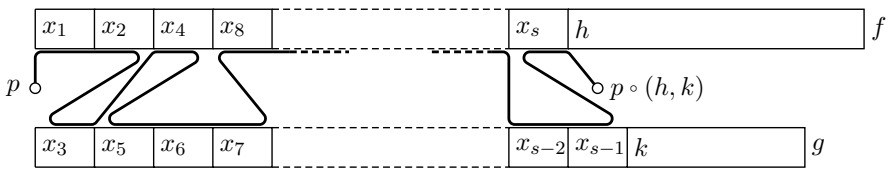


Figure 5.4: A computation of $\mathcal{T}$ which oscillates between f and g

We now summarise the consequences of this lemma in a form that is useful for further proofs, and gather them into a corollary.

Corollary 5.9 *Let $\mathcal{T}$ be a deterministic transducer on $A^* \times B^*$. Let p be a state of $\mathcal{T}$ and (u,v) and (h,k) two elements of $A^* \times B^*$.*

(i) *If $p \circ (u,v) = q$ and $q \circ (h,k) = r$, then $p \circ (u\,h, v\,k) = r$.*

(ii) *If $p \circ (u,v) = q$ and $p \circ (u\,h, v\,k) = r$, then $q \circ (h,k) = r$.*

(iii) *If $p \circ (u\,h, v)$ and $p \circ (u, v\,k)$ both exist, then $h = 1_{A^*}$ or $k = 1_{B^*}$.*

Proof. (i) is a direct consequence of the definition: $(u\,h, v\,k)$ is the label of the computation obtained by concatenating the two computations

$$p \xrightarrow[\mathcal{T}]{(u,v)} q \qquad \text{and} \qquad q \xrightarrow[\mathcal{T}]{(h,k)} r \quad .$$

(ii) Since (u,v) is the unique prefix of $(u\,h, v\,k)$ of length $|(u,v)|$ such that $p \circ (u,v)$ is defined, the computation $p \xrightarrow{(u,v)} q$ must be a prefix of the computation $p \xrightarrow{(u\,h, v\,k)} r$, which can therefore be written $p \xrightarrow{(u,v)} q \xrightarrow{(h,k)} r$.

(iii) Both $(u\,h, v)$ and $(u, v\,k)$ are prefixes of $(u\,h, v\,k)$; Lemma 5.5 implies that one must be a prefix of the other, and hence that $h = 1_{A^*}$ or $k = 1_{B^*}$. ∎

Exercises

• 5.1 Verify that $V_1 = \{(a^n b^m, c^n) \mid n, m \in \mathbb{N}\}$ and $W_1 = \{(a^n b^m, c^m) \mid n, m \in \mathbb{N}\}$ are the behaviours of deterministic transducers from $\{a, b\}^*$ to $\{c\}^*$.

5.2 Show that we can always assume that the initial state i of a deterministic transducer $\mathcal{T} = \langle Q_A, Q_B, A, B, E, i, T \rangle$ belongs to Q_A (or to Q_B), and that T is included in Q_A (or in Q_B).

5.3 Deterministic relations

The above definitions and properties about *transducers* enable us to define a new family *of rational relations* – in fact, two – and to prove some initial properties.

5.3.1 Definitions

Definition 5.3 A relation $\alpha\colon A^* \to B^*$ is *deterministic rational* if it is the pure behaviour of a deterministic transducer over $A^* \times B^*$ with an endmarker. □

Example 5.3 The universal relation, i.e. the relation whose graph is all of $A^* \times B^*$, is a deterministic rational relation (*cf.* Figure 5.5). This example shows better than any that the uniqueness of computations in a deterministic transducer does not imply that its behaviour is a function. □

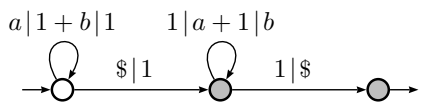


Figure 5.5: The universal relation is deterministic

By Proposition 5.1, a deterministic rational relation is a rational relation. In what follows, we shall abbreviate 'deterministic rational relation' to *deterministic relation*. We write $\mathrm{DRat}\, A^* \times B^*$ for the set of deterministic relations on $A^* \times B^*$.

There are two ingredients in the definition of a deterministic relation: the endmarker and the deterministic transducer, and it does not follow directly from Definition 5.3 that the behaviour of a deterministic transducer is a deterministic relation! However, an immediate construction which translates the forward direction of Proposition 5.1, as shown in Figure 5.6 (b), enables us to fall on our feet: the terminology is indeed consistent.

Intuitively, the endmarker \$ allows us to encode the information 'the entire word has been read', and this information is used by the transducer, which can change state when it reads the symbol. In fact, and still intuitively, it is this symbol which ensures that deterministic relations are closed under complement. To describe more formally the possibilities that this symbol provides, we distinguish the deterministic relations for which this is not really necessary.

cf. § 5.3.2 below

Definition 5.4 A relation α on $A^* \times B^*$ is *pure rational deterministic* if it is the behaviour of a deterministic transducer on $A^* \times B^*$ without an endmarker. □

As above, we shall say *pure deterministic relation* for a pure deterministic rational relation. We write $\mathrm{D'Rat}\, A^*{\times}B^*$ for the set of pure deterministic relations on $A^*{\times}B^*$. For the same reason as above, but not by definition, a pure deterministic relation is a deterministic relation. From Example 5.2 and from the generalisation which leads to Property 5.1, we deduce:

Property 5.2 *Every morphism (resp. every inverse morphism) from one free monoid to another is a pure deterministic relation.* ■

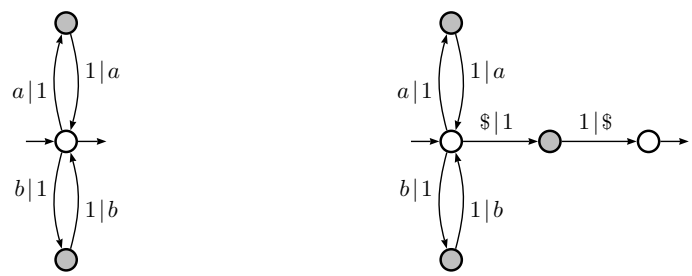
Corollary 5.9 (iii) becomes a necessary condition which pure relations must satisfy.

Lemma 5.10 *If θ is a pure deterministic relation on $A^* \times B^*$, we have, for all u and h in A^* and all v and k in B^*,*

$$(u\,h, v) \in \theta \quad \text{and} \quad (u, v\,k) \in \theta \quad \Longrightarrow \quad h = 1_{B^*} \quad \text{or} \quad k = 1_{B^*} \ . \qquad ■$$

The universal relation is a deterministic relation which does not satisfy this lemma and hence:

Proposition 5.11 $\mathrm{D'Rat}\, A^* \times B^*$ *is strictly contained in* $\mathrm{DRat}\, A^* \times B^*$. ■



(a) deterministic transducer (b) $\mathcal{I}_1$, deterministic transducer with endmarker

Figure 5.6: The identity is a pure deterministic relation.

5.3.2 Complement

A remarkable property of deterministic relations is

Proposition 5.12 $\mathrm{DRat}\, A^* \times B^*$ *is closed under complement.*

Proof. Let $\mathcal{T}$ be a deterministic transducer with an endmarker which recognises a deterministic relation τ. We 'complete' $\mathcal{T}$ by adding, if necessary, *two sink states*, s_A in Q_A, and s_B in Q_B, such that for all a in $A_\$$ and all p in Q_A, $(a, 1_{B^*})$ is the label of a transition with source p, and for all b in $B_\$$ and all p' in Q_B, $(1_{A^*}, b)$ is the label of a transition with source p'. The transitions which we thus add to $\mathcal{T}$ have destination s_A if their label is of the form $(a, 1_{B^*})$ or $(1_{A^*}, \$)$, or s_B if their label is of the form $(1_{A^*}, b)$ or $(\$, 1_{B^*})$. The effect of this construction on $\mathcal{I}_1$, the transducer with an endmarker in Figure 5.6, is shown in Figure 5.7 (a). Every element of $A^*_\$ {\times} B^*_\$$ is the label of a unique computation with source i in $\mathcal{T}$, a complete transducer with endmarker.

cf. Rem. 5.1, p. 579 We next form the product of $\mathcal{T}$ with (the trim automaton of) the action δ_3, which recognises the subset $A^*\$\times B^*\$$ of $A^*_\$\times B^*_\$$. We obtain $\mathcal{T}'$, a deterministic transducer with endmarker, whose behaviour is equal to $|\mathcal{T}| \cap (A^*\,\$\times B^*\,\$)$, hence whose pure behaviour is equal to that of $\mathcal{T}$, but also such that every element of $A^*\$\times B^*\$$ is the label of a unique computation with source i in $\mathcal{T}'$.

Let S be the set of states which are the destination of these computations and T' the final states of $\mathcal{T}'$ (obviously $T' \subseteq S$). The transducer $\mathcal{T}''$ obtained by taking the states $S \setminus T'$ as the terminal states has a pure behaviour which is the complement in $A^*\times B^*$ of the pure behaviour of $\mathcal{T}'$, hence of $\mathcal{T}$. ■

cf. Rem. 1.1, p. 526 Since the rational relations are not closed under complement:

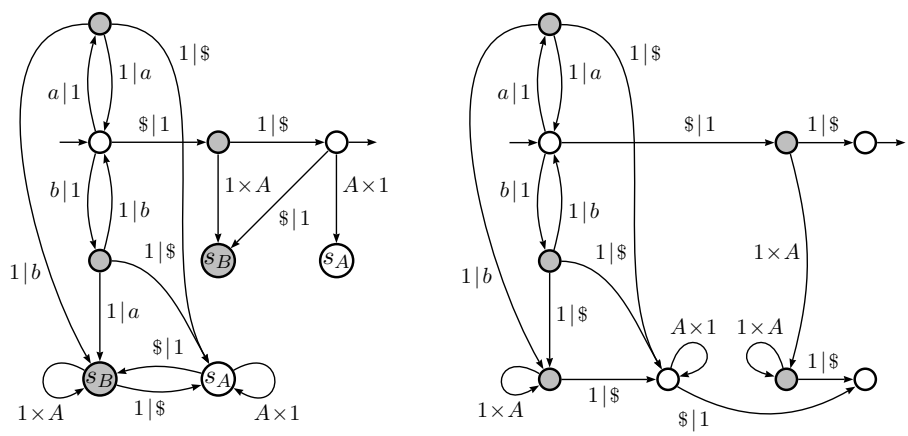
Corollary 5.13 $\mathrm{DRat}\, A^*\times B^*$ *is strictly contained in* $\mathrm{Rat}\, A^*\times B^*$. ■

Furthermore, in the example that we chose to show that the rational relations are not closed under intersection (hence under complement) the two relations under discussion are (pure) deterministic, hence the property:

cf. Exer. 5.1

Property 5.3 *The intersection of two elements of* $\mathrm{DRat}\, A^*\times B^*$ *is not necessarily in* $\mathrm{Rat}\, A^*\times B^*$. ■

Corollary 5.14 $\mathrm{DRat}\, A^*\times B^*$ *is not closed under union.* ■



(a) completion of the transducer $\mathcal{I}_1$ (b) product of $\mathcal{I}_1$ and δ_3

Figure 5.7: The complement of the identity is a deterministic relation

5.3.3 Iteration Lemma

cf. Cor. 5.13 We have just observed that $\mathrm{DRat}\, A^*\times B^*$ is a proper subset of $\mathrm{Rat}\, A^*\times B^*$, but to show that a particular rational relation is *non-deterministic*, we need a proposition that describes a property necessarily satisfied by the behaviour of a deterministic transducer. It will not surprise us that this proposition will take the form of a *Star Lemma*.

cf. p. 70

Lemma 5.15 *For every deterministic relation θ on $A^* \times B^*$, there exists a positive integer N which has the following property. Each pair (u, v) in $\widehat{\theta}$, where u and v are* both *of length greater than N, has a factorisation*

$$(u, v) = (u_1, v_1)\,(f, g)\,(u_2, v_2) \quad ,$$

which satisfies the conditions:

(i) $|(f, g)| \geqslant 0$ *;*

(ii) $|(u_1 f, v_1 g)| < N$ *;*

(iii) *for all h and k such that $(u_1 f h, v_1 g k)$ is in $\widehat{\theta}$, $(u_1 f^n h, v_1 g^n k)$ is in $\widehat{\theta}$ for all integers n.*

If θ is pure deterministic*, the assumption that $|u| > N$ and $|v| > N$ can be replaced by $|(u, v)| > N$, with the same conclusions.*

Proof. Let $\mathcal{T} = \langle\, Q, A_\$, B_\$, E, i, T \,\rangle$ be a deterministic transducer which realises θ and let $N = \|Q\|$. Let $(u, v) \in \widehat{\theta}$ such that $|u| > N$ and $|v| > N$; then

$$i \circ (u\,\$, v\,\$) = t \in T \quad .$$

Let (u', v') be *the* prefix of length N of *the* factorisation of $(u\,\$, v\,\$)$ determined by $\mathcal{T}$. Since the length of u and of v is greater than N, the final \$ belongs neither to u' nor to v'. The computation $i \circ (u', v')$ can thus be written

$$i \circ (u', v') \;:=\; i \xrightarrow{c_1} q_1 \xrightarrow{c_2} q_2 \cdots \xrightarrow{c_N} q_N \quad , \tag{5.1}$$

where all the c_i are in $(A \times \{1_{B^*}\}) \cup (\{1_{A^*}\} \times B)$. The proof then follows that of the classic iteration lemmas. Since $N = \|Q\|$, there exist two *distinct* indices i and j such that $q_i = q_j = q$. Write

$$(u_1, v_1) = c_1 \cdots c_i \qquad \text{and} \qquad (f, g) = c_{i+1} \cdots c_j \quad .$$

It then follows that $i \circ (u_1, v_1) = q$ and $q \circ (f, g) = q$. Thus, for all (h, k) such that $q \circ (h, k) = t' \in T$, we have

$$\forall n \in \mathbb{N} \qquad i \circ [(u_1, v_1)\,(f, g)^n\,(h, k)] = t' \quad ,$$

and the three conditions are satisfied.

If θ is pure deterministic, it is recognised by a deterministic transducer $\mathcal{T} = \langle\, Q, A, B, E, I, T \,\rangle$ and the computation (5.1) exists as soon as $|(u, v)| > N$. ■

Example 5.4 The classic example of a non-deterministic (but nonetheless unambiguous) rational relation is the function $\theta_1 \colon \{x\}^* \to \{x\}^*$ defined by

$$x^n \theta_1 = \begin{cases} x^n & \text{if} \quad n \text{ odd} \\ x^{n/2} & \text{if} \quad n \text{ even} \end{cases}$$

for all n in $\mathbb{N}$. The relation θ_1 is rational since

$$\widehat{\theta_1} = (x, x)\,(x^2, x^2)^* \cup (x^2, x)^* \quad ,$$

but it is not deterministic. Suppose in fact that it is, and consider a factorisation of the pair (x^{2N}, x^N) given, along with the integer N, by Lemma 5.15:

$$(x^{2N}, x^N) = (u_1, v_1)\,(f, g)\,(u_2, v_2) \quad .$$

The word f must be of even length, and g of half that length, as otherwise the pair $(u_1 f^2 u_2, v_1 g^2 v_2)$ would not be in $\widehat{\theta_1}$. The pair $(u_1 f u_2 x, v_1 g v_2 x^{N+1})$ is also in $\widehat{\theta_1}$ but not $(u_1 f^2 u_2 x, v_1 g^2 v_2 x^{N+1})$. Contradiction. □

Exercises

5.3 Show that the relation whose graph is defined by

$$\widehat{\theta_2} = \{(a^n b, a^n) \mid n > 0\} \cup \{(a^n, 1) \mid n \in \mathbb{N}\}$$

is a deterministic relation which satisfies Lemma 5.10, but is not pure deterministic.

5.4 Show that the relations $\theta_3, \theta_4 \colon \{x\}^* \to \{x\}^*$ defined by

$$x^n\theta_3 = \{x^m \mid n \leqslant m \leqslant 2n\} \qquad \text{and} \qquad x^n\theta_4 = \{x^n, x^{2n}\}$$

are rational but not deterministic.

• 5.5 Verify that the relations $\sigma_3, \sigma_4 \colon \{a, b\}^* \to \{a, b\}^*$ defined by

$$a^n\sigma_3 = \{a^n, b^{2n}\} \ , \qquad b^n\sigma_3 = \{a^n, b^n\} \qquad \text{and} \qquad a^{2p}\sigma_4 = a^{4p} \ , \qquad a^{2p+1}\sigma_4 = b^{2p+1} \ ,$$

are pure deterministic relations.

• 5.6 Let $\sigma \colon A^* \to B^*$ be a rational substitution. Give a sufficient condition for σ to be a deterministic relation.

• 5.7 Let K be a rational of A^*; show that $\iota_K \colon A^* \to A^*$, the intersection with K, is a pure deterministic relation.

5.4 Geography of $\mathrm{Rat}\, A^* \times B^*$ I

We have deduced fairly directly from the definitions the sequence of strict inclusions

$$\mathrm{D'Rat}\, A^* \times B^* \subsetneq \mathrm{DRat}\, A^* \times B^* \subsetneq \mathrm{Rat}\, A^* \times B^* \quad .$$

It remains to place the family $\mathrm{Rec}\, A^* \times B^*$ in this schema and to examine the corresponding problems of decidability. From Propositions 5.2 and II.3.10, we deduce the following:

cf. p. 579 and p. 268

Proposition 5.16 *The intersection of a deterministic relation and a recognisable relation is a deterministic relation.* ■

From this we deduce, since the *universal relation* is a deterministic relation:

Corollary 5.17 *Every recognisable relation from one free monoid to another is a deterministic relation.* ■

The explicit proof of the corollary, like that of Proposition 5.16, would be cumbersome without the notion of covering.

The families $\mathrm{Rec}\, A^* \times B^*$ and $\mathrm{D'Rat}\, A^* \times B^*$ are *incomparable* since identity is a pure deterministic relation and the universal relation is not.

Let us turn now to the result promised at the end of Section 4. From Corollary 5.8 we deduce:

Proposition 5.18 *If τ is a deterministic relation from A^* to B^*, then $\underline{\tau}$ is an $\mathbb{N}$-rational relation.* ■

Corollary 5.19 *The equivalence of deterministic transducers is decidable.*

Proof. By definition, two deterministic transducers $\mathcal{T}$ and $\mathcal{T}'$ are equivalent if $|\mathcal{T}| = |\mathcal{T}'|$ or $\underline{|\mathcal{T}|} = \underline{|\mathcal{T}'|}$; that is, by Corollary 5.8, if, and only if, $\mathcal{T}$ and $\mathcal{T}'$ are equivalent *as* $\mathbb{N}$-*automata over* $A^* \times B^*$, which is decidable by Theorem 4.4. ■

Note that this proposition is optimal since we have, *a contrario*:

Proposition 5.20 *The inclusion of deterministic relations is undecidable.*

Proof. Take two morphisms $\theta, \mu \colon A^* \to B^*$ which describe a Post Correspondence Problem. We have $U_\theta \subseteq \complement U_\mu$ if and only if $U_\theta \cap U_\mu = \emptyset$ and U_θ and $\complement U_\mu$ are both the graphs of deterministic relations. ■ *cf. p. 569*

Now we turn to some relative decidability problems.

Theorem 5.1 *If A is an alphabet with at least two letters, then* $\mathrm{DRat}\, A^* \times B^*$ *is undecidable in* $\mathrm{Rat}\, A^* \times B^*$ *.*

Proof. We use the notation and technique of Lisovik's Lemma to establish the statement by a standard procedure: with a Post Correspondence Problem we build a relation that is deterministic if, and only if, it is empty. We start with two morphisms $\theta \colon X^* \to Y^*$ and $\mu \colon X^* \to Y^*$, set $Z = X \cup Y$, and construct the subsets W_θ and W_μ of $Z^* \times c^*$ as in the proof of Theorem 4.3. If $\complement W_\theta \cup \complement W_\mu$ is deterministic, so is its complement *cf. Lem. 4.4, p. 571* *cf. p. 573*

$$W_\theta \cap W_\mu = \{(w\,(w\theta), c^{|w\theta|}) \mid w \in X^*,\ w\theta = w\mu\} \ .$$

If there exists a word w such that $w\theta = w\mu$ – that is, if $V_\theta \cap V_\mu$ is non-empty – then, for all n, $w^n \in V_\theta \cap V_\mu$. We choose n such that $|w^n| \geqslant N$, where N is the integer given by Lemma 5.15, and by the assumption that $W_\theta \cap W_\mu$ is deterministic. The factorisation

$$(w^n\,(w^n\theta), c^{n|w\theta|}) = (u_1, v_1)\,(f, g)\,(u_2, v_2)$$

given by the lemma is such that $u_1 f \in X^+$ and hence $u_1 f^2 u_2$ is not in the domain of $W_\theta \cap W_\mu$, which contradicts the lemma that ensures that $(u_1 f^2 u_2, v_1 g^2 v_2)$ is in $W_\theta \cap W_\mu$ if it is a deterministic relation.

Thus, $\complement W_\theta \cup \complement W_\mu$ is deterministic if and only if $V_\theta \cap V_\mu$ is empty, which is undecidable. This proves the result for an alphabet with at least 8 letters; in order to conclude in full generality, we need an injective map from $\{a, b\}^*$ to Z^* that is not only rational but also deterministic. ■

The other problems are, to my knowledge, still open. I lean towards a negative answer to the first and a positive to the second.

Problem 5.1 *Is* $\mathrm{Rec}\, A^* \times B^*$ *decidable in* $\mathrm{DRat}\, A^* \times B^*$ *?*

Problem 5.2 *Is* $\mathrm{D'Rat}\, A^* \times B^*$ *decidable in* $\mathrm{Rat}\, A^* \times B^*$ *?*

Exercises

5.8 Prove without using Proposition 5.16 that a recognisable relation is deterministic.

• 5.9 Give a sufficient condition for a recognisable relation to be a pure deterministic relation.

5.5 Matrix representations

Every rational $\mathbb{K}$-relation is recognised by a transducer in normal form, from which we deduce, by a computation which we have already had occasion to present several times, a *matrix representation* of the relation. As deterministic relations are defined by a property of transducers in normal forms which realise them, it is natural that they can also be defined by a property of matrix representations which realise them.

cf. Prop. III.2.14, p. 410, Sec. 1.5, p. 540 and Sec. 3.2, p. 561

5.5.1 Representation of a deterministic transducer

Recall that a *language* is *prefix* if it contains no proper prefix of any of its elements. Recall that we call a *prefix family of languages* a family of languages K_i such that:

cf. Exer. I.2.3, p. 85

(i) the K_i are pairwise disjoint;

(ii) the union of all the K_i is a prefix language.

We say a *matrix* m with coefficients in $\mathfrak{P}(A^*)$ is *prefix* if the coefficients of each row of m form a prefix family. We then easily show:

cf. Exer. I.2.3, ibid.

Property I.2.3 *The product of two prefix matrices is prefix.* ■

We shall say that a *representation* (λ, μ, ν) of A^* is *prefix* if the matrices λ, ν and $f\mu$, for all f in A^*, are prefix. Using the property, it is enough that the matrices λ, ν and $a\mu$, for all a in A, be prefix. We can then state:

Proposition 5.21 *Let* $\mathcal{T} = \langle Q_A, Q_B, A, B, E, i, T \rangle$ *be a deterministic transducer such that* T *is included in* Q_A*. We can construct* (λ, μ, ν)*, a prefix matrix representation of dimension* Q_A*, from* A^* *to* $\mathrm{Rat}\, B^*$ *which realises* $\mathcal{T}$*; that is,*

$$\forall (f, g) \in A^* \times B^* \qquad (f, g) \in |\mathcal{T}| \iff g \in \lambda \cdot f\mu \cdot \nu \ .$$

Conversely, we can associate every prefix representation (λ, μ, ν) *from* A^* *to* $\mathrm{Rat}\, B^*$ *with a deterministic transducer on* $A^* \times B^*$ *whose behaviour is realised by* (λ, μ, ν)*.*

Proof. The transition matrix E of $\mathcal{T}$, a square matrix of dimension $Q = Q_A \cup Q_B$, is written $E = X + Y$ where X and Y are matrices whose coefficients are respectively in $A \times 1_{B^*}$ and $1_{A^*} \times B$. Since $\mathcal{T}$ is deterministic, X, Y and E have a block decomposition of the form

$$X = \begin{pmatrix} X_A & X_B \\ 0 & 0 \end{pmatrix} , \qquad Y = \begin{pmatrix} 0 & 0 \\ Y_A & Y_B \end{pmatrix} \qquad \text{and hence} \qquad E = \begin{pmatrix} X_A & X_B \\ Y_A & Y_B \end{pmatrix} ,$$

where X_A (resp. Y_B) is a square matrix of dimension Q_A (resp. of dimension Q_B). Consider the $Q_B \times Q_A$ matrix

$$Z = Y_B^* Y_A \ .$$

For all q in Q_B, we have

$$g \in Z_{q,p} \qquad \text{if and only if} \qquad q \xrightarrow[\mathcal{T}]{(1_{A^*},g)} p \quad \text{or} \quad q \circ (1_{A^*}, g) = p \in Q_A \ .$$

The determinism of $\mathcal{T}$ ensures that if $p = q \circ (1_{A^*}, g)$ exists, it is unique and, by Corollary 5.9 (ii), that $q \circ (1_{A^*}, g) = p$ and $q \circ (1_{A^*}, g\,k) = p'$ implies $k = 1_{B^*}$, hence see p. 583 that the coefficients of row q of Z form a prefix family.

By writing $I = \begin{pmatrix} I_A & I_B \end{pmatrix}$ and $T = \begin{pmatrix} T_A \\ T_B \end{pmatrix}$ for the block decompositions congruent with that of E, we have $T_B = 0$ by the assumption that $T \subseteq Q_A$. A simple calculation then shows that

$$|\mathcal{T}| = I \cdot E^* \cdot T = \big(I_A + I_B Y_B^* Y_A\big)\big(X_A + X_B Y_B^* Y_A\big)^* T_A \ .$$

The unique non-zero coordinate of the boolean vector I is i and its value is 1 (which we identify with 1_{B^*}): if $i \in Q_A$, I_A is a prefix row vector and I_B is null, or if $i \in Q_B$, I_A is null and I_B is a prefix row vector; in either case,

$$\lambda = I_A + I_B Y_B^* Y_A$$

is a prefix row vector of dimension Q_A. We then set

$$X_A = \sum_{a \in A} (a, 1)\, a\xi_A \qquad \text{and} \qquad X_B = \sum_{a \in A} (a, 1)\, a\xi_B$$

where the $a\xi_A$ (resp. the $a\xi_B$) are row-monomial boolean $Q_A \times Q_A$ matrices (resp. $Q_A \times Q_B$ matrices) and the unique non-zero coefficient of each row, if it exists, is identified with 1_{B^*}. What is more, for each a, the matrix $\begin{pmatrix} a\xi_A & a\xi_B \end{pmatrix}$ is also row monomial; that is, for each p in Q_A, if row p of $a\xi_A$ is not uniformly zero, that of ξ_B is and *vice versa*. All this, of course, is just the transcription of the determinism of $\mathcal{T}$. It then follows that the matrix $a\xi_B Y_B^* Y_A$ is a prefix $Q_A \times Q_A$ matrix and also

$$a\mu = a\xi_A + a\xi_B Y_B^* Y_A \ .$$

We naturally set $\nu = T_A$ and obtain immediately that the representation (λ, μ, ν) realises $|\mathcal{T}|$.

The verification of the converse is left as an exercise. ∎

5.5.2 Representation of a deterministic relation

Deterministic relations are defined starting from deterministic transducers *with an endmarker*. It is thus necessary to define what the matrix representation of such a transducer is. Let $\mathcal{T}$ be a (deterministic) transducer on $A_\$^* \times B_\$^*$. From the above, we know that there exists a (prefix) matrix representation (λ, μ, ν) such that

$$\forall (f,g) \in A_\$^* \times B_\$^* \qquad (f,g) \in |\mathcal{T}| \quad \Longleftrightarrow \quad g \in \lambda \cdot f\mu \cdot \nu \ .$$

If we suppose that $|\mathcal{T}| \subseteq A^*\$ \times B^*\$$, the *pure* behaviour of $\mathcal{T}$ is thus described by

$$\forall (f,g) \in A^* \times B^* \qquad (f,g) \in |\mathcal{T}|_\$ \quad \Longleftrightarrow \quad g\$ \in \lambda \cdot f\mu \cdot (\$\mu \cdot \nu) \ .$$

Logically, the *matrix representation with endmarker* of a relation θ from A^* to B^* is a triple (λ, μ, ν), where μ is a morphism from A^* to $(\mathrm{Rat}\, B_\$^*)^{Q \times Q}$, and where λ and ν are respectively row and column vectors of dimension Q with coefficients in $\mathrm{Rat}\, B_\$^*$ such that

$$\forall (f,g) \in A^* \times B^* \qquad (f,g) \in \widehat{\theta} \quad \Longleftrightarrow \quad g\$ \in \lambda \cdot f\mu \cdot \nu \ . \tag{5.2}$$

Theorem 5.2 *A rational relation on $A^* \times B^*$ is deterministic if and only if it has a prefix matrix representation with endmarker.*

Proof. The condition is necessary as a direct consequence of Proposition 5.21 and the equation of Definition 5.2, together with the fact that we can always suppose that a deterministic relation on $A^* \times B^*$ is realised by a transducer with endmarker whose behaviour is contained in $A^*\$ \times B^*\$$ (Corollary 5.3).

The proof of the fact that the condition is sufficient is left as an exercise. ■

Exercise

- 5.10 Prove the sufficient conditions of Proposition 5.21 and Theorem 5.2.

5.6 An example: the map equivalence of a morphism

Morphisms, and hence inverse morphisms, are, like intersections with rational sets, (pure) deterministic relations. This implies furthermore that (pure) deterministic relations are not closed under composition. The following result gives a non-trivial example of a deterministic relation. Recall that a morphism φ from A^* to B^* is called *continuous* if $A\varphi \subset B^+$; that is, if no letter of A is sent by φ to 1_{B^*}.

Proposition 5.22 *The map equivalence of a morphism from one free monoid to another is a deterministic relation. If in addition the morphism is continuous, the relation is pure deterministic.*

Let us start by giving an example.

Example 5.2 (continued) Let us again consider the morphism $\varphi_1 \colon A^* \to B^*$:

$$a\varphi_1 = x \ , \quad b\varphi_1 = y\,x \quad \text{and} \quad c\varphi_1 = x\,y \ .$$

We saw that the relation $\varphi_1\varphi_1^{-1}$ from A^* to it itself realised by the transducer in Figure 5.8. *cf. Exam. 1.12, p. 544*

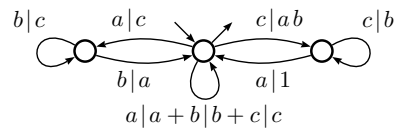


Figure 5.8: A transducer for $\varphi_1\varphi_1^{-1}$

A normalisation and a rearrangement of this automaton lead easily to the deterministic automaton shown in Figure 5.9. The states of Q_B are shown in grey and the names of the states correspond to the notation used below in the proof of the general case. □

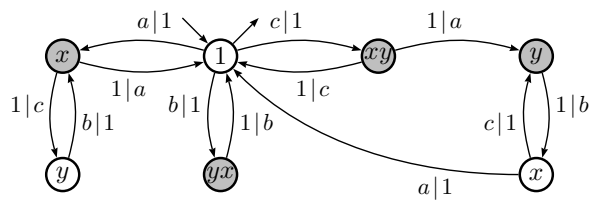


Figure 5.9: A deterministic automaton for $\varphi_1\varphi_1^{-1}$

Some examples also show that the proposition cannot be extended to arbitrary deterministic relations. *cf. Exer. 5.12*

Proof of Proposition 5.22. Let us first consider a continuous morphism $\varphi \colon A^* \to B^*$. Let S be the set of prefixes of $A\varphi$:

$$S = \{s \in B^* \mid \exists a \in A \quad s \leqslant a\varphi\} \ .$$

The transducer $\mathcal{T}$ which we shall construct will have the set of states

$$Q = S \times \{1, 2\}$$

and as $(1_{B^*}, 1)$, its unique initial state, which is also the unique final state. The transitions of $\mathcal{T}$, described in functional form, are given by

$$(s,1) \circ (a, 1_{A^*}) = \begin{cases} ((a\varphi)^{-1}s, 1) & \text{if} \quad a\varphi \leqslant s \\ (s^{-1}(a\varphi), 2) & \text{if} \quad s < a\varphi \\ \text{undefined} & \text{otherwise} \end{cases} \tag{5.3}$$

$$(s,2) \circ (1_{A^*}, a) = \begin{cases} ((a\varphi)^{-1}s, 1) & \text{if} \quad a\varphi \leqslant s \\ (s^{-1}(a\varphi), 2) & \text{if} \quad s < a\varphi \\ \text{undefined} & \text{otherwise} \ . \end{cases} \tag{5.4}$$

We show by induction on $|f| + |g|$ that for all s in S we have

$$s\,(g\varphi) = f\varphi \qquad \Longleftrightarrow \qquad (s,1) \circ (f,g) = (1_{B^*},1) \tag{5.5}$$

$$s\,(f\varphi) = g\varphi \quad \text{and} \quad s \neq 1_{B^*} \qquad \Longleftrightarrow \qquad (s,2) \circ (f,g) = (1_{B^*},1)\ . \tag{5.6}$$

The transitions defined by (5.3) gives (5.5) and the transitions defined by (5.4) gives (5.6) for $|f| + |g| = 1$.

Suppose that $s\,(g\varphi) = f\varphi$; then $|f| \geqslant 1$, $f = a\,f'$ with a in A and

(a) *either* $a\varphi \leqslant s$, in which case we have

$$(s,1) \circ (a,1_{A^*}) = (s',1) \qquad \text{and} \qquad s'\,(g\varphi) = f'\varphi\ ;$$

(b) *or* $s < a\varphi$, in which case we have

$$(s,1) \circ (a,1_{A^*}) = (s',2) \qquad \text{and} \qquad s'\,(f'\varphi) = g\varphi\ .$$

In both cases, the induction hypothesis gives

$$(s,1) \circ (f,g) = (1_{B^*},1)\ .$$

Suppose, conversely, that $(s,1) \circ (f,g) = (1_{B^*},1)$; from the definition of the transitions of $\mathcal{T}$, it follows that necessarily $f = a\,f'$ with a in A and

(c) *either* $a\varphi \leqslant s$, in which case

$$(s,1) \circ (a,1_{A^*}) = (s',1) \qquad \text{and} \qquad (s',1) \circ (f',g) = (1_{B^*},1)\ ;$$

(d) *or* $s < a\varphi$, in which case

$$(s,1) \circ (a,1_{A^*}) = (s',2) \qquad \text{and} \qquad (s',2) \circ (f',g) = (1_{B^*},1)\ .$$

In both cases, the induction hypothesis gives

$$s\,(g\varphi) = f\varphi\ .$$

The equivalence (5.6) is treated in exactly the same way.

Finally, the fact that $(1_{B^*},1)$ is both the unique initial state and the unique final state implies that (5.5), with $s = 1_{B^*}$, gives exactly $\varphi\varphi^{-1} = |\mathcal{T}|$.

We now consider the case where $\varphi\colon C^* \to B^*$ is not continuous. There exists a partition $C = A + D$ such that the restriction of φ to A^* is a continuous morphism and for all d in D, $d\varphi = 1_{B^*}$. It is clear that $(u,v) \in \varphi\varphi^{-1}$ if and only if u and v can be written

$$u = x_1 f_1 x_2 f_2 \cdots x_p f_p x_{p+1} \qquad \text{and} \qquad v = y_1 g_1 y_2 g_2 \cdots y_q g_q y_{q+1}\ , \tag{5.7}$$

with the f_i and g_j in A, the x_l and y_k in D^*, and such that

$$(f_1 f_2 \cdots f_p, g_1 g_2 \cdots g_q) \in \varphi\varphi^{-1}\ .$$

We construct the transducer defined above for the continuous restriction of φ to A^* and we add, for all s in S and all d in D, the transitions

$$(s,1)\circ(d,1_{C^*}) = (s,1) \qquad \text{and} \qquad (s,2)\circ(1_{C^*},d) = (s,2) \ .$$

It is not hard to verify that the behaviour of the transducer thus transformed is the set of pairs (u,v) such that $y_{q+1} = 1_{D^*}$ in the factorisation (5.7). To obtain the complete graph of $\varphi\varphi^{-1}$, so we can also accept the pairs (u,v) with $y_{q+1} \neq 1_{D^*}$, we must use an endmarker, and add to the transducer two other states t and t' with the transitions

$$(1_{B^*},1)\circ(\$,1_{C^*}) = t \ , \quad t\circ(1_{C^*},\$) = t' \ , \qquad \text{and} \qquad t\circ(1_{C^*},d) = t \quad \text{for all } d \text{ in } D \ ,$$

and take t' as the final state, instead of $(1_{B^*},1)$. ■

Exercises

5.11 Using the morphism φ_1 of Example 5.2, construct a morphism $\psi_1\colon A^* \to B^*$ such that $\varphi_1^{-1}\psi_1$ is *not* a deterministic relation from B^* to itself.

• 5.12 Give an example of a deterministic relation θ such that $\theta\theta^{-1}$ is not a deterministic relation.

5.13 We shall say that a morphism $\kappa\colon C^* \to A^*$ is *prefix* if $(C)\kappa$ is a prefix set of A^*. Prove the following property:

Property 5.4 *Let $\theta\colon C^* \to B^*$ be a deterministic relation and $\kappa\colon C^* \to A^*$ a prefix morphism. Then, the relation $\kappa^{-1}\theta\colon A^* \to B^*$ is deterministic.*

• 5.14 Prove:

Property 5.5 *Let $\theta\colon A^* \to B^*$ be a deterministic relation, K a rational of A^*, and L a rational of B^*. The relations $\iota_K\theta$ and $\theta\iota_L$ are deterministic.*

6 Synchronisation of transducers

This aim of this section is to study rational relations which are computable by transducers that we call *letter-to-letter*. These are automata whose transitions are labelled by pairs of letters (or by n-tuples of letters if we are discussing n-ary relations), or again, referring to the 2-tape (or k-tape) Rabin–Scott model, pairs of rational relations computable by automata whose read heads move simultaneously over the 2 (or k) tapes. *cf. Sec. 1.6, p. 545*

The point of departure is the fact that length-preserving relations, under which a word and its image have the same length, can always be recognised by such transducers. The main idea underlying this is that of *resynchronisation*, which can be applied just as well to relations (Lemma 6.1) as to transducers (Proposition 6.9). This idea translates the link which can be made between the *result of computations* of an automaton and the *properties of these computations themselves*; or, which comes to

the same thing, between the properties of a rational relation and those of a rational expression which denotes it; in other words, what we have described already several times as a *Fatou property*. All this applies in fact not only to relations that preserve length but to those where the difference between the length of a word and that of its image is bounded, relations which we shall call 'of bounded length discrepancy', or a 'bld-relation', with a suitable generalisation of letter-to-letter transducers.

One further step in the generalisation of letter-to-letter transducers will allow us to define a new family of rational relations which we shall call 'synchronous relations' and which form without doubt the largest subfamily of rational relations so far defined that is an effective Boolean algebra.

We shall consider mainly relations from A^* to B^*, i.e. binary relations, whose graphs are subsets of $A^*\times B^*$. On the whole, the definitions and results extend easily to k-ary relations, whose graphs are subsets of $A_1^*\times A_2^*\times\cdots\times A_k^*$. This general case will therefore remain essentially implicit. Some definitions will nonetheless be designed particularly to allow generalisation to k-ary relations, and could have been simpler if this was not our aim (for example, the length discrepancy of a relation, synchronous relations).

Throughout the section, relations are considered mostly via their graphs, of which we study, characterise and prove certain properties; they are written not as lower-case Greek letters as before but by upper-case Roman letters which primarily denote graphs, subsets of $A^*\times B^*$ (or $A_1^*\times A_2^*\times\cdots\times A_k^*$). Also for this reason we shall speak in this section of the *composition* of two relations to avoid confusion with the *product* of their graphs.

6.1 Rational relations of bounded length discrepancy

6.1.1 Definitions, notation and conventions

(i) The set of pairs in $A^*\times B^*$ of words of equal length is a *free submonoid* of $A^*\times B^*$, generated by $A\times B$.

(ii) By *letter-to-letter transducer* we mean an automaton over $A^*\times B^*$ whose transitions are labelled by elements of $A\times B$; a letter-to-letter transducer is therefore an automaton over a free monoid.

(iii) Every element $u=(f,g)$ of $A^*\times B^*$ factorises uniquely into $u=u'u''$ with u' in $(A\times B)^*$ and u'' in $(A^*\times 1_{B^*})\cup(1_{A^*}\times B^*)$.

(iv) To be more precise, we shall say that the *length discrepancy* of a pair of words $u=(f,g)$ is the pair of integers

$$p=(\mathsf{max}(|f|-|g|,0),\mathsf{max}(0,|g|-|f|))\quad;$$

in other words, p is an element of $\mathbb{N}^2$ one of whose coordinates is zero, or, which amounts to the same thing, p is an element of $\mathbb{N}^2\setminus\big[(1,1)+\mathbb{N}^2\big]$.

We could have called the 'length discrepancy' of a pair (f,g) the *difference in length* of f and g, or rather, the absolute value $\big||f|-|g|\big|$ of this difference. The notions and results that follow could then have been stated and proved just as well,

if not more simply. However, our definition enables the generalisation of the notion of length discrepancy to n-tuples of words: the length discrepancy of $(f_1, f_2, \ldots, f_n)$ is an element p of $\mathbb{N}^n \setminus [(1,1,\ldots,1) + \mathbb{N}^n]$ and

$$p_i = \max_{j \in [n]} \{|f_i| - |f_j|\} \; ;$$

that is, p_i is the difference between the length of f_i and that of the shortest of f_j, $1 \leqslant j \leqslant n$.

Having said that, to make our discussion easier we shall often use turns of phrase that refer directly to the case of binary relations. In particular, by *a length discrepancy is bounded by an integer* k we shall mean that all the components of the length discrepancy are less than or equal to k, which corresponds for pairs of words to the difference in length being less than or equal to k in absolute value.

The following notation follows the same idea:[24]

$$\mathrm{Diff}_k\,(A^* \times B^*) = (A^{\leqslant k} \times 1_{B^*}) \cup (1_{A^*} \times B^{\leqslant k}) \;\; .$$

Every element u whose length discrepancy is bounded by k factorises *uniquely* into

$$u = u'\,u'' \qquad \text{with} \qquad u' \in (A \times B)^* \quad \text{and} \quad u'' \in \mathrm{Diff}_k\,(A^* \times B^*)$$

and into

$$u = v'\,v'' \qquad \text{with} \qquad v' \in \mathrm{Diff}_k\,(A^* \times B^*) \quad \text{and} \quad v'' \in (A \times B)^* \;\; .$$

In other words, the products

$$(A \times B)^*\,(\mathrm{Diff}_k\,(A^* \times B^*)) \qquad \text{and} \qquad (\mathrm{Diff}_k\,(A^* \times B^*))\,(A \times B)^*$$

are *unambiguous* in $A^* \times B^*$. The uniqueness of this factorisation implies

$$\forall T_1, T_2 \subseteq (A \times B)^*\,,\; \forall S_1, S_2 \subseteq \mathrm{Diff}_k\,(A^* \times B^*) \qquad T_1\,S_1 \cap T_2\,S_2 = (T_1 \cap T_2)\,(S_1 \cap S_2) \;\; . \tag{6.1}$$

We abbreviate the notation to Diff_k when the alphabets A and B are understood.

(v) The *length discrepancy of a relation* R is the upper bound of the length discrepancy of the elements of R; the relation R is a *bounded-length discrepancy relation*, or *bld-relation*, if there exists an integer k such that the length discrepancy of the elements of R is bounded by k. The relation R *preserves the length* if its length discrepancy is zero. Obviously, we have:

Property 6.1 *The composition of two length-preserving relations preserves length, and the composition of two bld-relations is of bounded length discrepancy.* ■

[24] We could be more precise and distinguish between Diff_k and $\mathrm{Diff}_{\leqslant k}$ (with obvious definitions) but that would not add much.

6.1.2 Characterisation of rational bld-relations

It is easiest to characterise the relations of zero length discrepancy.

Theorem 6.1 *A length-preserving rational relation on $A^*\times B^*$ is a rational subset of $(A\times B)^*$; that is,*

$$R \in \mathrm{Rat}\, A^*\times B^* \quad \textit{and} \quad R \subseteq (A\times B)^* \quad \Longrightarrow \quad R \in \mathrm{Rat}\,(A\times B)^* \ .$$

Hence we deduce that a length-preserving rational relation can be realised by a letter-to-letter transducer. We shall not immediately roll out the series of corollaries which we can deduce from this result, but seek first to generalise it.

Theorem 6.2 *A rational relation on $A^*\times B^*$ of length discrepancy bounded by k is a finite union of products of rational subsets of $(A\times B)^*$ by subsets of $\mathrm{Diff}_k\,(A^*\times B^*)$ (respectively a finite union of products of subsets of $\mathrm{Diff}_k\,(A^*\times B^*)$ by rational subsets of $(A\times B)^*$).*

Theorem 6.1 is the instance of Theorem 6.2 for $k = 0$. The proof of the latter is based on a lemma:

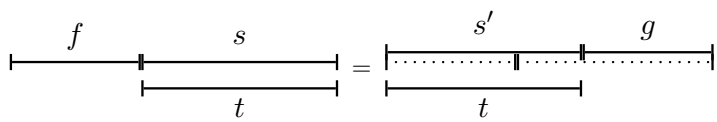
Lemma 6.1 [**Resynchronisation Lemma**] *For every H in $\mathrm{Rat}\,(A\times B)^*$ and every word f in A^*, there exists a family $\{H_{f,g} \mid g \in A^{|f|}\}$ of rationals $H_{f,g}$ of $(A\times B)^*$ indexed by words of the same length as f and such that*

$$(f, 1_{B^*})\, H = \bigcup_{g\in A^{|f|}} H_{f,g}\,(g, 1_{B^*}) \ .$$

Proof. Let us choose an integer k. For all f in A^k and all $u = (s,t)$ in $(A\times B)^*$, there exists a unique $v = (s',t)$ in $(A\times B)^*$ and a unique g in A^k such that

$$(f, 1_{B^*})\,(s,t) = (s',t)\,(g, 1_{B^*}) \ .$$

It is this equality that is shown in the following diagram:



Hence we deduce by additivity that, for every subset H of $(A\times B)^*$, there exists an $A^k\times A^k$ matrix of subsets of $(A\times B)^*$, written $\widetilde{H}$, such that

$$\forall f \in A^k \qquad (f, 1_{B^*})\, H = \sum_{g\in A^k} \widetilde{H}_{f,g}(g, 1_{B^*}) \ .$$

Still by additivity, it follows, for all G and H contained in $(A\times B)^*$ that

$$\widetilde{(H+G)} = \widetilde{H} + \widetilde{G} \ . \tag{6.2}$$

The associativity of product then gives, for all f in A^k,

$$\begin{aligned}(f,1_{B^*})\,H\,G &= [(f,1_{B^*})\,H]\;G = \left[\sum_{g\in A^k}\widetilde{H}_{f,g}\,(g,1_{B^*})\right]G = \sum_{g\in A^k}\left(\widetilde{H}_{f,g}\,(g,1_{B^*})\,G\right)\\ &= \sum_{g\in A^k}\left(\widetilde{H}_{f,g}\sum_{h\in A^k}\widetilde{G}_{g,h}\,(h,1_{B^*})\right) = \sum_{g,h\in A^k}\left(\widetilde{H}_{f,g}\,\widetilde{G}_{g,h}\,(h,1_{B^*})\right)\\ &= \sum_{h\in A^k}\left(\sum_{g\in A^k}\widetilde{H}_{f,g}\,\widetilde{G}_{g,h}\right)(h,1_{B^*}) = \sum_{h\in A^k}\left(\widetilde{H}\,\widetilde{G}\right)_{f,h}(h,1_{B^*})\ .\end{aligned}$$

That is

$$\widetilde{(H\,G)} = \widetilde{H}\,\widetilde{G}\ . \tag{6.3}$$

From (6.2) and (6.3), we conclude

$$\widetilde{(H^*)} = (\widetilde{H})^*\ . \tag{6.4}$$

Thus, the family of subsets H of $(A\times B)^*$, where the coefficients of $\widetilde{H}$ are in $\mathrm{Rat}\,(A\times B)^*$, contains the singletons and is closed under union, product and star: it therefore contains $\mathrm{Rat}\,(A\times B)^*$, which proves the lemma. ■

We then deduce, and it is in this form that we shall use the lemma, that for all u in $A^*\times B^*$ of length discrepancy equal to $(k,0)$, and for all H in $(A\times B)^*$ we have

$$u\,H = \bigcup_{g\in A^k} u'\,H_{u'',g}\,(g,1_{B^*})\ , \tag{6.5}$$

where $u = u'\,u''$ is the factorisation of u in $(A\times B)^*\,\mathrm{Diff}_k$.

Proof of Theorem 6.2. By induction on the star height of R in $\mathrm{Rat}\,A^*\times B^*$. If R is of height 0, then R is a finite set of elements whose length discrepancy is bounded by k and hence is in the required form. If R is of height $h+1$, then R is a finite union of monomials of the form *cf. § I.6.1.2, p. 159*

$$P = u_0\,R_1^*\,u_1\,R_2^*\,u_2\cdots u_{r-1}\,R_r^*\,u_r\ ,$$

with the u_i in $A^*\times B^*$, and the R_i in $\mathrm{Rat}\,A^*\times B^*$ of star height h. If any of the R_i contains an element of non-zero length discrepancy, then R_i^* is of unbounded length discrepancy, and hence so is R. Thus, each R_i preserves length, and by the induction hypothesis belongs to $\mathrm{Rat}\,(A\times B)^*$, hence so does R_i^*.

Lemma 6.1 thus ensures that $u_0\,R_1^*$ is a finite union of sets of the form $S\,v$ with S in $\mathrm{Rat}\,(A\times B)^*$ and the length discrepancy of v is equal to that of u_0. By distributivity, P is a finite union of monomials of the form

$$S\,v\,u_1\,R_2^*\,u_2\cdots u_{r-1}\,R_r^*\,u_r\ .$$

By iterating the process, we obtain that P is a finite union of monomials of the form $L\,w$ with L in $\mathrm{Rat}\,(A\times B)^*$ and w in Diff_k, where k is a bound on the length discrepancy of the element $u_0\,u_1\cdots u_r$. ■

Remark 6.1 This proof is identical in its principle, and also almost in its practice, to that of Proposition II.6.4. It is not however known how to give the two results a single statement. □

cf. p. 304

As all the steps of the proof of Lemma 6.1 and Theorem 6.2 are effective, we have:

Corollary 6.2 *We can decide whether a rational relation is of bounded length discrepancy and can compute the length discrepancy of a rational relation.* ■

We write $\mathrm{Rat}_{\mathsf{bld}}\,(A^*\times B^*)$ for the family of rational bld-relations.

The notion of a *stable monoid* in $\mathbb{N}^n$ enables the following generalisation of Theorem 6.1, a generalisation whose proof we shall leave as an exercise. Recall that we say a finitely generated submonoid M of $\mathbb{N}^n$ is *stable* if it is the intersection $M = C\cap H$ of a *closed cone* C of $\mathbb{Q}_+^n$ and of a subgroup H of $\mathbb{Z}^n$. We write $\lambda\colon A^*\times B^* \to \mathbb{N}^2$ for the *length* morphism: $\lambda\colon (f,g)\mapsto (|f|,|g|)$.

cf. Exer. 6.3

cf. Exer. II.7.10, p. 333 and p. 370

Theorem 6.3 *Let M be a* stable submonoid *of $\mathbb{N}^2$ and $S = M\lambda^{-1}$ its inverse image in $A^*\times B^*$. A rational relation on $A^*\times B^*$ whose graph is contained in S is a rational subset of S; that is:*

$$R\in \mathrm{Rat}\,A^*\times B^* \quad \textit{and} \quad R\subseteq S \quad\Longrightarrow\quad R\in \mathrm{Rat}\,S\ .$$ ■

6.1.3 Translation into automata theoretic terms, and corollaries

In the next subsection we shall rework the same results using automata rather than rational expressions, thus achieving less elegant but more efficient proofs. We can however directly transcribe Theorem 6.2 into automata-theoretic terms. Let $\mathcal{A} = \langle Q, A^*\times B^*, E, I, T\rangle$ be a finite automaton over $A^*\times B^*$. We know that it does not change the power of the model, i.e. the class of relations accepted by finite transducers, if the coefficients of the vectors I or T are rational subsets of $A^*\times B^*$. To distinguish this case from the simpler case where I and T are boolean vectors, we shall say that $\mathcal{A}$ is a transducer *with an initial function*, respectively *with a final function*, or *with initial and final functions*. Theorem 6.2 then implies:

Corollary 6.3 *A rational bld-relation with bound k is realised by a finite letter-to-letter transducer with a final function in $\mathfrak{P}(\mathrm{Diff}_k)$ (resp. by a finite letter-to-letter transducer with an initial function in $\mathfrak{P}(\mathrm{Diff}_k)$).* ■

A simple instance of the Fundamental Theorem, this proposition is not without interest since the letter-to-letter transducers are automata *over a free monoid*, hence susceptible to being determinised.[25]

Proposition 6.4 *Every finite letter-to-letter automaton with a final function in $\mathfrak{P}(\mathrm{Diff}_k)$ is equivalent to a deterministic complete finite letter-to-letter automaton with a final function in $\mathfrak{P}(\mathrm{Diff}_k)$.*

[25] In this case, I prefer to call them 'automata' rather than 'transducers'. In particular, deterministic letter-to-letter automata are not formally deterministic transducers, even if they can easily be transformed into them.

Proof. Let $\mathcal{A} = \langle Q, A\times B, E, I, T \rangle$ be a letter-to-letter automaton with a final function. We use the standard subset construction to compute the determinisation of $\mathcal{A}$: $\mathcal{A}_{\mathsf{det}} = \langle \mathfrak{P}(Q), A\times B, F, \{I\}, U \rangle$ and the final function U is defined by

$$\forall P \subseteq Q \qquad U_P = \bigcup_{q\in P} T_q \ .$$

The equivalence between $\mathcal{A}$ and $\mathcal{A}_{\mathsf{det}}$ is proved by the sequence of implications:

$$\begin{aligned} u = u'u'' \in |\mathcal{A}| \quad &\Longleftrightarrow \quad \exists i \in I\,,\ \exists p \in Q \qquad i \xrightarrow[\mathcal{A}]{u'} p \quad \text{and} \quad u'' \in T_p \\ &\Longleftrightarrow \quad I \xrightarrow[\mathcal{A}_{\mathsf{det}}]{u'} P \quad \text{and} \quad u'' \in U_P \quad \Longleftrightarrow \quad u = u'u'' \in |\mathcal{A}_{\mathsf{det}}| \ . \end{aligned}$$ ■

Corollary 6.5 *The family of rational relations with length discrepancy bounded by k is an effective Boolean algebra.*

Proof. Let $\mathcal{R}'$ and $\mathcal{R}''$ be the relations realised by the two letter-to-letter automata

$$\mathcal{A}' = \langle Q', A\times B, E', I', T' \rangle \qquad \text{and} \qquad \mathcal{A}'' = \langle Q'', A\times B, E'', I'', T'' \rangle \ .$$

The *cartesian product* of $\mathcal{A}'$ and $\mathcal{A}''$

$$\mathcal{A}'\times\mathcal{A}'' = \langle Q'\times Q'', A\times B, F, I'\times I'', T'\times T'' \rangle \ ,$$

is defined as usual as far as the initial states and transitions are concerned:

$$F = \{\big((p', p''), (a, b), (q', q'')\big) \mid \big(p', (a, b), q'\big) \in E' \quad \big(p'', (a, b), q''\big) \in E''\} \ .$$

The final function $\mathcal{A}'\times\mathcal{A}''$ is obtained by taking the intersection of the final functions of $\mathcal{A}'$ and $\mathcal{A}''$:

$$\forall q', q'' \in Q'\times Q'' \qquad (T'\times T'')_{q',q''} = T'_{q'} \cap T''_{q''} \ .$$

From the standard proof and from (6.1), it follows that

$$|\mathcal{A}'\times\mathcal{A}''| = |\mathcal{A}'| \cap |\mathcal{A}''| = R' \cap R'' \ .$$

By Proposition 6.4, we can assume that $\mathcal{A}'$ is complete deterministic. By replacing the final function T' by its complement with respect to Diff_k,

$$\forall q' \in Q' \qquad U'_{q'} = \mathrm{Diff}_k \setminus T'_{q'} \ ,$$

we obtain an automaton $\mathcal{C} = \langle Q', A\times B, E', I', U' \rangle$ whose behaviour is the complement with respect to $(A\times B)^*\,\mathrm{Diff}_k$ of that of $\mathcal{A}'$:

$$|\mathcal{C}| = ((A\times B)^*\,\mathrm{Diff}_k) \setminus |\mathcal{A}'| \ .$$

We can deduce closure under difference since, because R' and R'' are both contained in $(A\times B)^*\,\mathrm{Diff}_k$, we have

$$R'' \setminus R' = R'' \cap \big[((A\times B)^*\,\mathrm{Diff}_k) \setminus R'\big] \ .$$ ■

Remark 6.2 The Boolean algebra in question in Corollary 6.5 is not a subalgebra of $\mathfrak{P}(A^* \times B^*)$ because its maximum element is $(A \times B)^* \operatorname{Diff}_k$ and not $A^* \times B^*$. If we want to preserve the complement relation's intrinsic character (not depending on k), the statement of Corollary 6.5 must be rephrased as:

The family of rational relations with length discrepancy bounded by k is closed under union and difference, and these operations are effective. □

Exercises

• 6.1 Supposing that all the alphabets have two letters, how many n-tuples are there in $A_1^* \times A_2^* \times \cdots \times A_n^*$ whose length discrepancy is bounded by k?

• 6.2 Write $\operatorname{Fin}(A^* \times B^*) = \mathfrak{P}_{\mathsf{fin}}(A^* \times B^*)$ for the family of finite relations. Show

$$\operatorname{Fin}(A^* \times B^*) = \operatorname{Rec}(A^* \times B^*) \cap \operatorname{Rat}_{\mathsf{bld}}(A^* \times B^*) \ . \tag{6.6}$$

cf. Exer. II.7.10, p. 333 • 6.3 Let M be a *stable submonoid* of $\mathbb{N}^2$ and $S = M\lambda^{-1}$ be a submonoid of $A^* \times B^*$.

(a) Prove: $R \in \operatorname{Rat} A^* \times B^*$ and $R \subseteq S \implies R \in \operatorname{Rat} S$ (Theorem 6.3).

(b) Prove:

Corollary 6.6 *Given a stable monoid M of $\mathbb{N}^2$, we can decide whether a rational relation is contained in $M\lambda^{-1}$.*

(c) What can we deduce for relations such that every word in the image of any word f is of length less than or equal to that of f?

(d) Give a necessary and sufficient condition on M for $\operatorname{Rat} S$ to be a Boolean algebra.

6.4 (a) Give a proposition which generalises both Theorem 6.3 and Theorem 6.2.

(b) Deduce a characterisation of rational relations R such that there exists an integer k and a rational number p/q for which, for all (f, g) in R, we have $\big|\, q|f| - p|g| \,\big| \leqslant k$.

6.2 Transducers of bounded lag

We return to the properties of computations of transducers which realise bld-relations. By again using the terminology derived from the multi-tape automaton model of Rabin and Scott, we shall prove that *in the course of every computation of every automaton which realises a bld-relation, the distance between the read heads is bounded.* We first make some definitions to translate this proposition into the terms of automata considered as labelled graphs.

6.2.1 Lag in a computation or transducer

Let $\mathcal{A}$ be a transducer on $A^* \times B^*$ and c a computation of $\mathcal{A}$:

$$c \ := \ q_0 \xrightarrow{(f_1, g_1)} q_1 \xrightarrow{(f_2, g_2)} q_2 \cdots q_{n-1} \xrightarrow{(f_n, g_n)} q_n \ .$$

The *delay of c at step r* is the integer

$$\mathsf{del}(c, r) = |f_1 f_2 \cdots f_r| - |g_1 g_2 \cdots g_r| :$$

that is, the difference between the lengths of the two components of the label of the prefix of length r of the computation c. The *lag in* c is the maximum of the absolute values of the delays observed in the course of the computation c:

$$\mathsf{lag}(c) = \mathsf{max}\{|\,\mathsf{del}(c,r)| \;\big|\; 0 \leqslant r \leqslant |c|\} \ .$$

The *lag* of a transducer $\mathcal{A}$ is the upper bound of the lag in the successful computations of $\mathcal{A}$:

$$\mathsf{lag}(\mathcal{A}) = \mathsf{sup}\{\mathsf{lag}(c) \;\big|\; c \text{ successful computation of } \mathcal{A}\} \ .$$

If this bound is *finite*, we say that $\mathcal{A}$ 'is of bounded lag' or 'has a bounded lag'. In particular, a (trim) transducer has a lag equal to 0 if and only if the length discrepancy of the label of each transition is zero and hence if and only if $\mathcal{A}$ is trivially equivalent to a letter-to-letter transducer.

Lemma 6.7 *A trim transducer is of bounded lag if and only if the length discrepancy of the label of each circuit is zero.*

Proof. Let $l := p \xrightarrow{(u,v)} p$ be a circuit in $\mathcal{A}$ such that $|u| - |v| \neq 0$. Since $\mathcal{A}$ is trim, there exists a successful computation $c := i \xrightarrow{(f,g)} p \xrightarrow{(h,k)} t$ which goes through p and we can construct successful computations:

$$c_n := i \xrightarrow{(f,g)} p \xrightarrow{(u^n,v^n)} p \xrightarrow{(h,k)} t$$

with arbitrarily large lag (in absolute value).

Conversely, let N be the number of states in $\mathcal{A}$ which we suppose has the property that the length discrepancy of the label of each circuit is zero. If c is an arbitrary computation of $\mathcal{A}$,

$$c := q_0 \xrightarrow{(f_1,g_1)} q_1 \xrightarrow{(f_2,g_2)} q_2 \cdots q_{r-1} \xrightarrow{(f_r,g_r)} q_r \cdots q_{n-1} \xrightarrow{(f_n,g_n)} q_n \ ,$$

then $\mathsf{del}(c,r)$ is a function (from the interval $[0,n]$ to $\mathbb{N}$) which can have at most N distinct values. In fact, $\mathsf{del}(c,r) = \mathsf{del}(c,i)$ where q_i is the first occurrence of the state q_r in c. This value can be computed on any of the simple sub-computations of c which link q_0 to q_r. ■

Example 6.1 The transducer $\mathcal{F}_2$ in Figure 6.1 is of bounded lag since the labels of the three elementary circuits

$$q \xrightarrow{(a,a)} q \ , \quad q \xrightarrow{(ba,ab)} q \qquad \text{and} \qquad q \xrightarrow{(bb,ba)} q$$

have zero length discrepancy. □

Proposition 6.8 *A rational relation is of bounded length discrepancy if and only if every trim transducer which realises it has a bounded lag.*

Proof. The condition is obviously sufficient. Let $\mathcal{A}$ be a transducer on $A^* \times B^*$, finite and trim. We can associate each state p of $\mathcal{A}$ with a fixed computation which goes from p to a final state of $\mathcal{A}$. In this way, we obtain an integer N such that, whatever delay z is observed in a computation of $\mathcal{A}$, we are certain to be able to find in the relation recognised by $\mathcal{A}$ a pair whose length discrepancy is between $|z| - N$ and $|z| + N$. The outcome is that the two quantities 'lag of $\mathcal{A}$' and 'length discrepancy of $|\mathcal{A}|$' are either both bounded or both unbounded, and the condition is necessary. ∎

All these considerations allow us to rediscover what we know already: that the length discrepancy of a rational relation is computable. The effective transformation from an automaton to a rational expression enables us to transcribe Corollary 6.3 into an algorithm that transforms a transducer of bounded lag into a letter-to-letter transducer. However, it is more interesting to prove directly, by working on transducers, the following proposition, which is in itself a new proof of Theorem 6.2.

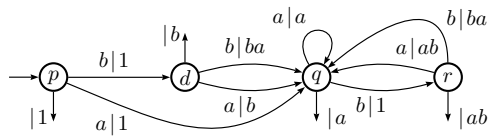


Figure 6.1: A transducer $\mathcal{F}_2$, of bounded lag

6.2.2 Resynchronisation algorithm for transducers

Proposition 6.9 *Every finite automaton over $A^* \times B^*$ of bounded lag is equivalent to a letter-to-letter transducer with a final function in $\mathfrak{P}(\mathrm{Diff}_k)$, for some k, and this letter-to-letter transducer is effectively computable.*

Proof. We start with an arbitrary *proper* automaton $\mathcal{A}$ over $A^* \times B^*$,

$$\mathcal{A} = \langle\, Q, A^* \times B^*, E, I, T \,\rangle \ .$$

As a preliminary, we note that thanks to Lemma 6.7 we can decide whether $\mathcal{A}$ is of bounded lag or not. We suppose from now on that $\mathcal{A}$ is of bounded lag. We also suppose that the initial function I takes values in $\mathbb{B}$ (that is, a state is either initial or not) but there is no label on the incoming arrows.

The first step of the algorithm consists of transforming $\mathcal{A}$ into an equivalent finite transducer $\mathcal{B}$ which has the property that we can assign to each state p a *rank* which is an integer, and which is the *common value* of the difference between the lengths $|f| - |g|$ of the label (f, g) of *all the computations* which go from an initial state to this state p. That is to say, the state 'indexes' the delay of every computation of which it is the destination (and whose source is an arbitrary initial state). Such a transducer is called *valued*.

The canonical procedure for this construction is to form the product of $\mathcal{A}$ with the *action 'length discrepancy'*; that is, to consider the *covering* of $\mathcal{A}$ where the pre-image of each state p is $p \times \mathbb{Z}$ and where the pre-image of each transition $e = \big(p, (f, g), q\big)$

cf. Exam. II.3.3, p. 268

is the set of transitions $\big((p,z),(f,g),(q,z+d)\big)$ for all z in $\mathbb{Z}$ and with $d=|f|-|g|$. Formally, let $\mathcal{A}'$ be the infinite automaton defined by

$$\mathcal{A}' = \langle\, Q\times\mathbb{Z}, A^*\times B^*, E', I\times\{0\}, T\times\mathbb{Z}\,\rangle \qquad \text{with}$$
$$E' = \{\big((p,z),(f,g),(q,z+d)\big) \mid \big(p,(f,g),q\big)\in E\,,\quad z\in\mathbb{Z}\,,\quad \text{and}\quad d=|f|-|g|\}\,.$$

The rank of each state (p,z) is z; each initial state of $\mathcal{A}'$ is of rank 0 (because of the assumption on the initial function). The following properties follow directly from the definitions:

Properties 6.2 (i) *$\mathcal{A}'$ is a covering of $\mathcal{A}$;* (ii) *$\mathcal{A}'$ is valued;*
(iii) *$\mathcal{A}$ is of bounded lag if and only if the accessible part $\mathcal{B}$ of $\mathcal{A}'$ is finite.* ■

By (i), $\mathcal{B}$ is equivalent to $\mathcal{A}$; it is finite by (iii) and valued by (ii). The rank of a state p in a valued transducer is written $\mathsf{r}(p)$.

Example 6.1 (continued) Figure 6.2 shows the effect of the above construction on the transducer $\mathcal{F}_2$ from Figure 6.1. □

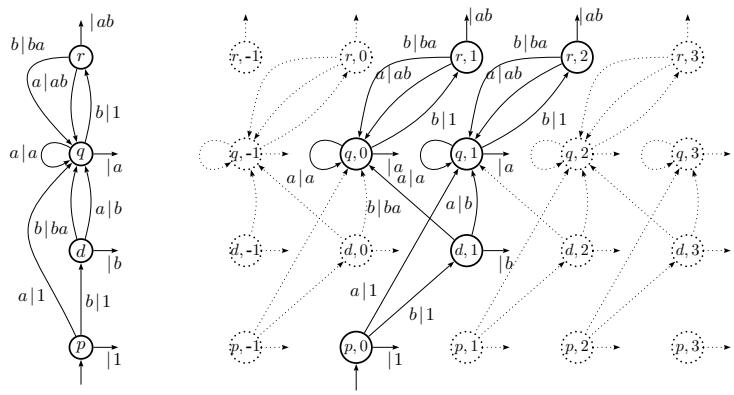


Figure 6.2: The transducers $\mathcal{F}_2$, on the left, and $\mathcal{F}_2'$, the valued covering of $\mathcal{F}_2$, on the right. The solid lines show $\mathcal{G}_2$, the (finite) accessible part of $\mathcal{F}_2'$, while the dotted lines show part of the rest of $\mathcal{F}_2'$.

Second step. We suppose from now on that $\mathcal{A}=\langle\, Q, A^*\times B^*, E, I, T\,\rangle$ is valued (and proper). We also suppose that the initial function I takes values in $\mathbb{B}$, and that there are no transitions between a state of strictly negative rank and a state of strictly positive rank, a technical requirement that is easy to fulfil by splitting any 'bad' transitions. It is clear that $\mathcal{A}$ is a letter-to-letter transducer – more precisely, an automaton over $(A\times B)^*$ – if and only if the rank of each state is 0. Hence suppose that there exist states in $\mathcal{A}$ whose rank is non-zero.

By swapping A and B if necessary – that is, by swapping the first and second components of the labels – we can assume that there exist states of strictly positive rank. The second step of the algorithm consists of constructing a kind of *co-covering* of $\mathcal{A}$ which is a valued transducer none of whose states is of strictly positive rank.

By swapping A and B again if necessary, and by restarting the procedure, we obtain a valued transducer equivalent to $\mathcal{A}$ which has neither any state of strictly positive rank nor any state of strictly negative rank; that is, a transducer with the property we require.

We set $Q = Q' \cup Q''$, where Q' is the set of states of positive or zero rank and Q'' is the set of states of strictly negative rank, and $E = E' \cup E''$, where E' is the set of transitions whose extremities are in Q'; by our hypothesis, E'' is the set of transitions one of whose extremities is in Q'' and the other is either in Q'' or is a state of rank zero.

We shall construct $\mathcal{B} = \langle R, A^* \times B^*, F, I, U \rangle$ equivalent to $\mathcal{A}$, with $R = R' \cup Q''$ and $F = F' \cup E''$, in the following manner:

$$R' = \{(p,u) \mid p \in Q',\ u \in B^{\mathsf{r}(p)}\} \qquad \text{and}$$
$$F' = \{\big((p,u),(f,u^{-1}g\,v),(q,v)\big) \mid \big(p,(f,g),q\big) \in E',\ u^{-1}g\,v \in B^*\}\ .$$

We verify:

(i) $|f| = |u^{-1}g\,v|$; indeed, $|u| = \mathsf{r}(p)$, $|v| = \mathsf{r}(q)$, $\mathsf{r}(q) = \mathsf{r}(p) + |f| - |g|$ and $|u^{-1}g\,v| = |g| + |v| - |u|$ if $u^{-1}g\,v$ is in B^*.

(ii) For each (q,v) in R' and each $\big(p,(f,g),q\big)$ in E', there exists a *unique* (p,u) in R' such that $u^{-1}g\,v$ is in B^*: thus u is the prefix of length $\mathsf{r}(p)$ of $g\,v$.

Suppose at first that the final function T satisfies

$$\forall q \in Q' \qquad T_q = (h,k) \qquad \text{with} \quad |k| \geqslant \mathsf{r}(q)\ . \tag{6.7}$$

Then, for the unique v, prefix of length $\mathsf{r}(q)$ of k, we have $U_{(q,v)} = (h, v^{-1}k)$ and for all the other v's, we have $U_{(q,v')} = 0$.

Under this assumption, and if we project the labels onto their first component, $\mathcal{B}$ is a co-covering of $\mathcal{A}$, by (ii). There is therefore a bijection between the successful computations of $\mathcal{B}$ and those of $\mathcal{A}$. It remains to verify that the second components of the label of two computations in bijection are equal, which is immediate by induction on the length of the parts of the computations that are in Q'.

Furthermore, by (i), all the states in R' have rank zero and hence $\mathcal{B}$ is a valued transducer none of whose states has a strictly positive rank. Before going further and finishing the proof without the assumption (6.7), let us give an example.

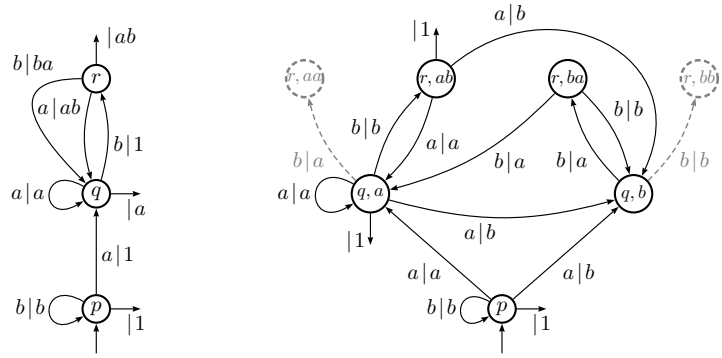
Example 6.2 We start with the valued transducer[26] $\mathcal{D}_2$: the rank of p is 0, that of q, 1, and that of r, 2. This transducer satisfies condition (6.7) and Figure 6.3 (b) shows the effect of the previous construction which transforms $\mathcal{D}_2$ into a letter-to-letter transducer. □

To deal with the case where, for some q in Q', T_q belongs to $A^* \times B^{<\mathsf{r}(q)}$, we shall use a trick.[27] Let \$ be a symbol that belongs to no alphabet, which we shall use here

[26] A simplified version of $\mathcal{F}_2$ from Example 6.1; *cf.* Exercise 6.7 below.

[27] Analogous to that of the Bedouin sage and his extra camel. *A priori*, I do not like using this sort of sleight of hand in proofs, but in this case the trick does not so much conceal a difficulty as it reveals the true nature of the algorithm.

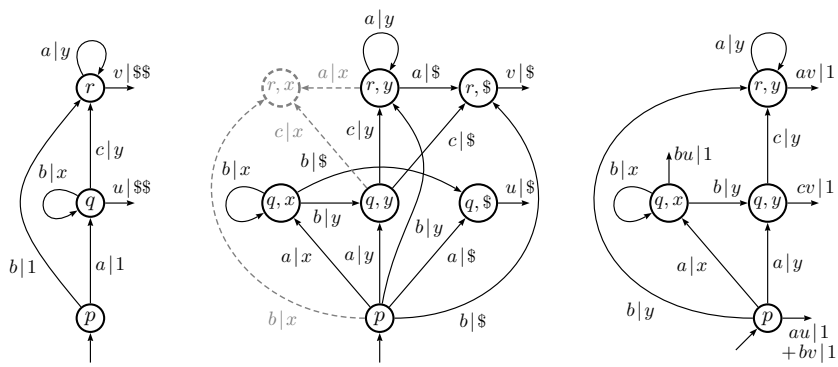
not as an endmarker but as a padding symbol. We start by transforming $\mathcal{A}$ into $\mathcal{A}'$ which satisfies (6.7) by adding a sufficient number of padding symbols \$ to the second component of every value of the final function. Then we construct a letter-to-letter transducer $\mathcal{B}'$ with the preceding algorithm.



(a) transducer $\mathcal{D}_2$ (b) result of the algorithm on $\mathcal{D}_2$

Figure 6.3: Resynchronisation of the transducer $\mathcal{D}_2$

In $\mathcal{B}'$, we identify the computations all of whose transitions have a label whose second component contains a \$. We note that they are finite and add their label to the final function of their source. Having done this, we remove all the states (q, w) where w contains a \$ and delete all the \$ symbols in the final functions. The resulting transducer $\mathcal{B}$ is letter-to-letter and equivalent to $\mathcal{A}$. (These three phases appear in the example shown in Figure 6.4.) ■



(a) (b) (c)

Figure 6.4: Usage of \$ to deal with the case of arbitrary final functions

6.2.3 Composition of letter-to-letter transducers

As we have already said, the composition of bld-relations is obviously a bld-relation, and the composition of two length-preserving relations preserves length. For the latter,

cf. Th. 3.6, p. 567

realised by letter-to-letter transducers, the Theorem of Composition of Representations takes a particularly simple form that we have already used twice as an example of general constructions, and of which we can give an elementary description:

cf. Exer. 1.26 and Exam. 3.1, p. 565

Proposition 6.10 *Let $\alpha\colon A^* \to B^*$ and $\beta\colon B^* \to C^*$ be two rational relations which are length-preserving, and that are realised by the letter-to-letter transducers $\mathcal{A} = \langle\, Q, A\times B, E, I, T \,\rangle$ and $\mathcal{B} = \langle\, R, B\times C, F, J, U \,\rangle$ respectively.*

The relation $\alpha \circ \beta\colon A^* \to C^*$ is realised by the transducer:

$$\mathcal{C} = \langle\, Q\times R, A\times C, H, I\times J, T\times U \,\rangle \ ,$$
$$H = \{\big((p,r),(a,c),(q,s)\big) \mid \exists b \in B \quad (p,(a,b),q) \in E \quad \textit{and} \quad (r,(b,c),s) \in F\} \ . \ \blacksquare$$

Example 6.3 Let α_1, the relation from $\{a,b\}^*$ to itself which replaces every second a by a b and β_1 that which replaces every other b with an a and which does not accept words with an odd number of b's. Figure 6.5 represents the transducers that realise α_1 (on the left, vertically), β_1 (at the top, horizontally), and $\alpha_1 \circ \beta_1$. □

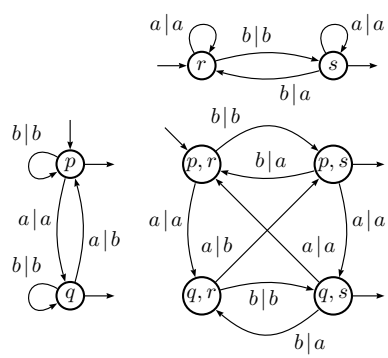


Figure 6.5: A composition of letter-to-letter transducers

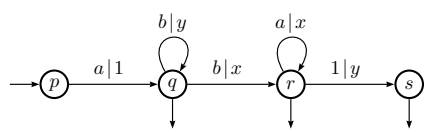
Exercises

6.5 Explain why a transducer has zero lag if, and only if, it is 'trivially equivalent to a letter-to-letter transducer' and not simply if, and only if, it *is* a letter-to-letter transducer.

• 6.6 Explain why the last sentence of the proof of Lemma 6.7, '*This value can be computed on any of the simple sub-computations of c which link q_0 to q_r*', is necessary; that is, there could be several distinct simple sub-computations which link q_0 to q_r, and true; in other words, these sub-computations all have the same delay at the destination.

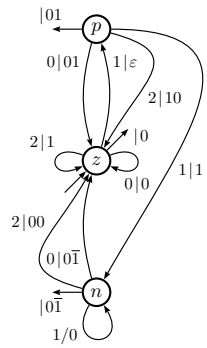
6.7 Which relations do the transducers $\mathcal{F}_2$ and $\mathcal{D}_2$ of Examples 6.1 and 6.2 realise?

• 6.8 Apply the resynchronisation algorithm to the transducer in the adjacent figure.



6.9 Prove Properties 6.2.

6.10 Apply the resynchronisation algorithm to the transducer in the adjacent figure (the input alphabet is $\{0,1,2\}$, the output alphabet is $\{\bar{1},0,1\}$; because of this, the empty word is written ε).

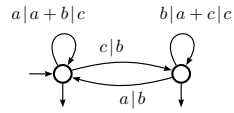


6.11 Estimate the complexities of the resynchronisation algorithms that can be deduced from the proofs of Theorem 6.2 and Proposition 6.9.

• 6.12 (a) Let $\sigma\colon A^* \to A^*$ be the map which relates a word f to the word obtained by placing the last letter of f at the beginning: $(h\,a)\sigma = a\,h\,,\ \forall a \in A\,,\ \forall h \in A^*$. Give the deterministic letter-to-letter transducer which realises σ.
(b) Show that $\overset{\star}{\sigma}$ (the iteration of σ) *is not* a rational relation.

• 6.13 Let α_1 and β_1 be the relations defined in Example 6.3; construct the transducers which realise
(a) $\alpha_1 \circ \alpha_1$; (b) $\beta_1 \circ \alpha_1$.

• 6.14 Let δ_1 be the relation realised by the transducer in the adjacent figure. Compute the transducer which realises $\delta_1 \circ \delta_1$.



6.3 Synchronous relations

We have seen that every family of rational bld-relations with bound k is closed under intersection and difference but is not a Boolean subalgebra of $\mathfrak{P}(A^* \times B^*)$. Our aim in this subsection is to overcome this difficulty and to define a family of rational relations which directly generalises bld-relations and which, like deterministic relations, is closed under complement. We write *cf. Rem. 6.2*

$$\mathrm{Diff}_{\mathsf{rat}}\,(A^* \times B^*) = \{(S \times 1) \cup (1 \times T) \mid S \in \mathrm{Rat}\,A^*\ ,\quad T \in \mathrm{Rat}\,B^*\}\ .$$

In some sense, $\mathrm{Diff}_{\mathsf{rat}}\,(A^* \times B^*)$ is the generalisation of $\mathrm{Diff}_k\,(A^* \times B^*)$ of finite sets to rational sets. As above, we write $\mathrm{Diff}_{\mathsf{rat}}$ when the alphabets A and B are, or can be, implicit.

6.3.1 Another family of rational relations

Definition 6.1 We call a rational relation *(left) synchronous* if it is recognised by a letter-to-letter transducer whose final function takes values in $\mathrm{Diff}_{\mathsf{rat}}$. □

Using Kleene's Theorem (and Theorem 6.1), this definition is paraphrased as:

Proposition 6.11 *A rational relation is (left) synchronous if and only if its graph is a finite union of products of graphs of length-preserving rational relations with elements of* $\mathrm{Diff}_{\mathsf{rat}}$. ■

To simplify discussion, we say *synchronous relation* for *synchronous rational relation* and also that *a relation is synchronous* for *a relation is rational and synchronous* for 'synchronous' necessarily refers to a realisation by automata. We shall also continue in what follows to identify a relation with its graph. (Similarly, we shall speak explicitly of the product *by composition* of relations.) We now state, for example:

Corollary 6.12 *A relation whose graph is the product of a length-preserving relation with a synchronous relation is synchronous.* ■

It follows directly from Definition 6.1 that:

Property 6.3 *The inverse of a synchronous relation is synchronous.* ■

Examples 6.4 (i) The relation $R_1 = (a,a)^*(1,b)^*$ is synchronous.

cf. Exam. 1.1, p. 526

(ii) More generally, let $K \in \operatorname{Rat} A^*$; the relation $\omega_K \colon A^* \to A^*$ – that is, the 'product on the right by K', $f\omega_K = f\,K$ – is synchronous: $\widehat{\omega_K} = \mathcal{D}^*(1_{A^*} \times K)$, where $\mathcal{D} = \bigcup_{a \in A}(a,a)$.

(iii) Conversely, let $L \in \operatorname{Rat} A^*$; the relation $\psi_L \colon A^* \to A^*$ – that is, the 'right quotient by L', $f\psi_L = f\,L^{-1}$ – is synchronous: $\psi_L = \omega_L^{-1}$ (or $\widehat{\psi_L} = \mathcal{D}^*(L \times 1_{A^*})$).

(iv) Suppose that the alphabet A is totally ordered. The relation τ_1 which makes a word f correspond to the set of words which are greater than f in the lexicographic order is a synchronous relation:

cf. Exer. 1.3, p. 528

$$\widehat{\tau_1} = \mathcal{D}^*\mathcal{V} + \mathcal{D}^*\mathcal{S}\mathcal{U}\,, \quad \text{with} \quad \mathcal{V} = 1_{A^*} \times A^+\,, \quad \mathcal{S} = \sum_{a,b \in A,\, a \prec b} (a,b) \quad \text{and} \quad \mathcal{U} = A^* \times A^*\,.$$

Likewise, the relation σ_1 which makes a word f correspond to the set of words which are greater than f for the radix order $\sqsubseteq$ is a synchronous relation:

$$\widehat{\sigma_1} = \mathcal{E}^*\mathcal{V} + \mathcal{D}^*\mathcal{S}\,\mathcal{E}^*\ , \qquad \text{with} \quad \mathcal{E} = A \times A\ .$$

(v) It follows from Corollary 6.3 (but not from Definition 6.1 itself) that a rational bld-relation is synchronous. □

Since a letter-to-letter transducer (with a final function) realises a relation which is both of *finite image* and of *finite pre-image* if and only if the final function sends each state to a pair of finite sets (and thus realises in this case a bld-relation) we have:

Corollary 6.13 *A rational relation is of bounded length discrepancy if and only if it is synchronous and of finite image and pre-image.* ■

Remark 6.3 *Right synchronous* rational relations are defined symmetrically: relations realised by letter-to-letter transducers whose *initial function* takes values in $\mathrm{Diff}_{\mathsf{rat}}$. Contrary to what holds for bld-relations (which are of bounded left and right length discrepancy at the same time) a right synchronous rational relation is not necessarily left synchronous, and *vice versa*: for example, $R_1 = (a,a)^*(1,b)^*$ is left synchronous but not right synchronous. □

Remark 6.4 The preceding definition holds for binary relations. While the definition of rational bld-relations generalises directly from binary relations to relations of arbitrary arity, the same generalisation for synchronous relations, while not being difficult, requires some preparation.

We first observe that *a relation of arity* n *can also be considered as a relation of arity* m *, for all* $m \geqslant n$ *, with* $m-n$ *components equal to* 1. We write Syn_n for the family of n-ary synchronous rational relations. With this convention, $\mathrm{Syn}_n \subseteq \mathrm{Syn}_m$ for $m \geqslant n$, and we see Syn_1 is not only another notation for Rat but also for $\mathrm{Diff}_{\mathsf{rat}}$, *etc.*

We shall then say that an n-ary rational relation is synchronous if it is recognised by a finite letter-to-letter automaton over $A_1^* \times A_2^* \times \cdots \times A_n^*$ whose final function takes values in Syn_{n-1} . □

Proposition 6.14 *A recognisable relation is (left* and *right) synchronous.*

Proof. Since the synchronous relations are closed under union, and recognisable relations are finite unions of products of rationals, we can assume that the relation is of the form $K \times L$ with $K \in \mathrm{Rat}\, A^*$ and $L \in \mathrm{Rat}\, B^*$. The language K is recognised by $\mathcal{A} = \langle Q, A, E, I, T \rangle$ and L by $\mathcal{B} = \langle R, B, F, J, U \rangle$. *Cor. II.2.20, p. 254*

For each q in Q, let K_q be the set of words which are the label of a computation of $\mathcal{A}$ whose source is q and destination is in T, and let K'_q be the set of words which are the label of a computation of $\mathcal{A}$ whose source is in I and whose destination is q. We have

$$K = \bigcup_{q \in I} K_q = \bigcup_{q \in T} K'_q = \bigcup_{q \in Q} K'_q K_q \ .$$

In the same way we define L_r and L'_r for all r in R, to get

$$L = \bigcup_{r \in J} L_r = \bigcup_{r \in U} L'_r = \bigcup_{r \in R} L'_r L_r \ .$$

Then

$$K \times L = \left[\bigcup_{q \in T, r \in R} \left((K'_q \times L'_r) \cap (A \times B)^* \right) (1_{A^*} \times L_r) \right] \bigcup \left[\bigcup_{q \in Q, r \in U} \left((K'_q \times L'_r) \cap (A \times B)^* \right) (K_q \times 1_{B^*}) \right] \ .$$

Since $(A \times B)^*$ is a rational of $A^* \times B^*$ and each $(K'_q \times L'_r)$ is recognisable, their intersection is rational in $A^* \times B^*$ (Proposition II.2.6), hence in $(A \times B)^*$ (Theorem 6.1) and $K \times L$ is synchronous by Proposition 6.11. *cf. p. 239*

Likewise, we have:

$$K \times L = \left[\bigcup_{q \in T, r \in R} (1_{A^*} \times L'_r) \left((K_q \times L_r) \cap (A \times B)^* \right) \right] \bigcup \left[\bigcup_{q \in Q, r \in U} (K'_q \times 1_{B^*}) \left((K_q \times L_r) \cap (A \times B)^* \right) \right]$$

and $K \times L$ is right synchronous. ■

Corollary 6.15 *A rational relation is* left synchronous *if and only if it is recognised by a letter-to-letter transducer whose final function takes values in* $\operatorname{Rec} A^* \times B^*$.

Proof. The condition is necessary since $\operatorname{Diff}_{\mathsf{rat}}(A^* \times B^*) \subseteq \operatorname{Rec} A^* \times B^*$; it is sufficient by the preceding proposition and Corollary 6.12. ∎

Proposition 6.16 *The composition of two synchronous relations is a synchronous relation.*

We start by proving a special case:

Lemma 6.17 *Let* $\alpha\colon A^* \to B^*$ *and* $\beta\colon B^* \to C^*$ *be two length-preserving relations,* L *a rational of* B^* *and* $\omega_L\colon B^* \to B^*$ *the relation 'right product by* L*'. Then,* $\alpha \circ \omega_L \circ \beta\colon A^* \to C^*$ *is a synchronous relation.*

cf. Exam. 6.4 (ii)

Proof. Let $\mathcal{A} = \langle Q, A \times B, E, I, T \rangle$ and $\mathcal{B} = \langle R, B \times C, F, J, U \rangle$ be the two letter-to-letter transducers which realise α and β respectively, and let, as in Proposition 6.10, $\mathcal{C} = \langle Q \times R, A \times C, H, I \times J, T \times U \rangle$ be the letter-to-letter transducer which realises $\alpha \circ \beta$.

For each r in R, let β_r be the relation realised by $\mathcal{B}_r = \langle R, B \times C, F, \{r\}, U \rangle$, the transducer obtained from $\mathcal{B}$ by taking r as the initial state, and let $W_r = (L)\beta_r$.

The relation $\alpha \circ \omega_L \circ \beta$ is realised by the letter-to-letter transducer with final function $\mathcal{C}' = \langle Q \times R, A \times C, H, I \times J, Y \rangle$ where

$$Y_{q,r} = \begin{cases} 1 \times W_r & \text{if} \quad T_q = 1 \\ 0 & \text{otherwise} \end{cases}$$

an assertion which should be obvious from Figure 6.6. ∎

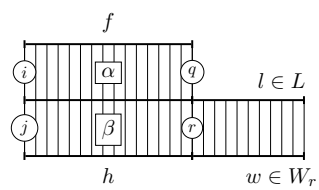


Figure 6.6: Composition of two letter-to-letter transducers

Proof of Proposition 6.16. Let $\sigma\colon A^* \to B^*$ and $\tau\colon B^* \to C^*$ be two synchronous relations. The graph of each of them is a finite union of products of length-preserving relations $\alpha\colon A^* \to B^*$ (and $\beta\colon B^* \to C^*$) with sets $(K \times 1_{B^*})$ where K is in $\operatorname{Rat} A^*$ or $(1_{A^*} \times L)$ where L is in $\operatorname{Rat} B^*$ (and, respectively, with sets $(V \times 1_{C^*})$ where V is in $\operatorname{Rat} B^*$ or $(1_{B^*} \times W)$ where W is in $\operatorname{Rat} C^*$).

By distributivity, it is enough to verify that the proposition holds for the four cases: $\widehat{\sigma} = \widehat{\alpha}(K \times 1_{B^*})$ or $\widehat{\sigma} = \widehat{\alpha}(1_{A^*} \times L)$, and $\widehat{\tau} = \widehat{\beta}(V \times 1_{C^*})$ or $\widehat{\tau} = \widehat{\beta}(1_{B^*} \times W)$.

(i) $\widehat{\sigma} = \widehat{\alpha}(K \times 1_{B^*})$ and $\widehat{\tau} = \widehat{\beta}(1_{B^*} \times W)$. Then

$$\widehat{\sigma \circ \tau} = \widehat{\alpha \circ \beta}(K \times W)$$

and $\sigma \circ \tau$ is synchronous by Corollary 6.15.

(ii) $\widehat{\sigma} = \widehat{\alpha}\,(1_{A^*} \times L)$ and $\widehat{\tau} = \widehat{\beta}\,(1_{B^*} \times W)$. Then

$$\widehat{\sigma \circ \tau} = (\alpha \circ \widehat{\omega_L} \circ \beta)\,(1_{A^*} \times W)$$

and $\sigma \circ \tau$ is synchronous by Lemma 6.17 (and Corollary 6.15).

(iii) $\widehat{\sigma} = \widehat{\alpha}\,(K \times 1_{B^*})$ and $\widehat{\tau} = \widehat{\beta}\,(V \times 1_{C^*})$. By inversion we arrive at case (ii).

(iv) $\widehat{\sigma} = \widehat{\alpha}\,(1_{A^*} \times L)$ and $\widehat{\tau} = \widehat{\beta}\,(V \times 1_{C^*})$. It follows that

$$\widehat{\sigma \circ \tau} = (\widehat{\sigma' \circ \beta}) \cup (\widehat{\alpha \circ \tau''}) \qquad \text{with}$$
$$\widehat{\sigma'} = \widehat{\alpha}\,(1_{A^*} \times (L\,V^{-1})) \qquad \text{and} \qquad \widehat{\tau''} = \widehat{\beta}\,((V\,L^{-1}) \times 1_{C^*}) \ ,$$

which is the composition of cases (ii) and (iii). ■

6.3.2 Determinisation and minimisation

As in Proposition 6.4, we use the subset construction to determinise every letter-to-letter automata whose final function takes values in $\mathrm{Diff}_{\mathsf{rat}}$. We deduce:

Proposition 6.18 *The family of synchronous rational relations forms an effective Boolean algebra.*

Proof. Let $\mathcal{A} = \langle\, Q, A \times B, E, \{i\}, T \,\rangle$ be a complete deterministic letter-to-letter automaton which realises a synchronous rational relation R. The relation $\overline{R}$, the complement of R – that is, the relation whose graph is the complement of the graph of R in $A^* \times B^*$ – is realised, as in Corollary 6.5, by the automaton $\mathcal{A}' = \langle\, Q, A \times B, E, \{i\}, T' \,\rangle$ where, for each p in Q, T'_p is the complement of T_p in $(A^* \times 1_{B^*}) \cup (1_{A^*} \times B^*)$. If T_p is in $\mathrm{Diff}_{\mathsf{rat}}$, then so is T'_p since $\mathrm{Rat}\,A^*$ and $\mathrm{Rat}\,B^*$ are closed under complement.

The legitimacy of these two constructions comes, as in the aforementioned previous results, from the uniqueness of the factorisation of the elements of $A^* \times B^*$ in the product $(A \times B)^*\,[(A^* \times 1_{B^*}) \cup (1_{A^*} \times B^*)]$. *cf. § 6.1.1, p. 596*

It is easy to decide whether a relation realised by an automaton is empty. Two relations are equivalent if their symmetric difference is empty. ■

We can continue the analogy between letter-to-letter automata with final functions and classic automata over a free monoid.

Proposition 6.19 *Every letter-to-letter automaton with a final function is equivalent to a minimal deterministic letter-to-letter automaton (unique up to isomorphism) which can be effectively computed.*

Proof. We know that we can always suppose that a letter-to-letter automaton with a final function is deterministic. The minimisation process for deterministic letter-to-letter automata will follow step for step that of the minimisation of deterministic automata over a free monoid. We content ourselves with a rough sketch.

(i) Let R be an arbitrary subset of $A^* \times B^*$. For all u in $(A \times B)^*$, we define the set of *continuations* of u in R:

$$\mathrm{Cont}_R(u) = \{w \in (A^* \times 1_{B^*}) \cup (1_{A^*} \times B^*) \mid u\,w \in R\} \ .$$

The relation R thus defines an equivalence relation ρ_R on $(A\times B)^*$ by

$$u \simeq v \bmod \rho_R \quad \Longrightarrow \quad \forall t \in (A\times B)^* \qquad \mathrm{Cont}_R(ut) = \mathrm{Cont}_R(vt) \ .$$

The continuations of a pair u in R will play the role of right contexts of a word in a language. We should note that if in the latter case, the word, the right contexts and the language are all taken from the same set, a free monoid, in the case that now interests us, u is in $(A\times B)^*$, the continuations are in $(A^*\times 1_{B^*}) \cup (1_{A^*}\times B^*)$ and R is in $A^*\times B^*$.

(ii) We verify easily that ρ_R is a right regular equivalence of $(A\times B)^*$, and that it is coarser than the equivalence relation defined on $(A\times B)^*$ by any deterministic letter-to-letter automaton with final function that realises R.

The equivalence ρ_R defines an automaton $\mathcal{A}_R = \langle Q_R, A\times B, E_R, \{[1]_{\rho_R}\}, T_R \rangle$ where Q_R is the quotient of $(A\times B)^*$ by ρ_R,

$$E_R = \{\big([u]_{\rho_R}, (a,b), [u(a,b)]_{\rho_R}\big) \mid [u]_{\rho_R} \in Q_R\,,\ (a,b) \in A\times B\} \ ,$$

and $T_R([u]_{\rho_R}) = \mathrm{Cont}_R(u)$. The automaton $\mathcal{A}_R$ is the *minimal* letter-to-letter automaton of R. It is finite and the final function takes values in $\mathrm{Diff}_{\mathsf{rat}}$ if and only if R is a synchronous rational relation.

(iii) The minimal automaton $\mathcal{A}_R$ is effectively computable from any deterministic letter-to-letter automaton $\mathcal{A} = \langle Q, A\times B, E, \{i\}, T \rangle$ that realises R, by a procedure which is exactly that of the computation of Nerode's equivalence. Let $\mathcal{P}_0$, $\mathcal{P}_1$, ... be the sequence of partitions of Q defined by

cf. § I.3.3.3, p. 114

- $\forall p, q \in Q \quad p \simeq q \bmod \mathcal{P}_0$ if and only if $T(p) = T(q)$;
- $\mathcal{P}_{i+1}$ is the refinement of $\mathcal{P}_i$ obtained by separating p and q (congruent modulo $\mathcal{P}_i$) if there exists (a,b) in $A\times B$ such that $T\big(p\cdot(a,b)\big) \neq T\big(q\cdot(a,b)\big)$.

The length of such a sequence of partitions is bounded by $\|Q\|$, and if $\mathcal{P}_{i+1} = \mathcal{P}_i$ then $\mathcal{P}_{i+n} = \mathcal{P}_i$ for all n. The minimal automaton $\mathcal{A}_R$ is equal to the quotient of $\mathcal{A}$ by the finest equivalence $\mathcal{P}_i$. ■

6.3.3 Geography of $\mathrm{Rat}\, A^*\times B^*$ II

cf. Sec. 5.4, p. 588

We can enrich and refine the description that we gave in the previous section of the family of rational relations by incorporating that of synchronous relations.

Theorem 6.4 $\mathrm{Rec}\,(A^*\times B^*) \subsetneq \mathrm{Syn}\,(A^*\times B^*) \subsetneq \mathrm{DRat}\,(A^*\times B^*) \subsetneq \mathrm{Rat}\,(A^*\times B^*)$.

The inclusions are strict, whatever the alphabets A and B: the identity belongs to $\mathrm{Syn}\, A^*\times B^* \setminus \mathrm{Rec}\, A^*\times B^*$, and the relation $R_2 = (a, aa)^*$, for example, belongs to $\mathrm{DRat}\, A^*\times B^* \setminus \mathrm{Syn}\, A^*\times B^*$. The first inclusion was proved by Proposition 6.14, the second is left as an exercise and the third was seen in Section 5.

cf. Cor. 5.13, p. 586

Remark 6.5 The inclusions of Theorem 6.4 extend to n-ary relations and we have

$$\mathrm{Rec}_n \subsetneq \mathrm{Syn}_n \subsetneq \mathrm{DRat}_n \subsetneq \mathrm{Rat}_n \ . \qquad \square$$

There are no decidability problems inside Syn $(A^* \times B^*)$ itself, since it is an effective Boolean algebra. As to relative decidability, the negative result below follows on from Corollary 4.2 and Theorem 5.1, and is proved in the same way as the latter. *cf. p. 570 and p. 589*

Proposition 6.20 Syn $A^* \times B^*$ *is undecidable in* Rat $A^* \times B^*$*: that is, it is not decidable whether a rational relation is synchronous.* ■

The preceding results about the place of synchronous relations in the family of general rational relations actually raise more questions than they answer.[28]

Problem 6.1 *Is* Rec $A^* \times B^*$ *decidable in* Syn $A^* \times B^*$*?*

The recognisable relations and bld-relations are left *and* right synchronous.

Problem 6.2 *Can we decide whether a left synchronous relation is right synchronous? Can we characterise the relations that are both left and right synchronous?*

It is not hard to convince oneself that a recognisable relation is not only synchronous (Proposition 6.14), i.e. realised by an automaton over $A \times B$ with a final function in $\mathrm{Diff}_{\mathsf{rat}}$, which we can call synchronised *with a synchronisation ratio* equal to 1, but also synchronised with a synchronisation ratio equal to r, for any rational number r; that is, realised by an automaton over $A^n \times B^m$ (with $r = n/m$) with a final function in $\mathrm{Diff}_{\mathsf{rat}}$.

Problem 6.3 *Is a rational relation which is synchronous with any given synchronisation ratio recognisable? What can we say if it is synchronous with two distinct synchronisation ratios?*

Exercises

6.15 Verify that letter-to-letter automata with final functions in $\mathrm{Diff}_{\mathsf{rat}}$ have exactly the same behaviour as 2-tape letter-to-letter automata in which we arrange that the shorter word is padded on the right by a suitable number of occurrences of a fixed symbol (that does not belong to the initial alphabet) up to the length of the word on the other tape.[29]

6.16 Verify that the assertion of Remark 6.3: $R_1 = (a,a)^*(1,b)^*$ is a left synchronous rational relation but not a right synchronous rational relation.

---

[28] *Note added in translation:* The French edition also raised interest in these problems. Ch. Choffrut and O. Carton have answered two of the questions stated there: Problem 6.1 is solved positively in [52] and Problem 6.3 is solved positively in [51] under the stronger form – that is, a rational relation which is synchronous with two distinct synchronisation ratios is recognisable – giving the result almost a taste of Cobham's Theorem (*cf.* Note I.72, p. 186). To my knowledge, Problem 6.2 is still open.

As I decided that the English edition should be a translation of the French one, and not a new edition, I have left the statements as 'problems'; in a future edition, they will become theorems.

[29] This is the version of synchronised automata used in [84] and by more logic-oriented authors (see the bibliographic notes).

• 6.17 (a) Describe the synchronised transducer which realises the recognisable relation $K \times L$ starting with the automata $\mathcal{A}$ and $\mathcal{B}$ over A^* and B^* that recognise K and L respectively.

(b) Construct this automaton in the case where $A = B = \{a, b\}$ and where

$$K = \{f \mid |f|_b \equiv 1 \bmod 2\} \quad \text{and} \quad L = \{g \mid |g|_a \equiv 2 \bmod 3\} \ .$$

6.18 Write out the proof of Proposition 6.20 based on that of Theorem 5.1.

• 6.19 Prove:

Proposition 6.21 *A synchronous rational relation is a deterministic rational relation.*

6.20 Verify that a recognisable relation is synchronous with a synchronisation ratio equal to r, for any rational number r.

7 Malcev–Neumann series

The purpose of this section is to present a proof of:

Theorem 4.5 *The semiring* $\mathbb{N}\mathrm{Rat}\, B^*$ *is a sub-semiring of a skew field and so is, more generally, the semiring* $\mathbb{N}\mathrm{Rat}\,(A_2^* \times \cdots \times A_k^*)$ *.*

cf. Sec. 4.3, p. 574

We have already said that this result is based on two theorems of algebra, whose statement and the definitions they require are repeated below:

Definition 4.1 *A total order relation* $\leqslant$ *on a group* G *makes* G *an* ordered group *if it is compatible, left and right, with multiplication in* G*; that is, if the following holds:*

$$\forall a, b, c \in G \qquad a \leqslant b \quad \Longrightarrow \quad a\,c \leqslant b\,c \quad \textit{and} \quad c\,a \leqslant c\,b \ . \qquad \square$$

Theorem 4.6 [**Birkhoff–Tarski–Neumann–Iwazawa**] *A finite direct product of free groups is an ordered group.*

Definition 4.2 [**Hahn–Malcev–Neumann**] *Let* $\mathbb{K}$ *be a semiring and* G *an ordered group. We write*[30] $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ *for the set of series on* G *with coefficients in* $\mathbb{K}$ *whose support is a well ordered subset of* G*.* $\square$

If there is no ambiguity we call $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ simply the set *of series on* G *with coefficients in* $\mathbb{K}$ or the set of $\mathbb{K}$-series on G, in contrast with the general definition, because they are the only series which we can reasonably consider on G if $\mathbb{K}$ is not a continuous semiring.

cf. Sec. III.1.1 and Note III.11, p. 381

Theorem 4.7 [**Malcev–Neumann**] *If* $\mathbb{K}$ *is a skew field and* G *an ordered group, then* $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ *is a skew field.*

[30] The notation $\mathbb{K}_{\mathsf{m}}\langle\langle G\rangle\rangle$ is used by some authors, in homage to Malcev.

7.1 Order on the free group

We seek here to attain and reconcile two distinct and almost contradictory objectives: on one hand to prove in the most elementary possible way (that is, using the simplest notions and most direct reasoning) that the free group can be ordered, and on the other to describe and understand the nature of the order with which we shall equip it.[31]

We start with some elementary properties on ordered groups, then we introduce the Magnus representation of the free group, which is vital to the achievement of our two objectives. We then prove Theorem 4.6 by making a detour via ordered rings, a detour which turns out to be a fiendish shortcut. In the fourth paragraph, we return to the description of the order we have thus constructed, which brings into play the central descending series of the free group, which we shall also define with the aid of the Magnus representation.

7.1.1 On ordered groups

We begin by showing that the definition of a total order on a group is equivalent to finding sets of *positive* and *negative* elements; that is, of elements respectively greater than and less than the neutral element of the group.

Proposition 7.1 *A group G is ordered if and only if there exist two subsets P and N of G which satisfy the following*[32] *conditions:*
(i) $G = P \oplus N \oplus 1_G$; (ii) $P^2 \subseteq P$ *and* $N^2 \subseteq N$; (iii) $\forall t \in G \quad t^{-1} P t \subseteq P$.

Proof. Suppose that G is ordered; we take

$$P = \{g \in G \mid 1_G \lneq g\} \qquad \text{and} \qquad N = \{g \in G \mid g \lneq 1_G\} \ .$$

This choice implies (i). The definition of the order on G implies

$$1_G \lneq g \quad \text{and} \quad 1_G \lneq h \quad \Longrightarrow \quad 1_G \lneq g\,h :$$

that is, (ii); and

$$1_G \lneq g \quad \Longrightarrow \quad t \lneq g\,t \quad \Longrightarrow \quad 1_G \lneq t^{-1} g\,t :$$

that is, (iii).

Suppose conversely that there exist P and N which verify (i), (ii) and (iii); we define a relation $\lneq$ on G by

$$g \lneq h \quad \text{if and only if} \quad g^{-1} h \in P \ ;$$

Hypothesis (i) implies that $\lneq$ is total and anti-symmetric; (ii) implies the transitivity of $\lneq$ and (iii) ensures its compatibility with multiplication in the group. ■

[31] The result of this contradiction is long-windedness.

[32] The $\oplus$ denotes here the *disjoint* union.

Proposition 7.2 *A subgroup of an ordered group is ordered.* ■

Proposition 7.3 *If G and H are ordered groups, the group $G \times H$ is ordered.*

cf. Sec. II.7.2, p. 326

Proof. We define the *lexicographic order*[33] on $G \times H$ by

$$(g,h) \lneq (g',h') \qquad \text{if and only if} \qquad \begin{cases} g \lneq g' \\ \quad \text{or} \\ g = g' \quad \text{and} \quad h \lneq h' \ . \end{cases}$$

We verify immediately that this order makes $G \times H$ an ordered group. ■

Remark 7.1 The result is actually much more general and extends to infinite products (on a set of well-ordered indices). □

Corollary 7.4 *The free commutative group $\mathbb{Z}^k$ is ordered.* ■

7.1.2 Representation of the free group

Let $\mathbb{Z}\langle A^* \rangle$ be the algebra of polynomials on A^* with coefficients in $\mathbb{Z}$, and $\mathbb{Z}\langle\langle A^* \rangle\rangle$ the algebra of series on A^* with coefficients in $\mathbb{Z}$ (also called the (large) algebra of the free monoid). The representation of the free group $\mathrm{F}(A)$ described in the following theorem is known as the *Magnus representation.*

Theorem 7.1 *The morphism $\mu\colon A^* \to \mathbb{Z}\langle A^* \rangle$ defined by $a\mu = 1 + a$ extends to an* injective morphism

$$\mu\colon \mathrm{F}(A) \longrightarrow \mathbb{Z}\langle\langle A^* \rangle\rangle \ ,$$

by taking $a^{-1}\mu = 1 - a + a^2 - \cdots + (-1)^n a^n + \cdots$.

Proof. Since $(1+a)(1-a+a^2-\cdots) = (1-a+a^2-\cdots)(1+a) = 1$, the choice of $a^{-1}\mu$ indeed makes μ a morphism. It suffices to show that $1_{\mathbb{Z}\langle\langle A^* \rangle\rangle}\mu^{-1} = 1_{\mathrm{F}(A)}$ and for that that $w\mu \neq 1$ if $w \neq 1_{\mathrm{F}(A)}$. We identify $\mathrm{F}(A)$ with the set of reduced words: an element w in $\mathrm{F}(A)$ other than $1_{\mathrm{F}(A)}$ is written (uniquely) in the form

$$w = a_1^{i_1} a_2^{i_2} \cdots a_n^{i_n} \ ,$$

with a_j in A, i_j in $\mathbb{Z} \setminus 0$, and $a_j \neq a_{j+1}$ for all $j \in [n]$. We shall show that the support of $w\mu$ contains $a_1 a_2 \cdots a_n$. We write

$$a_j^{i_j}\mu = 1 + i_j\, a_j + {a_j}^2\, S_{i_j}(a_j) \ , \tag{7.1}$$

where $S_{i_j}(a_j)$ is a series of $\mathbb{Z}\langle\langle A^* \rangle\rangle$. By assumption, the word $a_1 a_2 \cdots a_n$ contains no factor of the form ${a_j}^2$ and the only way to obtain $a_1 a_2 \cdots a_n$ in the product

$$w\mu = \big(1 + i_1\, a_1 + {a_1}^2\, S_{i_1}(a_1)\big) \cdots \big(1 + i_n\, a_n + {a_n}^2\, S_{i_n}(a_n)\big)$$

is to form the product of the terms $i_j\, a_{i_j}$, and its coefficient is the product of the i_j, all different from 0. ■

[33] N.B. this is not the *product order* on $G \times H$: that is, the order obtained as the product of the orders on G and H.

7.1.3 A detour via ordered rings

Definition 7.1 A total order $\leqslant$ on a ring R makes R an *ordered ring* if it is compatible with addition in R and left and right multiplication by the 'positive' elements of R; that is, if the following holds:

$$\forall a, b, c \in R\,,\ \forall d \in R\,,\ 0_R \leqslant d \qquad a \leqslant b \quad \Longrightarrow \quad \begin{cases} a + c \leqslant b + c\ , \\ a\,d \leqslant b\,d \quad \text{and} \quad d\,a \leqslant d\,b\ . \end{cases} \qquad \square$$

We then have the equivalent of Proposition 7.1.

Proposition 7.5 *A ring R is ordered if and only if there exists a subset C of R which satisfies the following conditions:*

(i) $R = C \oplus 0_R \oplus -C$; (ii) *C is closed for $+$ and $\times$.*

Proof. Suppose that R is ordered; we take

$$C = \{x \in R \mid 0_R \lneqslant x\}\ .$$

The definition of the order on R implies directly

$$0_R \lneqslant x \quad \text{and} \quad 0_R \lneqslant y \quad \Longrightarrow \quad 0_R \lneqslant x + y \quad \text{and} \quad 0_R \lneqslant x\,y\ .$$

Conversely, suppose that the subset C verifies (i) and (ii); we define the ordering relation $\lneqslant$ on R by

$$x \lneqslant y \qquad \text{if and only if} \qquad y - x \in C\ .$$

Condition (i) implies that $\lneqslant$ is total and antisymmetric; (ii) implies the transitivity of $\lneqslant$ as shown by the following:

$$y - x \in C \quad \text{and} \quad z - y \in C \quad \Longrightarrow \quad z - x = (z - y) + (y - x) \in C\ ,$$

and, together with the distributivity of $\times$ over $+$, it also implies compatibility with multiplication in R. ■

Let $\mathsf{U}_+(R)$ be the intersection of the group of units of R with C:

$$\mathsf{U}_+(R) = \{x \in R \mid 0_R \lneqslant x \quad \text{and} \quad x \text{ invertible}\}\ .$$

It is a simple exercise to verify:

Lemma 7.6 *Let R be an ordered ring. The subset $\mathsf{U}_+(R)$ is a group and the restriction of the order on R to $\mathsf{U}_+(R)$ makes $\mathsf{U}_+(R)$ an ordered group.* ■

Proof of Theorem 4.6. We choose a total order on A^* which is a well ordering and is compatible with multiplication; that is,

$$a \leqslant b \quad \text{and} \quad c \leqslant d \quad \Longrightarrow \quad a\,c \leqslant b\,d\ ,$$

for example, the radix order. We then order $\mathbb{Z}\langle\langle A^*\rangle\rangle$ by choosing as the positive cone the set of series such that the coefficient of the smallest monomial with a non-zero coefficient is positive.

The group $\mathsf{U}_+(\mathbb{Z}\langle\langle A^*\rangle\rangle)$ is thus the set of series whose constant term is equal to 1. The free group $\mathrm{F}(A)$ is a subgroup of $\mathsf{U}_+(\mathbb{Z}\langle\langle A^*\rangle\rangle)$, ordered by Proposition 7.2. A finite direct product of free groups is ordered by Proposition 7.3. ■

7.1.4 Order on the free group

We shall now describe more explicitly this order on the free group whose existence we have just proved. To this end we shall use the Magnus representation to define a sequence of normal subgroups of the free group. For all u and v in $\mathrm{F}(A)$, we write $[u,v] = u^{-1}v^{-1}uv$ for the *commutator* of u and v and, if D and E are subsets of $\mathrm{F}(A)$,

$$[D,E] = \{[u,v] \mid u \in D\,, \quad v \in E\} \ .$$

Theorem 7.2 *Let D_n be the subset of $\mathrm{F}(A)$ defined by*

$$D_n = \{w \in \mathrm{F}(A) \mid w\mu = 1 + S_w \quad \textit{and} \quad \mathbf{v}(S_w) \geqslant n\} \ ,$$

cf. Rem. III.1.5, p. 387 *where $\mathbf{v}(S_w)$ is the valuation of the series S_w. Then, for all n in $\mathbb{N}_*$,*

(i) *D_n is a normal subgroup of $\mathrm{F}(A)$ ($D_1 = \mathrm{F}(A)$);*

(ii) $\displaystyle\bigcap_{n\in\mathbb{N}} D_n = 1_{\mathrm{F}(A)}$ *;*

(iii) *D_n/D_{n+1} is a finitely generated free commutative group; more precisely, for $n = 1$, we have: $D_1/D_2 = \mathbb{Z}^{\|A\|}$;*

(iv) $[D_n,D_1] \subseteq D_{n+1}$.

Proof. (i) Since $\mathbf{v}(st) = \mathbf{v}(s) + \mathbf{v}(t)$ for all s and t in $\mathbb{Z}\langle\langle A^*\rangle\rangle$, D_n is a subgroup of $\mathrm{F}(A)$, indeed, a normal subgroup, for all n in $\mathbb{N}_*$.

(ii) By definition, $1_{\mathrm{F}(A)}$ belongs to the intersection of all the D_n and we saw with Theorem 7.1 that $w\mu$ is different from 1 for all w in $\mathrm{F}(A)$ other than $1_{\mathrm{F}(A)}$.

(iii) For all w in D_n, we write $w\mu$ in the form

$$w\mu = 1 + Q_w + S'_w \ , \tag{7.2}$$

where Q_w is a homogeneous polynomial of degree n of $\mathbb{Z}\langle A^*\rangle$ and $\mathbf{v}(S'_w)$ is greater than or equal to $n+1$. If w and w' are in D_n, we have

$$(w\,w')\mu = 1 + Q_w + Q_{w'} + S'_{w\,w'}$$

and the map $\pi_n\colon w \mapsto Q_w$ is a morphism from D_n to the *additive group* H_n of homogeneous polynomials of degree n over $\mathbb{Z}\langle A^*\rangle$. The kernel of π_n is D_{n+1} and D_n/D_{n+1} is isomorphic to $(D_n)\pi_n$, a subgroup of H_n.

Of course H_1, the group of homogeneous polynomials *of degree* 1, is isomorphic to $\mathbb{Z}^{\|A\|}$ and it is again the proof of Theorem 7.1, Equation (7.1), which ensures the surjectivity of π_1 onto H_1.

(iv) Let w be in D_n and t in $D_1 = \mathrm{F}(A)$; we write

$$t\mu = 1 + U_t + T_t \ ,$$

where U_t is a polynomial whose monomials are of degree between 1 and n (inclusive) and T_t is a series with valuation greater than or equal to $n+1$. By writing $w\mu$ in the form of (7.2), we obtain

$$(w\,t)\mu = 1 + U_t + Q_w + T_{w\,t} \qquad \text{and} \qquad (t\,w)\mu = 1 + U_t + Q_w + T_{t\,w} \ .$$

Thus, $(w\,t)\mu$ and $(t\,w)\mu$ are two series that are equal modulo A^{n+1}, hence so are $((w\,t)\mu)^{-1}$ and $((t\,w)\mu)^{-1}$. It follows that $((t\,w)\mu)^{-1}(w\,t)\mu = ([w,t])\mu$ is equal to 1 modulo A^{n+1}; that is, $[w,t]$ belongs to D_{n+1}. ■

Remark 7.2 We have not really proved (iii), but only that D_n/D_{n+1} is a *subgroup* of a finitely generated commutative group. It is a classic result in algebra that such a subgroup is itself a finitely generated free commutative group. Formally, though, we do not need to prove more than we already have in order to prove Theorem 4.6 (or Theorem 7.3). *cf. Exer. 7.4*

Furthermore, it is standard, but rather more difficult to prove, that we have not just the inclusion of $[D_n, D_1]$ in D_{n+1} but also equality of the two, and thus that D_n is *the nth term of the central descending series* of $\mathrm{F}(A)$. We shall not need this beautiful result either. □

Theorem 7.3 *Let A be a finite alphabet of cardinal k. For all group orders on $\mathbb{Z}^k$, we can choose on $\mathrm{F}(A)$ an order which makes $\mathrm{F}(A)$ an ordered group such that the canonical morphism $\gamma\colon \mathrm{F}(A) \to \mathbb{Z}^k$ is a morphism of ordered groups.*

Proof. We use the notation of Theorem 7.2 (and sketch the construction in Figure 7.1); for each n, H_n is a free commutative group (of rank k^n), hence ordered, and $E_n = (D_n)\pi_n$ is an ordered group (Proposition 7.2). Let $E_n = X_n \oplus Y_n \oplus 1_{E_n}$ be the partition constructed on this order (Proposition 7.1). Set $P_n = (X_n){\pi_n}^{-1}$ and $N_n = (Y_n){\pi_n}^{-1}$ (we have $D_{n+1} = (1_{E_n}){\pi_n}^{-1}$) and

$$P = \bigcup_{n\in\mathbb{N}_*} P_n \qquad \text{and} \qquad N = \bigcup_{n\in\mathbb{N}_*} N_n \ .$$

It remains to show that P and N satisfy the three conditions of Proposition 7.1 and hence define an order on $\mathrm{F}(A)$.

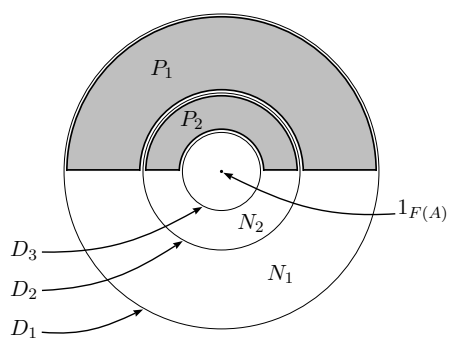


Figure 7.1: Construction of an order on $\mathrm{F}(A)$

From Theorem 7.2 (ii), we deduce that $P \cup N = \mathrm{F}(A) \setminus 1_{\mathrm{F}(A)}$.

Let u and v be in P, u be in D_n, hence in P_n, and v be in D_m, hence in P_m. If $n = m$, $(u\,v)\pi_n = u\pi_n\, v\pi_n$ and $u\,v$ is in P_n by the property of X_n. If $n \not\leqslant m$, then $u\,v$ is equivalent to u modulo D_{n+1} hence belongs to P_n.

Let u be in P, hence in P_n for some n, and t be in D_1. Since $t^{-1}ut = u[u,t]$, and $[u,t]$ is in D_{n+1} (Theorem 7.2 (iv)), u and $t^{-1}ut$ are equivalent modulo D_{n+1} and $t^{-1}ut$ is in P_n, hence in P.

By definition, the canonical morphism $\gamma\colon \mathrm{F}(A) \to \mathbb{Z}^k$ is equal to π_1, and the order on $D_1 \setminus D_2$, hence on $\mathrm{F}(A)$, is constructed in such a way that π_1 is a morphism for the order. ■

Exercises

7.1 Verify Proposition 7.2 and that the lexicographic order indeed makes the product of two ordered groups an ordered group.

7.2 Show that we can order $\mathbb{Z}^n$ other than by the lexicographic order, by projection on to a line of 'completely' irrational gradient.

7.3 Verify Lemma 7.6.

7.4 Show that a subgroup of a free commutative group of rank k (that is, generated by k elements) is a free commutative group of rank $r \leqslant k$.

7.2 Series on an ordered group

We now come to the proof of Theorem 4.7. It relies on the notion of a *well partial ordering* and on Higman's Theorem, which we saw in Section II.5. Hence, we freely use the definitions and results proved in that section for *well quasi-orderings* and which hold identically for well partial orderings. Recall in particular:

Th. II.5.2, p. 295

cf. p. 293

Theorem II.5.2 [**Higman**] *If X is a set equipped with a well partial ordering, then division is a well partial ordering on* $\mathrm{V}(X)$.

Properties II.5.2 *Let E be a set equipped with a well partial ordering $\leqslant$.*

(i) *The trace of $\leqslant$ on every subset of E is a well partial ordering.*

(ii) *Every image of E under a morphism is equipped with a well partial ordering.*

7.2.1 The semiring $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$

cf. Def. II.5.3, p. 298

Recall that if E and F are partially ordered sets, a map $\alpha\colon E \to F$ is a *morphism* (we understand 'for the order') if

$$a \leqslant b \Rightarrow a\alpha \leqslant b\alpha \ .$$

The morphism α is called *strict* if, in addition,

$$a \lneqslant b \Rightarrow a\alpha \lneqslant b\alpha \ .$$

In other words, the morphism α is strict if $a \leqslant b$ and $a\alpha = b\alpha$ imply $a = b$.

Lemma 7.7 *Let $\alpha\colon E \to F$ be a strict morphism between two ordered sets. If the order on E is a well partial ordering, then, for all f in F, $f\alpha^{-1}$ is a finite set.*

Proof. By definition of a strict morphism, the elements of $f\alpha^{-1}$ are *pairwise incomparable*, and hence finite in number. ■

cf. Th. II.5.1 (vi), p. 294

We now know enough to show that the series on an ordered group form a semiring. We first prove a small lemma.

Lemma 7.8 *Let G be an ordered group. The map $\pi\colon G \times G \to G$ defined by $(g,h)\pi = g\,h$ is a strict morphism (of (partially) ordered sets).*[34]

Proof. That π is a morphism is a direct consequence of the definitions: $(g,h) \leqslant (g',h')$ implies $g \leqslant g'$ and $h \leqslant h'$, from which $g\,g' \leqslant h\,h'$.

Since $(g,h) \leqslant (g',h) \leqslant (g',h')$, $g\,h = g'\,h'$ implies $g\,h = g'\,h = g'\,h'$ and hence $g = g'$ and $h = h'$, since G is a group. ■

Proposition 7.9 *If $\mathbb{K}$ is a semiring (resp. a ring) and G a (totally) ordered group, $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ is a semiring (resp. a ring).*

Proof. We need to show that $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ is closed under product. Therefore, let s and t be series on G (with coefficients in $\mathbb{K}$) whose supports, respectively S and T, are two well ordered subsets of G. For each $g\,h$ in $S\,T$, the set $(S \times T) \cap (g\,h)\pi^{-1}$ is finite (Lemma 7.7) and the Cauchy product of s and t is well defined. The support of $s\,t$ is contained in $(S \times T)\pi$ and is hence a well ordered set. ■

cf. Propty II.5.2 (i) and (ii), p. 298

7.2.2 Ordered semigroups

The last step is to define *ordered semigroups*, which will be an essential ingredient of the proof of Theorem 4.7.

Definition 7.2 A semigroup P is called *ordered* if there exists a *total* order $\leqslant$ on P such that, for all r, s and t in P we have

(i) $r \leqslant s \implies r\,t \leqslant s\,t$ and $t\,r \leqslant t\,s$; (ii) $r \lneqslant r^2$. □

Condition (i) expresses the regularity of the order relation for multiplication. It entails, as for a congruence, that

$$r \leqslant r' \quad \text{and} \quad s \leqslant s' \quad \Longrightarrow \quad r\,s \leqslant r'\,s \leqslant r'\,s' \ .$$

From condition (ii) we deduce:

Lemma 7.10 *In an ordered* simplifiable *semigroup a product is strictly greater than each of its factors.*

Proof. Let r and s be elements of P, an ordered semigroup, such that $r \lneqslant s$. From (ii) and (i), we obtain $r \lneqslant r^2 \leqslant r\,s$.

[34] But not at all a morphism of groups.

Furthermore, by multiplying (ii) on the right by s, we have $r\,s \leqslant r^2\,s$. If $r\,s \leqslant s$, it follows, by multiplying on the left by r, that $r^2\,s \leqslant r\,s$ from which $r\,s = r^2\,s$ or $r = r^2$ since P is simplifiable; this is a contradiction, hence $s \not\leqslant r\,s$. ■

Remark 7.3 Definition 7.2 is inconsistent in the sense that a group G is a semigroup but at the same time an order that makes G an ordered group does not make it an ordered semigroup, because of condition (ii). This does not matter: it is in fact simply taking liberties with language, but it does require that we take care to give separate statements for analogous properties of ordered groups and ordered semigroups (for example, Lemma 7.8 and Lemma 7.11).

On the other hand, if G is an ordered group, the trace of the order of G on the set P of 'positive' elements of G certainly makes P an ordered semigroup, and this is what we have in mind. □

With the same proof as that of Lemma 7.8 we have:

Lemma 7.11 *Let P be an ordered semigroup. The map $\pi\colon P \times P \to P$ defined by $(g,h)\pi = g\,h$ is a morphism (of (partially) ordered sets). If P is* simplifiable *the morphism π is strict.* ■

Lemma 7.12 *Let P be a simplifiable ordered semigroup. The canonical morphism (of semigroups) $\alpha\colon \mathrm{V}(P) \to P$ is a strict morphism (of ordered sets).*

Proof. Suppose $(s_1, s_2, \ldots, s_p) \trianglelefteq (t_1, t_2, \ldots, t_n)$; there then exists an ascending sequence of indices $i_1, i_2, \ldots, i_p$ such that, for all j, $s_j \leqslant t_{i_j}$. Thus

$$(s_1, s_2, \ldots, s_p)\alpha = s_1 s_2 \cdots s_p \leqslant t_{i_1} t_{i_2} \cdots t_{i_p} \quad .$$

Lemma 7.10 used $p+1$ times gives

$$t_{i_1} t_{i_2} \cdots t_{i_p} \leqslant t_1 t_2 \cdots t_n = (t_1, t_2, \ldots, t_n)\alpha \quad .$$

Lemma 7.11 and Lemma 7.10 ensure that we cannot have the equality

$$(s_1, s_2, \ldots, s_p)\alpha = (t_1, t_2, \ldots, t_n)\alpha$$

unless $p = n$ and $s_j = t_j$ for all j, hence α is strict. ■

7.2.3 The field $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$

From the above results we now easily deduce the crucial result for the embedding of $\mathrm{F}(A)$ in a field.

Proposition 7.13 [**Neumann**] *Let P be an ordered simplifiable semigroup and S a well ordered subset of P. Then*

(i) *The semigroup generated by S, S^+, is a well ordered subset of P.*

(ii) *The family $\{S^n\}_{n\in\mathbb{N}}$ is locally finite: every p in S^+ belongs to a finite number of the S^n.*

Proof [Higman]. We write, as in Lemma 7.12, $\alpha\colon \mathrm{V}(P) \to P$ for the canonical morphism.

(i) Since S is well ordered, $\mathrm{V}(S)$ is a well (partially) ordered set by Higman's Theorem; the semigroup $S^+ = (\mathrm{V}(S))\alpha$ is well ordered by Lemma 7.12 and Properties II.5.2.

(ii) For all p in S^+, $(p)\alpha^{-1}$ is finite by Lemma 7.7 and since, under our assumptions, the morphism α is *strict. A fortiori*, $(p)\alpha^{-1} \cap \mathrm{V}(S)$ is finite, and this is exactly the property sought. ∎

Proof of Theorem 4.7. We already know that $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ is a ring; to prove that it is a skew field, we must show not only that every non-zero element is invertible but also that this inverse belongs to $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$. Let, as in Proposition 7.1, P be the set of 'positive' elements of G; as we have already noted, the trace on P of the order of G makes P an ordered semigroup (and P is *simplifiable* since it is a sub-semigroup of a group).

Let t be an arbitrary non-zero element of $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$. Let $T = \mathsf{supp}\, t$ be the (non-empty) support of t, and $g_t = \min T$ the smallest element of T: this exists since, by definition of $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$, T is a well ordered set. We can then write

$$t = \langle t, g_t\rangle g_t\,(1-s): \quad \text{that is, written otherwise,} \qquad s = 1_G - \frac{1}{\langle t, g_t\rangle} {g_t}^{-1}\, t \ ,$$

and the series s thus defined satisfies

$$\langle s, 1_G\rangle = 0_{\mathbb{K}} \qquad \text{and} \qquad \langle s, g\rangle = \frac{-1}{\langle t, g_t\rangle}\langle t, g_t\, g\rangle \ ,$$

for all g in G. Hence we deduce that $\langle s, g\rangle \neq 0_{\mathbb{K}}$ implies that $g_t \lneq g_t\, g$ – that is, $1_G \lneq g$ – and the support $S = g_t^{-1}T \setminus 1_G$ of s is hence contained in P.

The family $\{s^n\}_{n\in\mathbb{N}}$ is locally finite by Proposition 7.13 (ii); it is thus summable and the support of its sum s^* is a subset of S^* hence a well ordered subset of G: then s^* belongs to $\mathbb{K}_{\mathsf{wo}}\langle\langle G\rangle\rangle$ and so does

$$t^{-1} = s^* {g_t}^{-1} \frac{-1}{\langle t, g_t\rangle} \ . \qquad ∎$$

7.2.4 A last inclusion

The theorem that we want to prove requires a final lemma on well ordered sets. *Th. 4.5, p. 616*

Lemma 7.14 *For each i in a finite set I, let F_i be a well ordered subset of a (totally) ordered set E_i. Then $\prod_{i\in I} F_i$ is a well ordered subset of $\prod_{i\in I} E_i$, ordered lexicographically.* ∎

Note that this lemma is not redundant with respect to Property II.5.2 (recalled above) since the order considered here on $\prod_{i\in I} E_i$ is not the product order but the lexicographic order. We then deduce from Theorem 7.3:

Corollary 7.15 *Let A be a finite alphabet. We can choose an order on* $\mathrm{F}(A)$ *such that the canonical injection makes A^* a well ordered subset of* $\mathrm{F}(A)$.

Proof. Let $k = \|A\|$. Since $\mathbb{N}$ is a well ordered submonoid of $\mathbb{Z}$, we see $\mathbb{N}^k$ is a well ordered submonoid of $\mathbb{Z}^k$. We choose an order on $\mathrm{F}(A)$ such that $\gamma\colon \mathrm{F}(A) \to \mathbb{Z}^k$ is a morphism for the lexicographic order on $\mathbb{Z}^k$.

Let $f_1, f_2, f_3, \ldots$ be a sequence of words in A^*, strictly descending for the order of $\mathrm{F}(A)$. Its image under γ is a descending sequence of $\mathbb{N}^k$, hence eventually stationary. As the inverse image under γ *in* A^* of an arbitrary element of $\mathbb{N}^k$ is finite, the sequence f_i is also eventually stationary; we have a contradiction, and A^* is a well ordered subset of $\mathrm{F}(A)$. ∎

Theorem 4.5 is then only a formality since we have

$$\mathbb{N}\mathrm{Rat}\, A^* \subseteq \mathbb{N}\langle\langle A^* \rangle\rangle \subseteq \mathbb{Q}\langle\langle A^* \rangle\rangle \subseteq \mathbb{Q}_{\mathsf{wo}}\langle\langle F(A) \rangle\rangle \ ,$$

the first three inclusions being obvious and Corollary 7.15 ensuring the last one. ∎

Exercises

7.5 Verify Lemma 7.11.

7.6 Justify our assertion in Theorem 4.7 that S^* is a well ordered subset, since we have only proved the property for S^+ in Proposition 7.13.

7.7 Verify Lemma 7.14.

SOLUTIONS TO THE EXERCISES

Automata, considered to be curiosities of creation,
also figured in the collections of monarchs
alongside marvels of nature.

from the catalogue of the exhibition see bibliographic notes
'L'étrange et le merveilleux en terres d'Islam'

1. Rational relations: an introduction

A and B are two finite alphabets; K and L are two rationals of A^* and B^* respectively. Show that the following relations from A^* to A^*, or from A^* to B^*, are rational.

1.1 ***Operations with constants.***

(a) *The constant relation* γ_L *defined by* $\forall f \in A^* \quad f\gamma_L = L$.

(b) *The relation* ω'_K *'left (or right) product with the rational* K*':* $f \longmapsto K\,f$.

(c) *The relation* ψ'_K *'left (or right) quotient by the rational* K*':* $f \longmapsto K^{-1}f$.

Solution: (a) $\widehat{\gamma_L} = A^* \times L$; (b) $\widehat{\omega'_K} = (1_{A^*} \times K)\,\mathcal{D}^*$ with $\mathcal{D} = \bigcup_{a\in A}(a,a)$; *cf. Exam. 1.2, p. 526*

(c) $\psi'_K = (\omega'_K)^{-1}$.

1.2 ***Factors and subwords.*** (a) *The relation* π *(resp.* σ*) which makes a word in* A^* *correspond to the set of its prefixes, respectively of its suffixes.*

(b) *The relation which makes a word* A^* *correspond to the set of its subwords.*

Solution: (a) $\widehat{\pi} = \mathcal{D}^*(A^* \times 1_{A^*})$ and $\widehat{\sigma} = (A^* \times 1_{A^*})\,\mathcal{D}^*$, with $\mathcal{D}$ as above.

(b) Set $\mathcal{H} = \big(\bigcup_{a\in A}(a, 1_{A^*})\big)$; the graph of the relation is $(\mathcal{D} \cup \mathcal{H})^*$.

1.3 ***Orders: predecessors and successors.***

The alphabet $A = \{a_1, a_2, \ldots, a_n\}$ *is* totally ordered*:* $a_1 < a_2 < \cdots < a_n$.

(a) *Show that the relations that make a word of* A^* *correspond to its predecessor, successor, set of predecessors and set of successors in the lexicographic order are rational.*

(b) *Do the same considering the* radix order *on* A^*, $\sqsubseteq$.

Solution: (a) Set $\mathcal{S} = \bigcup_{a,b\in A,\, a<b} (a,b)$, $\mathcal{U} = A^* \times A^*$, $\mathcal{V} = 1_{A^*} \times A^+$ and $\mathcal{D}$ as above.

The graph of the relation that makes a word f correspond to all the words of which f is a (strict) prefix, is $\mathcal{D}^*\mathcal{V}$. The set of pairs (f,g) such that there exists a decomposition $f = h\,a\,u$ and $g = h\,b\,v$ with $h, u, v \in A^*$, $a, b \in A$ and $a < b$ is described by the expression $\mathcal{D}^*\mathcal{S}\mathcal{U}$. Thus, the definition of $\preccurlyeq$ translates into an expression for the graph of τ_1 , the relation that makes a word f correspond to the set of words which are strictly greater than f:

$$\widehat{\tau_1} = \mathcal{D}^*\mathcal{V} + \mathcal{D}^*\mathcal{S}\mathcal{U} \ .$$

The relation τ_2 which makes a word f correspond to the set of words which are strictly less than it is the inverse of τ_1: *cf. Propty 1.1, p. 526*

$$f\tau_2 = \{g \mid g \not\preccurlyeq f\} = \{g \mid f \in g\tau_1\} = f\tau_1^{-1} \ .$$

Let $a = a_1$ be the smallest letter of A. The word $f\,a$, which is greater than f, is the successor of f; in fact, every word g that is less than $f\,a$ is:

(i) either a prefix of $f\,a$; that is, f, or a prefix of f;

(ii) or has a factorisation $g = h\,x\,u$ and $f\,a$ factorises in turn into $f\,a = h\,y\,v$ with $x < y$. As a is the smallest letter of A we cannot have $h = f$, $y = a$ and $v = 1$; h is thus a prefix of f and $g \preccurlyeq f$.

Therefore all the words less than $f\,a$ are less than f, except f itself; $f\,a$ is indeed the successor of f, and the graph of the successor relation is $\mathcal{D}^*(1,a)$.

The predecessor relation is the inverse of the successor relation: the predecessor of f is the word $f\,a^{-1}$, defined only if the last letter of f is an a: $f = g\,a$ and $f\,a^{-1} = g$.

(b) Set, in addition to the above,

$$\mathcal{E} = A \times A \qquad \text{and} \qquad \mathcal{F} = \bigcup_{i=1}^{i=n-1} (a_i, a_{i+1}) \ .$$

The definition of $\sqsubseteq$ translates into an expression for the graph of σ_1, the relation which makes a word f correspond to the set of words which are greater than f for $\sqsubseteq$:

$$\widehat{\sigma_1} = \mathcal{E}^*\mathcal{V} + \mathcal{D}^*\mathcal{S}\,\mathcal{E}^* \ .$$

The graph of the successor relation for $\sqsubseteq$ is $\mathcal{D}^*\mathcal{F} + (a_n, a_1)^*(1_{A^*}, a_1)$.

cf. Def. I.1.7 and Def. I.1.8, p. 66

1.5 ***Shuffle and infiltration.*** *Show that the operations of* shuffle *and* infiltration, *defined respectively, for all f and g in A^*, by*

$$f \between g = \{h \in A^* \mid \exists I, J \subset [|h|]\,,\ I \cup J = [|h|]\,,\ I \cap J = \emptyset \qquad f = h_I \quad \text{and} \quad g = h_J\}$$

$$f \uparrow g = \{h \in A^* \mid \exists I, J \subset [|h|]\,,\ I \cup J = [|h|] \qquad f = h_I \quad \text{and} \quad g = h_J\}$$

define two rational relations from $A^* \times A^*$ to A^*.

Solution: The graph of the shuffle operation is

$$\left[\left(\bigcup_{a\in A}(a, 1_{A^*}, a)\right) \bigcup \left(\bigcup_{a\in A}(1_{A^*}, a, a)\right)\right]^* \ ,$$

and that of the infiltration operation is

$$\left[\left(\bigcup_{a\in A}(a, 1_{A^*}, a)\right) \bigcup \left(\bigcup_{a\in A}(1_{A^*}, a, a)\right) \bigcup \left(\bigcup_{a\in A}(a, a, a)\right)\right]^* \ .$$

1.6 *Let $\theta_1 \colon \{a\}^* \to \{x\}^*$ be defined by $(a^{2p})\theta_1 = x^{4p}$ and $(a^{2p+1})\theta_1 = x^{2p+1}$; show that θ_1 is a rational relation.*

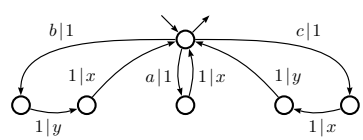
Solution: $\widehat{\theta_1} = (a^2, x^4)^* \cup (a, x)\,(a^2, x^2)^*$.

1.7 *Let $\tau \colon A^* \times A^* \to B^*$ be a rational relation; show by means of an example that the 'diagonal' of τ, $\delta_\tau \colon A^* \to B^*$, defined by $f\delta_\tau = (f, f)\tau$ for all f in A^*, is not necessarily a rational relation.*

Solution: Set $\tau \colon \{a\}^* \times \{a\}^* \to \{a, b\}^*$ defined by $\widehat{\tau} = (a, 1_{A^*}, a)^*(1_{A^*}, a, b)^*$; for all n, $a^n\delta_\tau = a^n b^n$ and $\mathsf{Im}\,\delta_\tau = \{a^n b^n \mid n \in \mathbb{N}\}$.

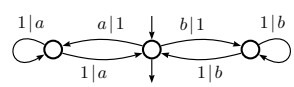
1.8 *Let $\varphi_1 \colon \{a, b, c\}^* \to \{x, y\}^*$ be the morphism defined by $a\varphi_1 = x$, $b\varphi_1 = y\,x$, $c\varphi_1 = x\,y$. Give a transducer in normal form which realises φ_1.*

Solution: The transducer in the adjacent figure.



1.9 *Do the same for the substitution σ_1 defined by $a\sigma_1 = a^+$ and $b\sigma_1 = b^+$.*

Solution: The transducer in the adjacent figure.



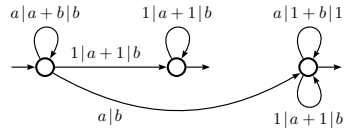
1.10 *Give a transducer which realises:* (a) *the relation 'lexicographic order' (that is, one which relates a word to all those which are greater than it);*

(b) *the relation 'radix order';*

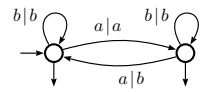
(c) *the relation* α_1 *from* $\{a,b\}^*$ *to itself which replaces every second* a *by* b.

Solution:

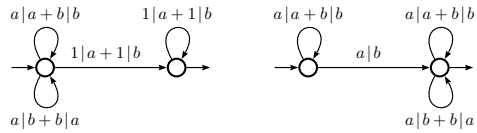
(a)



(c)



(b)



1.12 ***Quotient.*** *Show that the left (and right) quotient of a fixed rational,* $f \longmapsto f^{-1}K$, *for* $f \in A^*$ *and* $K \in \mathrm{Rat}\, A^*$, *is a rational relation from* A^* *to itself.*

Solution: Write $f\varphi_K = f^{-1}K = \{g \mid f\,g \in K\}$; hence

$$\widehat{\varphi_K} = \{(f,g) \mid f\,g \in K\} = K\pi^{-1} \ ,$$

where $\pi\colon A^*\times A^* \to A^*$ is the relation 'product': π^{-1} is rational and the image of a rational, hence recognisable, of A^* is a rational of $A^*\times A^*$. *cf. Exam. 1.5, p. 528*

1.13 ***Büchi systems.*** *Let* K *be a* Büchi system *on* A^*. *Show that the relation* β_K *from* A^* *to itself defined by this system,* $f\beta_K = \{g \mid f \overset{\star}{\underset{\mathcal{B}\ K}{\vdash\!\!\dashv}} g\}$, *is rational.* *cf. Sec. II.6.4, p. 319*

Solution: We know that there exists a rational R_K of $\widetilde{A}^*$, contained in $\overline{A}^*A^*$, such that *cf. Prop. II.6.25, p. 319*

$$\forall f \in A^* \qquad f\beta_K = \{g \mid g \in (f R_K)\,\tau_A \cap A^*\} \ .$$

We also know, since $\overline{A}^*$ and A^* are two *recognisable* subsets of $\widetilde{A}^*$, that R_K is a *finite* union: *cf. Prop. II.2.19, p. 253*

$$R_K = \bigcup_{i\in I} \overline{S_i}\,T_i \ ,$$

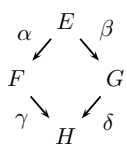
where the S_i are T_i are rationals of A^*. Lemma II.6.26 allows us to write

$$(f\,\overline{S_i}\,T_i)\tau_A \cap A^* = \big(\big((f\,\overline{S_i})\tau_A \cap A^*\big)\,T_i\big)\,\tau_A \cap A^* = \big((f\,\overline{S_i})\tau_A \cap A^*\big)\,T_i \ .$$

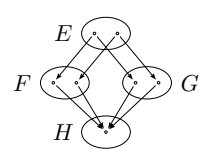
Then $f \mapsto (f\,\overline{S})\tau_A \cap A^*$ is rational for S rational in A^* since we have (Lemma II.6.27) *cf. p. 320*

$$(f\,\overline{S})\tau_A \cap A^* = \{g \in A^* \mid h \in S \quad \text{and} \quad g\,h = f\} = f\,S^{-1}.$$

1.14 *Verify that the commutative diagram*



Solution:



that is, $\alpha\gamma = \beta\delta$, *does not in general imply that* $\gamma\delta^{-1} = \alpha^{-1}\beta$.

1.15 *Give an example of rational relations* $\theta\colon A^* \to B^*\times C^*$ *and* $\sigma\colon B^*\times C^* \to D^*$ *where* $\theta\circ\sigma$ *is not a rational relation.*

Solution: Take $A = B = \{a\}$, $C = \{b\}$ and $D = \{a,b\}$, then define θ and σ by

$$\widehat{\theta} = \big(a,(a,b)\big)^* \qquad \text{and} \qquad \widehat{\sigma} = \big((a,1),a\big)^*\big((1,b),b\big)^* \ .$$

It follows that $(a^n,b^m)\sigma = a^nb^m$ and $\mathsf{Im}\,(\theta\circ\sigma) = \{a^nb^n \mid n\in\mathbb{N}\}$ is not rational.

1.16 ***On the Star Lemma.*** *Let* $\theta\colon A^* \to B^*$ *be a rational relation. Show that* it is not true *that there exists an integer* N *such that for all* (r,w) *in* $\widehat{\theta}$ *and* for every factorisation

cf. Lem. I.1.13, p. 71

$$(r,w) = (r',w')(f,g)(r'',w'')$$

such that $|f|+|g| > N$ *(or even* $|f| > N$ *and* $|g| > N$ *), there exists a factorisation*

$$(f,g) = (h,k)(u,v)(s,t)$$

such that (i) $1 \leqslant |u|+|v| < N$ *and* (ii) $(r',w')(h,k)(u,v)^*(s,t)(r'',w'') \subseteq \widehat{\theta}$.

Solution: Let the relation $\iota_L\colon A^* \to A^*$ be the intersection with the set $L = a^*b^*$. If N is the integer suggested in the statement, we consider the factorisation

$$(a^Nb^N, a^Nb^N) = (a^N,1)(b^N,a^N)(1,b^N) \ .$$

Since for every factorisation

$$(b^N,a^N) = (b^i,a^l)(b^j,a^m)(b^k,a^n)$$

such that $1 \le j+m < N$, we have

$$(a^N,1)(b^i,a^l)(b^j,a^m)^2(b^k,a^n)(1,b^N) = (a^Nb^{N+j}, a^{N+m}b^N) \ ,$$

a pair which does not belong to $\widehat{\iota_L}$ as soon as $m \neq j$ and the proposed statement is false.

1.17 ***Conjugacy.*** *Show that the relation* $\mathsf{Conj}\colon A^* \to A^*$ *which relates a word* f *to the set of its* conjugates

$$\mathsf{Conj}(f) = \{v\,u \mid u,v\in A^* \quad u\,v = f\} \ ,$$

is not a rational relation.

Solution: We indeed showed in Example 1.10 that $\iota_L\circ\rho\circ\iota_K$ is not rational (by setting $L = a^*b^*$ and $K = b^*a^*$). Then we have

cf. p. 537

$$\iota_L\circ\mathsf{Conj}\circ\iota_K = \iota_L\circ\rho\circ\iota_K \ .$$

1.18 ***Relation between commutative monoids.*** *Show that a relation* $\tau\colon M \to N$ *between commutative monoids is rational if and only if there exists an integer* k*, two morphisms* $\varphi\colon \mathbb{N}^k \to M$ *and* $\psi\colon \mathbb{N}^k \to N$ *, and a rational* K *of* $\mathbb{N}^k$ *such that*

$$\text{(i)}\ \ \widehat{\tau} = (K)\,\varphi\times\psi \qquad\qquad \text{or} \qquad\qquad \text{(ii)}\ \ \tau = \varphi^{-1}\circ\iota_K\circ\psi \ .$$

Solution: We repeat *verbatim* the proof of Theorem 1.2: the two conditions (i) and (ii) are equivalent by Lemma 1.2.

cf. p. 532

The condition is sufficient: if K is rational, $(K)\,\varphi\times\psi$ is also. The condition is necessary. Let $\widehat{\tau}$ be a rational subset of $M\times N$; by Property 1.3, we can assume that M and N are finitely generated. We thus have two surjective morphisms $\varphi\colon \mathbb{N}^{k'} \to M$ and $\psi\colon \mathbb{N}^{k''} \to N$, from which we obtain another

Prop. II.1.7 (a), p. 223

$$\varphi\times\psi\colon \mathbb{N}^{k'+k''} \longrightarrow M\times N \ .$$

There exists a rational K of $\mathbb{N}^k$ such that $(K)\,\varphi\times\psi = \widehat{\tau}$.

Prop. II.1.7 (b)

1.19 ***Iterations of rational relations.*** *Determine whether the iteration of each of the following rational relations (from $\{a,b\}^*$ to itself) is rational or not:*

(a) $\varphi : f\varphi$ *is the set of words obtained by replacing* one *factor* $a\,a$ *of* f *by* 1_{A^*} *;*

(b) $\chi : f\chi$ *is the set of words obtained by replacing* one *factor* $a\,b$ *of* f *by* $b\,a$ *;*

(c) $\psi : f\psi$ *is the set of words obtained by replacing* one *factor* $a\,b\,b$ *of* f *by* $b\,a\,a$ *;*

(d) $\omega : f\omega$ *is the set of words obtained by replacing* one *factor* $a\,b\,b$ *of* f *by* $b\,a$ *.*

Solution:[35] (a) $\overset{\star}{\varphi}$ is rational; (b) $\overset{\star}{\chi}$ is not rational, since the function τ defined by $f\tau = f\overset{\star}{\chi} \cap b^*a^* = b^{|f|_b}\,a^{|f|_a}$ is not rational; (c) $\overset{\star}{\psi}$ is rational; (d) $\overset{\star}{\omega}$ is not rational.

1.20 ***Recurrent and adult words.*** *Show that if σ is a rational relation, the relation which associates a word A^* with the set of its recurrent (resp. adult) descendants is a rational relation (Proposition 1.7).*

Solution : A complicated proof whose details are found in [214].

1.21 ***Iteration of a recognisable relation.*** *Show that the proper iteration of a recognisable relation is a recognisable relation. What can be said about the iteration?*

Solution : A recognisable relation θ (from A^* to A^*) is of the form

$$\widehat{\theta} = \bigcup_{i\in I} R_i \times S_i$$

with I finite, and the R_i and S_i in $\operatorname{Rat} A^*$. Let γ be the relation from I to itself defined by

$$i\,\gamma\,j \qquad \text{if and only if} \qquad S_i \cap R_j \neq \emptyset \ .$$

Property 1.7 implies that
$$\widehat{\theta^2} = \bigcup_{i\,\gamma\,j} R_i \times S_j \ ,$$
so, by induction on n
$$\widehat{\theta^{n+1}} = \bigcup_{i\,\gamma^n\,j} R_i \times S_j \ ,$$
and hence
$$\widehat{\overset{+}{\theta}} = \bigcup_{i\,\overset{\star}{\gamma}\,j} R_i \times S_j \ .$$

Then $\overset{+}{\theta}$ is recognisable since $\overset{\star}{\gamma} = \gamma^{\leqslant \|I\|}$.

The iteration $\overset{\star}{\theta} = \overset{+}{\theta} \cup \iota$ is recognisable if and only if $\iota \setminus \overset{+}{\theta}$ is finite.

1.24 *Give a rational expression for the graph of $\varphi_1\varphi_1^{-1}$ (cf. Example 1.12, p. 544). What is the interest of this exercise?*

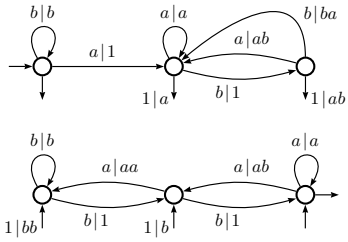
Solution : $\widehat{\varphi_1\varphi_1^{-1}} = [(a,a) + (b,b) + (c,c) + (a,c)(b,c)^*(b,a) + (c,a\,b)(c,b)^*(a,1)]^*$.

This computation conceals no mystery, but it makes us realise that, even in such a simple case as this, it is necessary to use realisation by representation to compute the expression of the composition of two rational relations.

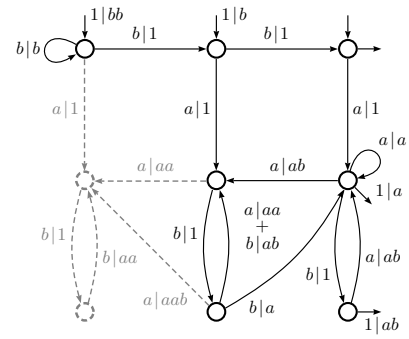
---

[35] *Note added in translation:* To tell the truth, these are more answers than solutions... The proof for $\overset{\star}{\omega}$ goes roughly as the one for $\overset{\star}{\chi}$: $[b^m a^n](\overset{\star}{\omega})^{-1} \cap a^*b^* = a^n\,b^{m2^n}$. The one for $\overset{\star}{\psi}$ requires more care. It can be shown that $f\,\overset{\star}{\psi} \cap [A^* \setminus A^*a\,b\,b\,A^*]$ is computed by the transducer constructed in Exer. 1.25, a result attributed to Schützenberger (*cf.* [96]), and the whole $\overset{\star}{\psi}$ is obtained as the composition of a modified version of the two transducers given in that exercise.

1.25 *Give a transducer which realises the composition of the relations realised by the transducers below.*



Solution:



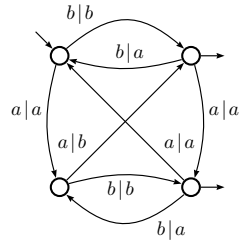
1.26 *Do the same for the transducers below.*

Solution: We immediately write the matrix representation of α_1:

$$\lambda = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad a\mu = \begin{pmatrix} 0 & a \\ b & 0 \end{pmatrix}, \quad b\mu = \begin{pmatrix} b & 0 \\ 0 & a \end{pmatrix}, \quad \nu = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and that of β_1:

$$\eta = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad a\kappa = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \quad b\kappa = \begin{pmatrix} 0 & b \\ a & 0 \end{pmatrix}, \quad \chi = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$



from which we obtain the matrix representation of $\alpha_1 \circ \beta_1$:

$$\zeta = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}, \quad a\pi = \begin{pmatrix} 0 & \begin{matrix} a & 0 \\ 0 & a \end{matrix} \\ \begin{matrix} 0 & b \\ a & 0 \end{matrix} & 0 \end{pmatrix}, \quad b\kappa = \begin{pmatrix} \begin{matrix} 0 & b \\ a & 0 \end{matrix} & 0 \\ 0 & \begin{matrix} a & 0 \\ 0 & a \end{matrix} \end{pmatrix}, \quad \omega = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix},$$

which gives the transducer above.

1.27 *If τ is a relation from A^* to M, we define τ_{f} and τ_∞ to be the 'finite and infinite parts' of τ respectively as follows:*

if $\|f\tau\|$ is finite, then $f\tau_{\mathsf{f}} = f\tau$ *and* $f\tau_\infty = \emptyset$ *otherwise* $f\tau_{\mathsf{f}} = \emptyset$ *and* $f\tau_\infty = f\tau$.

Show that τ is rational, and that τ_{f} and τ_∞ are rational and effectively computable from τ.[36]

Solution: Suppose that τ is recognised by a real-time transducer $\mathcal{T}$; a transducer whose label has a second component which is an infinite subset of M will be called 'red'.

The relation τ_∞ is recognised by the set of computations of $\mathcal{T}$ which contain *at least one* red transition. By the Coloured Transition Lemma, it is rational and we can construct from $\mathcal{T}$ a transducer which recognises it.

cf. Lem. II.3.11, p. 269

The relation τ_{f} is then given by $\widehat{\tau_{\mathsf{f}}} = \widehat{\tau} \bigcap [(\mathsf{Dom}\,\tau \setminus \mathsf{Dom}\,\tau_\infty) \times M]$, which is rational.

cf. Propty 1.8, p. 540

[36] Result due to Elgot and Mezei, [84]; *cf.* also [27, Exer. III.4.8].

1.29 *Give a simple example of a 2-tape automaton which can move its read head in both directions on one of two tapes and which accepts a non-rational subset of $A^* \times B^*$.*

Solution: Without embarking on a formal description, consider a 2-tape machine which accepts words of the form $a^n b^m$ on the first tape and words of the form c^p on the second, on which the read head can move in both directions.

At first the machine reads alternately a's and c's in such a way that it can test if there are as many c's as a's. If that is not the case, the computation blocks and the pair is not accepted. If it is the case, then reading the first b puts the machine into a state which makes the read head on the second tape return to the starting point. Then alternately reading b's and c's allows it to test whether there are as many c's as b's. At the end, the only pairs accepted are those of the form $(a^n b^n, c^n)$ which do not form a rational set.

2. $\mathbb{K}$-relations

2.2 ***Uniform convergence.*** *Let θ be a $\mathbb{K}$-relation from M to N and $\{s_n\}_{n\in\mathbb{N}}$ a sequence of series of $\mathbb{K}\langle\langle M\rangle\rangle$ that converge uniformly to a series s, such that $(s_n)\theta$ is defined for all n.*

(a) *Give an example, for $\mathbb{K} = \mathbb{Q}$, of such a $\mathbb{Q}$-relation, and such a sequence for which the sequence $\{s_n\theta\}_{n\in\mathbb{N}}$ does not converge in $\mathbb{Q}\langle\langle N\rangle\rangle$.*

(b) *Show that if the sequence $\{s_n\theta\}_{n\in\mathbb{N}}$ converges to a series t in $\mathbb{K}\langle\langle N\rangle\rangle$, then $s\theta$ is defined and is equal to t (Proposition 2.4).*

Solution: (a) Set $M = a^*$, and for θ take the $\mathbb{Q}$-relation 'sum of coefficients'. For the sequence of series take the polynomials *cf. Exer. III.1.20, p. 392*

$$p_n = 1 + \sum_{k=1}^{k=n} \frac{1}{k} a^k \quad , \quad \text{a sequence which converges uniformly to} \quad s = 1 + \sum_{k=1}^{k=\infty} \frac{1}{k} a^k \quad .$$

The $(p_n)\theta$ are all defined but tend to infinity and their limit is undefined.

(b) Proved by standard calculus techniques.

2.5 *Give the statement of Proposition 2.5 for the case of classic relations between words (that is, $\mathbb{K} = \mathbb{B}$).*

Solution:

$$\begin{aligned}\widehat{\theta\sigma} &= \{(m,p) \mid \exists n \in N \quad (m,n) \in \widehat{\theta} \quad \text{and} \quad (n,p) \in \widehat{\sigma}\} \\ &= \{(m,p) \mid \exists n \in N \quad (n,m) \in \widehat{\theta^{-1}} \quad \text{and} \quad (n,p) \in \widehat{\sigma}\} \\ &= \{(m,p) \mid \exists (n,m,p) \in \widehat{\theta^{-1}} \times P \cap M \times \widehat{\sigma}\} \;=\; \left(\widehat{\theta^{-1}} \times P \cap M \times \widehat{\sigma}\right) \pi_N \quad .\end{aligned}$$

2.7 *Show that* if $\theta\colon M \to N$ is a *function* and T a subset of M in bijection with N under θ, then, for all s in $\mathbb{K}\langle\langle N\rangle\rangle$, $s = \left[\left(s\underline{\theta^{-1}}\right) \odot \underline{T}\right] \underline{\theta}$ *(Lemma 2.8).* *cf. p. 555*

Solution: The equation holds for the characteristic series of each n in N and extends by linearity and continuity to all series in $\mathbb{K}\langle\langle N\rangle\rangle$.

3. Rational $\mathbb{K}$-relations

3.3 *Give a direct proof (that is, without using Proposition III.3.12) of Proposition III.3.13:* *cf. p. 436*
if M is rationally enumerable, then $\mathbb{K}\mathrm{Rec}\, M \subseteq \mathbb{K}\mathrm{Rat}\, M$.

Solution: By the definition of a rationally enumerable monoid, let $\alpha\colon A^* \to M$ be a surjective morphism and T a rational (hence recognisable) set of A^* such that α is a bijection from T to M.

Cor. III.2.20, p. 415 Let $s \in \mathbb{K}\mathrm{Rec}\, N$; by Proposition 2.12, $s\underline{\alpha^{-1}} \in \mathbb{K}\mathrm{Rec}\, A^*$. The characteristic series $\underline{T}$ is a
Th. III.3.2, p. 428 $\mathbb{K}$-rational, hence $\mathbb{K}$-recognisable, series of A^*. By Schützenberger's Theorem, $s\underline{\alpha^{-1}} \odot \underline{T}$ is a $\mathbb{K}$-recognisable, hence $\mathbb{K}$-rational, series on A^* and finally, $s = \left[\left(s\underline{\alpha^{-1}}\right) \odot \underline{T}\right] \underline{\alpha}$ (Lemma 2.8) is $\mathbb{K}$-rational by Theorem 3.1 (i).

3.4 ***Support of $\mathbb{R}$-rational series on A^*.*** *Prove:*

cf. Prop. III.4.23 **Proposition 3.4** [**Fliess**] *Let $\mathbb{K}$ be a sub-semiring of $\mathbb{R}$. The set of supports of the $\mathbb{K}$-rational series on A^* is closed under morphism and inverse morphism.*

Solution:[37] Let $\alpha\colon A^* \to B^*$ be a morphism of monoids. If $t \in \mathbb{K}\langle\langle B^*\rangle\rangle$, the equality
cf. p. 557 $\mathsf{supp}(t\underline{\alpha^{-1}}) = (\mathsf{supp}\, t)\alpha^{-1}$ is immediate and Proposition 2.12 ensures closure under inverse morphism.

If α is *continuous*, then $s\underline{\alpha}$ is defined and, if all of s's coefficients are positive (which we can
cf. p. 464 always assume, *cf.* the proof of Proposition III.4.23), $\mathsf{supp}(s\underline{\alpha}) = (\mathsf{supp}\, s)\alpha$. Theorem 3.1 (i) gives us the conclusion for this family of morphisms.

This result's interest therefore comes from the case where α is *not* continuous and where, given s, $s\underline{\alpha}$ is not necessarily defined. As every morphism (from A^* to B^*) is the composition
cf. Exer. 0.4.4, p. 26 of an injective (and hence continuous) morphism and a sequence of elementary projections, it suffices to prove the result for $A = B \cup \{x\}$, with α the identity on B and $x\alpha = 1_{B^*}$.

Suppose that s is realised by a representation of dimension n and let

$$L = x^{<n}\left(B\, x^{<n}\right)^* = A^* \setminus A^* x^n A^*$$

be the (rational) language of words which contain no factor x^m with $m \geqslant n$. The relation
cf. p. 428 $\iota_{\underline{L}} \circ \underline{\alpha}$ is of finite co-image, and $s\iota_{\underline{L}} \circ \underline{\alpha} = (s \odot \underline{L})\underline{\alpha}$ is defined. By Theorem III.3.2, $s \odot \underline{L}$ is rational, and we have $\mathsf{supp}(s \odot \underline{L}) = \mathsf{supp}\, s \cap L$. Since $\mathsf{supp}\,((s \odot \underline{L})\underline{\alpha}) = (\mathsf{supp}(s \odot \underline{L}))\,\alpha = (\mathsf{supp}\, s \cap L)\alpha$, it remains to show

$$(\mathsf{supp}\, s)\alpha = (\mathsf{supp}\, s \cap L)\alpha \quad . \tag{SE.3.1}$$

Let $f = g\, x^n h$ be in $\mathsf{supp}\, s$. By reasoning identical to that in the proof of Theo-
cf. p. 457 rem III.4.9, $\langle s, g\, x^n h\rangle$ is a linear combination of $\langle s, g\, x^i h\rangle$ for $0 \leqslant i \leqslant n-1$ and if $\langle s, g\, x^n h\rangle$ is non-zero, then so is at least one $\langle s, g\, x^i h\rangle$. An induction on the length of f completes the proof of (SE.3.1).

3.5 *Let $(\zeta, \pi, \omega) = (\lambda, \mu, \nu) \circ (\eta, \kappa, \chi)$ be the composition product of representations. Verify that if all the coordinates of ν and χ (resp. of λ and η) are equal to 1, then so are the components of ω (resp. of ζ).*

cf. p. 567 *Solution*: Since κ is a morphism, $(1)\kappa = 1$, the identity matrix of size R, and Equations (3.5) become respectively $\omega_{(q,R)} = \chi$ and $\zeta_{(q,R)} = \eta$, for all q in Q.

4. Equivalence of finite $\mathbb{K}$-transducers

4.4 *Prove that if A is an alphabet with at least two letters,* $\mathrm{Rec}\,(A^* \times c^*)$ *is undecidable*
cf. p. 573 *in* $\mathrm{Rat}\,(A^* \times c^*)$ *(Corollary 4.6).*

cf. pp. 570–573 *Solution*: We use the proof of Corollary 4.2 and the notation of that of Theorem 4.3.

If $\complement W_\theta \cup \complement W_\mu$ is recognisable, its complement $W_\theta \cap W_\mu$ is also. We note on one hand that $V_\theta \cap V_\mu$ is the graph of a function, as the intersection of the graphs of two functions, hence so is $W_\theta \cap W_\mu$. On the other hand, if $V_\theta \cap V_\mu$ is non-empty, this function has an infinite image (since if $(f, f\theta)$ belongs to $V_\theta \cap V_\mu$, then so does $(f^n, (f^n)\theta)$), and this also
cf. p. 540 holds for $W_\theta \cap W_\mu$. This contradicts Property 1.9.

[37] Adapted from [31].

Thus, $\complement W_\theta \cup \complement W_\mu$ is recognisable if and only if $V_\theta \cap V_\mu$ is empty, which we know to be equivalent to the Post Correspondence Problem.[38]

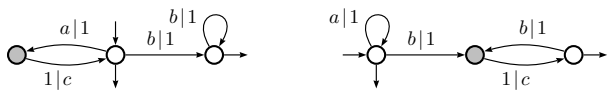
4.5 *Explain why the fact that we can decide if the label of a computation of $\mathcal{V}$ is equal to 0 (in the proof of Proposition 4.3) does not contradict the fact that it is undecidable whether a $\mathbb{Z}$-rational series on A^* has a coefficient equal to 0.* *cf. p. 571* *cf. Th. III.4.17, p. 468*

Solution: It is not contradictory because there is no connection between the two facts. The automaton $\mathcal{V}$ is an *automaton over* $\mathbb{Z}$: that is it accepts *subsets* of $\mathbb{Z}$. We do not take the *sum* of the labels of the successful computations but their *union*. On the contrary, in the case of $\mathbb{Z}$-automata over A^*, we take, for each word f, *the sum of the coefficients* of the successful computations whose label is f and what is undecidable is whether there exists an f such that this sum is zero.

5. Deterministic rational relations

5.1 *Verify that $V_1 = \{(a^n b^m, c^n) \mid n, m \in \mathbb{N}\}$ and $W_1 = \{(a^n b^m, c^m) \mid n, m \in \mathbb{N}\}$ are the behaviours of deterministic transducers from $\{a, b\}^*$ to $\{c\}^*$.*

Solution: V_1 and W_1 are realised respectively by the transducers

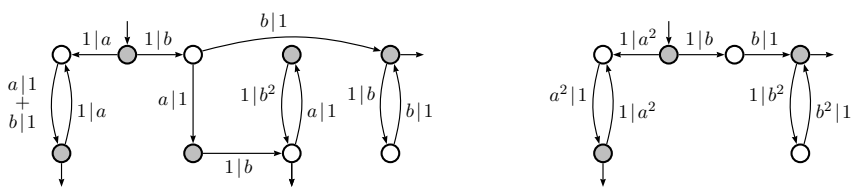


5.5 *Verify that the relations $\sigma_3, \sigma_4 \colon \{a, b\}^* \to \{a, b\}^*$ defined by*

$$a^n\sigma_3 = \{a^n, b^{2n}\} \ , \qquad b^n\sigma_3 = \{a^n, b^n\} \qquad \text{and} \qquad a^{2p}\sigma_4 = a^{4p} \ , \qquad a^{2p+1}\sigma_4 = b^{2p+1} \ ,$$

are pure deterministic rational.

Solution: *cf.* Figure SE.5.1.



(a) transducer for σ_3 (b) transducer for σ_4

Figure SE.5.1: Deterministic transducers without an endmarker

5.6 *Let $\sigma \colon A^* \to B^*$ be a rational substitution. Give a sufficient condition for σ to be a deterministic relation.*

Solution: If $a\sigma$ is a *prefix* rational set of B^*, for each a in A, then σ is a deterministic relation.

[38] *Note added in translation:* The proposed solution yields the answer for alphabets with at least 8 letters (*cf.* proof of Theorem 4.3, p. 573); the extension to alphabets with only 2 letters uses a rational injective map.

5.7 *Let K be a rational of A^*; show that $\iota_K : A^* \to A^*$, the intersection with K, is a pure deterministic relation.*

Solution: Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ be a deterministic automaton over A which realises K. The transducer $\mathcal{T} = \langle Q \times \{0,1\}, A, A, F, I \times \{0\}, T \times \{0\} \rangle$ defined by

$$F = \{\big((p,0),(a,1_{A^*}),(p,1)\big) \mid (p,a,q) \in E\} \cup \{\big((p,1),(1_{A^*},a),(q,0)\big) \mid (p,a,q) \in E\}$$

is deterministic, because $\mathcal{A}$ is, and realises ι_K.

5.9 *Give a sufficient condition for a recognisable relation to be a pure deterministic relation.*

Solution: The recognisable relation τ, given by

$$\widehat{\tau} = \bigcup_{i \in I} R_i \times S_i \ .$$

cf. Sec. 5.5, p. 590 is pure deterministic if the R_i, or the S_i, form a prefix family of languages.

cf. pp. 590–592 **5.10** *Prove the sufficient conditions of Proposition 5.21 and Theorem 5.2.*

Solution: A somewhat cumbersome construction, but which presents no special difficulty, and which is explained in [181, pp. 33 *et seq.*].

5.12 *Give an example of a deterministic relation θ such that $\theta\theta^{-1}$ is not a deterministic relation.*

Solution: If we take for θ the relation σ_4 from Exercise 5.5, there follows

$$a^n \sigma_4 \sigma_4^{-1} = \{a^n, b^n, b^{2n}\} \ ,$$

a relation which we can show is not deterministic as in Exercise 5.4.

5.14 *Prove that if $\theta : A^* \to B^*$ is a deterministic relation, K a rational of A^* and L a rational of B^*, then the relations $\iota_K \theta$ and $\theta \iota_L$ are deterministic (Property 5.5).*

Prop. 5.16, p. 588 *Solution*: $\widehat{\iota_K \theta} = \widehat{\theta} \cap K \times B^*$, $\widehat{\theta \iota_L} = \widehat{\theta} \cap A^* \times L$, and the intersection of a deterministic relation with a recognisable relation is a deterministic relation.

6. Synchronisation of transducers

6.1 *Supposing that all the alphabets have two letters, how many n-tuples are there in $A_1^* \times A_2^* \times \cdots \times A_n^*$ whose length discrepancy is less than or equal to k?*

Solution: $\|\mathrm{Diff}_k\| = n\left(\left(2^{k+1}-1\right)^{n-1} - 1\right) + 1$.

6.2 *Show that* $\mathrm{Fin}\,(A^* \times B^*) = \mathrm{Rec}\,(A^* \times B^*) \cap \mathrm{Rat}_{\mathsf{bld}}\,(A^* \times B^*)$.

Solution: The inclusion $\subseteq$ is obvious. If $R \in \mathrm{Rec}_2$, then $R = \bigcup_I T_i \times S_i$; if there exists an i for which T_i is not finite, there exist some arbitrarily long words f in T_i, and hence some pairs (f,g) of arbitrarily large length discrepancy in R.

6.3 *We write* $\lambda\colon A^*\times B^* \to \mathbb{N}^2$ *for the* length *morphism:* $(f,g)\lambda = (|f|,|g|)$. *Let* M *be a* stable submonoid *of* $\mathbb{N}^2$ *and* $S = M\lambda^{-1}$ *be a submonoid of* $A^*\times B^*$. *cf. Exer. II.7.10, p. 333, and p. 370*

(a) *Prove that* a rational relation on $A^*\times B^*$ whose graph is contained in S is a rational subset of S *(Theorem 6.3):*[39] *p. 600*

$$R \in \operatorname{Rat} A^*\times B^* \quad \text{and} \quad R \subseteq S \quad \Longrightarrow \quad R \in \operatorname{Rat} S \ .$$

(b) *Show that, given a stable monoid* M *in* $\mathbb{N}^2$, *we can decide whether a rational relation is contained in* $M\lambda^{-1}$ *(Corollary 6.6).* *p. 602*

(c) *What can we deduce for relations such that the image of every word* f *is formed of words of length less than or equal to that of* f*?*

(d) *Give a necessary and sufficient condition on* M *for* $\operatorname{Rat} S$ *to be a Boolean algebra.*

Solution: (a) The proposition is analogous to Theorem 6.1, and its proof is a simple rewording of that of Theorem 6.2. We start by proving the corresponding generalisation of the Resynchronisation Lemma: *p. 598*

Lemma 6.1 *For all* $H \in \operatorname{Rat} S$ *and every word* f *in* A^*, *there exists a family of rationals* $H_{f,g}$ *of* S, *indexed by the words of the same length as* f, $\{H_{f,g} \mid g \in A^{|f|}\}$, *and such that*

$$(f, 1_{B^*})\, H = \bigcup_{g\in A^{|f|}} H_{f,g}\,(g, 1_{B^*}) \ .$$

Proof. Choose an integer k. For all f in A^k and all $u = (s,t)$ in $A^*\times B^*$, there exists a unique $v = (s',t)$ in $A^*\times B^*$ and a unique g in A^k such that

$$(f, 1_{B^*})\,(s,t) = (s',t)\,(g, 1_{B^*}) \ .$$

We have $(s',t)\lambda = (s,t)\lambda$ and hence (s',t) is in S with (s,t).

The rest of the proof is identical to that of the original lemma. ■

Proof of Theorem 6.3. By induction on the star height of R in $\operatorname{Rat} A^*\times B^*$. If R of height 0, then R is a finite set of elements of S and hence a rational of S. If R is of height $h+1$, then R is a finite union of monomials of the form

$$P = u_0\, R_1^*\, u_1\, R_2^*\, u_2 \cdots u_{r-1}\, R_r^*\, u_r \ ,$$

with the u_i in $A^*\times B^*$, and the R_i in $\operatorname{Rat} A^*\times B^*$ of star height h. Let w be in an arbitrary R_i; we have

$$u_0 \cdots u_{i-1}\, w^*\, u_i \cdots u_r \subseteq P \subseteq R \subseteq S \ ,$$

by which, taking the image under λ,

$$(u_0 \cdots u_r)\lambda + [(w)\lambda]^{\oplus} \subseteq M \ ,$$

and hence, since M is *stable*, $(w)\lambda$ is in M: that is, w is in S.

Thus, each R_i is included in S and, by the induction hypothesis, belongs to $\operatorname{Rat} S$, hence so does R_i^*. With the new version of Lemma 6.1, we complete the proof in exactly the same way as that of Theorem 6.2. ■

(b) As every step of the proof of Lemma 6.1 and Theorem 6.3 is effective, we deduce Corollary 6.6.

(c) $R \in \operatorname{Rat} A^*\times B^*$ is such that $(f,g) \in R \Rightarrow |g| \leqslant |f|$ if and only if $R\lambda$ is included in the stable submonoid $\{(1,1),(1,0)\}^{\oplus}$. Hence R can be realised[40] by an automaton labelled by elements of $(A\times B)\cup(A\times 1_{B^*})$.

(d) $\operatorname{Rat} S$ is a Boolean algebra if and only if M is a free monoid; that is, generated by a unique element.

---

[39] A result stated (in a slightly erroneous form, *mea culpa*) in [94].

[40] Proof of a result due to J. Leguy, *RAIRO Inf. Théor.*, 1981 [147]; *cf.* also [13, § 1.4].

p. 603 **6.6** *Explain why the last sentence of the proof of Lemma 6.7,* 'This value can be computed for any of the simple sub-computations of c which links q_0 to q_r', *is necessary; that is, there could be several distinct simple sub-computations which link q_0 to q_r, and true; that is, these sub-computations all have the same delay at the destination.*

Solution: Consider the computation

$$c \;:=\; q_0 \xrightarrow{(u_1,v_1)} p \xrightarrow{(f,g)} q \xrightarrow{(x,y)} p \xrightarrow{(h,k)} q \xrightarrow{(u_2,v_2)} q_r \;\;.$$

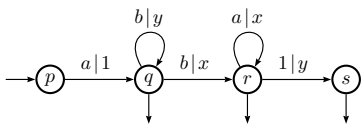
The two computations

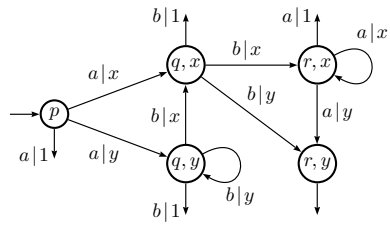
$$c \;:=\; q_0 \xrightarrow{(u_1,v_1)} p \xrightarrow{(h,k)} q \xrightarrow{(u_2,v_2)} q_r \quad\text{and}\quad c \;:=\; q_0 \xrightarrow{(u_1,v_1)} p \xrightarrow{(f,g)} q \xrightarrow{(u_2,v_2)} q_r$$

can be distinct. However, $|f| - |g| = |h| - |k| = -\big(|x| - |y|\big)$ because of the two loops

$$p \xrightarrow{(f,g)} q \xrightarrow{(x,y)} p \quad\text{and}\quad q \xrightarrow{(x,y)} p \xrightarrow{(h,k)} q \;\;.$$

6.8 *Apply the resynchronisation algorithm to the following transducer:*

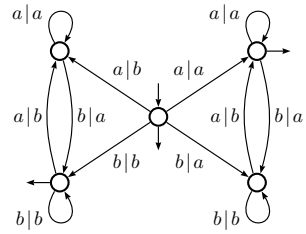




Solution: The transducer in the adjacent figure.

6.12 (a) *Let $\sigma\colon A^* \to A^*$ be the map which relates a word f to the word obtained by placing the last letter of f at the beginning: that is, $(h\,a)\sigma = a\,h$, for all a in A and all h in A^*. Give the deterministic letter-to-letter transducer which realises σ.*

(b) *Show that $\overset{\star}{\sigma}$ (the iteration of σ)* is not *a rational relation.*



Solution: (a) The transducer in the adjacent figure.

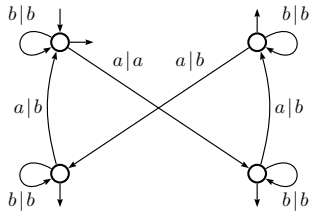
(b) $\overset{\star}{\sigma}$ is the conjugacy relation Conj, which is not rational (*cf.* Exercise 1.17, p. 538 and p. 630).

6.13 *Let α_1 and β_1 be the relations defined by the following transducers:*

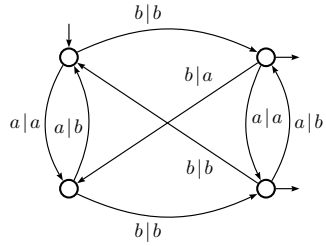
Construct the transducers which recognise: (a) $\alpha_1 \circ \alpha_1$; (b) $\beta_1 \circ \alpha_1$.

Solution:

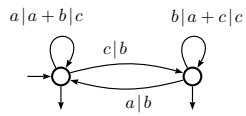
(a)

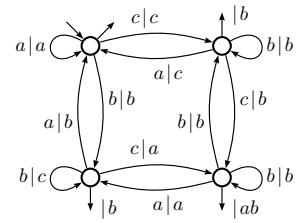


(b)



6.14 *Let δ_1 be the relation realised by the transducer below. Compute the transducer which realises $\delta_1 \circ \delta_1$.*





Solution: The transducer in the adjacent figure.

6.17 (a) *Describe the synchronised transducer which realises the recognisable relation $K \times L$ starting with the automata $\mathcal{A}$ and $\mathcal{B}$ over A^* and B^* which recognise K and L respectively.*

(b) *Construct this automaton in the case where $A = B = \{a, b\}$ and where*

$$K = \{f \mid |f|_b \equiv 1 \bmod 2\} \quad \textit{and} \quad L = \{g \mid |g| \equiv 2 \bmod 3\} \ .$$

Solution: (a) Let $\mathcal{A} = \langle Q, A, E, I, T \rangle$ and $\mathcal{B} = \langle R, B, F, J, U \rangle$ be the two automata which recognise K and L respectively. For each q in Q, let K_q be the language recognised by the automaton $\langle Q, A, E, \{q\}, T \rangle$, and for each r in R, let L_r be the language recognised by the automaton $\langle R, B, F, \{r\}, U \rangle$.

A letter-to-letter automaton with final function $\mathcal{C}$ whose behaviour is $K \times L$ is obtained by a sort of product which we can call the 'synchronised product' of $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{C} = \mathcal{A} \bowtie \mathcal{B}$, defined in the following manner: $\mathcal{C} = \langle Q \times R, A \times B, H, I \times J, V \rangle$ with

$$H = \{((p, r), (a, b), (q, s)) \mid (p, a, q) \in E \quad \text{and} \quad (r, b, s) \in F\}$$

and
$$V_{q,r} = \begin{cases} (1_{A^*} \times L_r) & \text{if} \quad T_q = 1 \ \text{and} \ U_r = 0\,, \\ (K_q \times 1_{B^*}) & \text{if} \quad T_q = 0 \ \text{and} \ U_r = 1\,, \\ 1 & \text{if} \quad T_q = 1 \ \text{and} \ U_r = 1\,, \\ 0 & \text{if} \quad T_q = 0 \ \text{and} \ U_r = 0\,. \end{cases}$$

(b) We obtain the automaton of Figure SE.6.1 by setting $K_0 = \{f \mid |f|_b \equiv 0 \bmod 2\}$, $L_0 = \{g \mid |g| \equiv 0 \bmod 3\}$ and $L_1 = \{g \mid |g| \equiv 1 \bmod 3\}$.

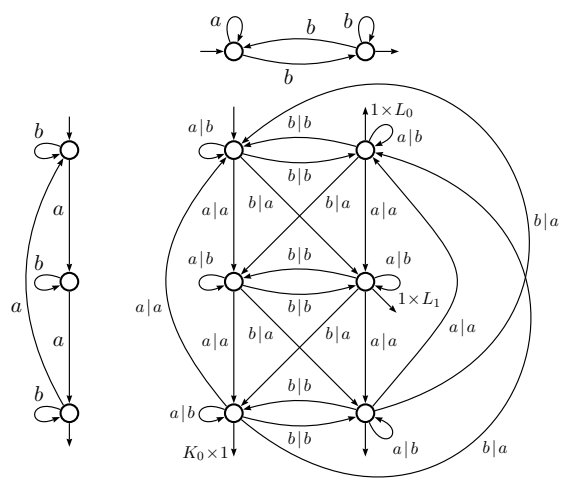


Figure SE.6.1: The synchronised product of Exercise 6.17

6.19 *Prove that* a synchronous rational relation is a deterministic rational relation *(cf. Proposition 6.21).*
p. 616

Solution: A synchronous rational relation is realised by a letter-to-letter automaton (which we can assume to be *deterministic*) over $A \times B$ with a final function in $\mathrm{Diff}_{\mathsf{rat}}(A^* \times B^*)$:

$$\mathcal{A} = \langle Q, A \times B, E, \{i\}, T \rangle \ .$$

The two-tape deterministic automaton $\mathcal{B}$ equivalent to $\mathcal{A}$ has two parts: the first performs the transitions of $\mathcal{A}$, and the second deals with the recognition of all the rational sets which appear as the image of the final function ω. It switches from one to the other when the endmarker \$ is read. More precisely, let

$$\mathcal{B} = \langle Q \cup R \cup S, A_{\$}, B_{\$}, F \cup H, \{i\}, U \rangle$$

be the automaton defined by:

(i) $R = Q \times A$ and, for every transition $(p, (a, b), q)$ in E, we have the transitions $(p, (a, 1), (p, a))$ and $((p, a), (1, b), q)$ in F.

(ii) For each q in Q, we have $T_q = (K_q \times 1) \cup (1 \times L_q)$. Let $\mathcal{T}_q$ be the deterministic automaton (whose initial state is j_q) which recognises L_q and, for each a in A, let $\mathcal{S}_{a,q}$ be the deterministic automaton (whose initial state is $l_{a,q}$) which recognises $a^{-1} K_q$.

The set S is the union of the states of all the $\mathcal{T}_q$ and of all the $\mathcal{S}_{a,q}$.

(iii) The set of transitions H is formed from:

(a) the union of the transitions of all the $\mathcal{T}_q$, read from the second tape: that is, if (x, a, y) is a transition of $\mathcal{T}_q$, then $\big(x, (1, a), y\big)$ is a transition in H, augmented with the transitions $\big(q, (\$, 1), j_q\big)$, for all q in Q;

(b) the union of the transitions of all the $\mathcal{S}_{a,q}$, read from the first tape; that is, if (x, b, y) is a transition of $\mathcal{S}_{a,q}$, then $\big(x, (b, 1), y\big)$ is a transition in H, augmented with the transitions $\big((q, a), (1, \$), l_{a,q}\big)$, for all a in A and all q in Q.

(iv) The set U of final states of $\mathcal{B}$ is the union of final states of all the $\mathcal{T}_q$ and of all the $\mathcal{S}_{a,q}$.

It is not hard to verify that $\mathcal{B}$ is deterministic and equivalent to $\mathcal{A}$.

NOTES & REFERENCES

The study of transducers or automata with k tapes or even automata with output is almost as old as that of finite automata. Several works are devoted to it, at least in part.

S. Ginsburg, *The Mathematical Theory of Context-Free Languages*. McGraw-Hill, 1966. [99]

S. Eilenberg, *Automata, Languages and Machines*, Volume A. Academic Press, 1974. [80]

J. Berstel, *Transductions and Context-Free Languages*. Teubner, 1979. [27]

J.-M. Autebert and L. Boasson, *Transductions rationnelles*. Masson, 1988. [13]

The study of weighted relations is also tackled in the following works:

A. Salomaa and M. Soittola, *Automata-Theoretic Aspects of Formal Power Series*. Springer, 1977. [218]

W. Kuich and A. Salomaa, *Semirings, Automata, Languages*. Springer, 1986. [142]

1. Rational relations: an introduction

The material of this section is now standard and draws on that of the works cited above. In particular, a number of examples and exercises are borrowed from [27].

If, as we said above, the automata which realise relations, or functions, have been studied since the founding of the subject – we could mention Mealy [170], refer to Raney [192], recall and cite Rabin and Scott [190], and suggest Schützenberger [220] – the first systematic study of rational relations is due to Elgot and Mezei [84]. They prove, among other things, Theorems 1.1 and 1.4, and most of the properties of synchronous relations (see below) and the theory of decomposition of rational functions (see the next chapter).

As its usual name indicates, Theorem 1.2 is due to M. Nivat [178]; the proof here is that from Eilenberg's treatise [80].

2.-3. $\mathbb{K}$-relations and rational $\mathbb{K}$-relations

Weighted relations were studied in [178] then, in the works of Fliess [87, 88, 90], in relation to the study of algebraic series (which are not covered in the present work). The use of the Evaluation and Composition Theorems as guide when making choices is that of Fliess and the definition of 'reasonable' semirings was inspired by Proposition 3.2 which is also due to him [90].

Another successful theory of weighted relations, slightly different from that which I have tried to present here, is that of Jacob [126, 127]. It consists of defining with 'regulated rational transductions' the largest possible family of relations which satisfy the pivotal theorems and which correspond to total maps (and hence maps whose composition is also always defined), and to do that *independently of the semiring of coefficients*. This point of view was adopted in related works [218, 142] which popularised the work of Jacob.

Proposition 2.9 was borrowed from [31]. The composition product of representations (Theorem 1.8 and Theorem 3.6) is due to Schützenberger [220].

4. Equivalence of finite $\mathbb{K}$-transducers

Theorem 4.1 (the undecidability of intersection) is due to Rabin and Scott [190], and Theorem 4.2 (universality is undecidable), and Corollaries 4.1 and 4.2, to Fischer and Rosenberg [86].

Proposition 4.3 is due to Gibbons and Rytter [98]. The proof that they give relies on the decidability of the intersection of (one-counter) context-free languages with the rationals.

Note that this result was generalised by Aalbersberg and Hoogeboom, who described partially commutative monoids in which intersection is decidable [1].

Theorem 4.3 is due, independently, to Ibarra [125] and Lisovik [151]. The proof given here reproduces that of [151].

The decidability of the equivalence of automata over a product of k free monoids and with weights in $\mathbb{N}$ (Theorem 4.4) is due to T. Harju and J. Karhumäki [111]. They use in particular the notion of a strictly graded monoid to deduce the result directly from Theorem 4.7. The problem was posed for *deterministic* k-tape automata in [190] and was solved for $k = 2$ by M. Bird [33] by an *ad hoc* method, then by L. Valiant [244] as a corollary of the decidability of the equivalence of 'finite-turn' deterministic pushdown automata. The problem remained open for $k \geqslant 3$.

5. Deterministic rational relations

The definition of deterministic relations goes back to the dawn of automata theory [190] and most of the results about them – with the exception of the decidability of equivalence (Corollary 5.19) sketched above, and the Theorem of Lexicographic Selection (Theorem V.3.3) which we shall prove in the next chapter – are considered as folklore. The presentation given here, and in particular Theorem 5.2, is taken from an article written with Maryse Pelletier [181].

6. Synchronisation of transducers

This section was adapted from an article written with Christiane Frougny [95]. The Fundamental Theorem (Theorem 6.1) is due to Elgot and Mezei [84], the Resynchronisation Lemma (Lemma 6.1) and the proof of Theorem 6.1 based on it are due to Eilenberg and Schützenberger (*cf.* [80]). The resynchronisation algorithm for automata with bounded lag is taken from [95] and the proof of Theorem 6.1 based on it has a lower complexity.[41] The many generalisations of Theorem 6.1 were mentioned in Exercise 6.3, with corresponding references.

Synchronous relations are the 'finite-automata-definable relations' of [84] and the 'sequential relations' of Büchi [44] (*cf.* also [242]). The developments of the theory, which includes other relations defined by finite automata, as well as other relations which could be called sequential (see the next chapter) made the introduction of a new term necessary.[42] Corollary 6.13 and Proposition 6.18 can be found in [84].

7. Malcev–Neumann series

The material of this section is standard if not elementary algebra, and is explained in sufficiently comprehensive treatises such as [65]. The description of the order on the free group is taken from [177]. As already mentioned, Theorem 5.2 was proved by Higman [118] to simplify the original proof of Theorem 4.7 by Neumann [176]. The Russian version of the same result was proved in [163].

*

The epigraph of the Solution to the Exercises section (p. 627) is taken from the catalogue of the exhibition 'L'étrange et le merveilleux en terres d'Islam', 23 Avril–23 Juillet 2001, Musée du Louvre, ©Réunion des Musées Nationaux.

[41] *Cf.* the comment at the beginning of Section 0.8, p. 34.

[42] *Note added in translation:* Synchronous relations were also studied and given a logical characterisation in [82].

Chapter V

The simplicity of functional transducers

Functions realised by finite automata

As announced in the last chapter, this chapter deals with functional relations, or *functions*, which can be realised by finite automata, and which, by habit and for convenience, and despite some confusion with other areas of mathematics, we shall call *rational functions*. The assumption of functionality combined with that of rationality will give remarkable structural results.

We start by proving that the fact of being functional is a *decidable property* for a rational relation. Then we define some families of functions that we will use in what follows, namely *sequential* (and *co-sequential*) functions, which are realised by transducers which, roughly, are to arbitrary functional transducers what deterministic (and co-deterministic) automata are to arbitrary automata.

The second section deals with the *uniformisation of rational relations by rational functions*. We deduce this result from the construction of a *Schützenberger covering* of an automaton, and this proof method will permit us to use it as the basis of all succeeding developments. In particular, we deduce simply that every every rational function has a *semi-monomial matrix representation*, which gives the principal structural result, namely that every rational function is the product of a sequential function and a co-sequential function, and it is this which allows us to say that rational functions 'are simple'.

These results continue into the following section; first with the notion of the cross-section, which is dual to that of uniformisation, a reversal of the point of view which turns out to be fruitful. We continue with a closer and more technical study of the way in which uniformisations and cross-sections can be constructed.

The fourth and final section is devoted to the study of sequential functions. We first show that the quality of being sequential is a decidable property for a rational function, by a procedure which directly extends that which we used to decide whether a transducer is functional. We then develop the parallel which can be made between sequential functions and rational (or recognisable) languages.

Sequential functions are amenable to two characterisations. The first starts from the definition of *translations of a function*, which is the exact analogue of the *quotients of a language*. A function is sequential (rational) if and only if the set of its translations is finite, and this set allows us to construct canonically a *minimal transducer* for the function, just as the set of quotients allows us to define the minimal automaton of a language.

The second characterisation is topological in nature and is based on the definition of what we call *uniformly bounded* functions. This property characterises sequential (rational) functions in the family of rational functions (Little Sequentiality Theorem) and indeed in the family of functions whose inverse preserves rationality (Great Sequentiality Theorem).

Contents of the chapter

1. Functionary
2. Uniformisation of rational relations
3. Cross-section of rational functions
4. Sequential functions

Solutions to the exercises

Notes & references

1 Functionary

In the last chapter we saw several examples of rational relations which were functions (morphisms, intersection with rationals *etc.*). Recall that of Exercise IV.1.6: *cf. p. 529*

Example 1.1 The function $\theta_1 \colon \{a\}^* \to \{x\}^*$, defined by

$$(a^{2p})\theta_1 = x^{4p} \qquad \text{and} \qquad (a^{2p+1})\theta_1 = x^{2p+1} \ ,$$

has a graph which is the rational set $\widehat{\theta_1} = (a^2, x^4)^* \cup (a, x)\,(a^2, x^2)^*$: it is a rational function. □

Its definition implies that the relation θ_1 is a function; what we must verify is that its graph is indeed a rational set. However, the problem can occur the other way around.

Example 1.2 Consider the transducer $\mathcal{G}_1$ in Figure 1.1; by definition the relation γ_1 which it realises is rational. What we must verify, and it is not obvious in this case, is that γ_1 is a *function*. □

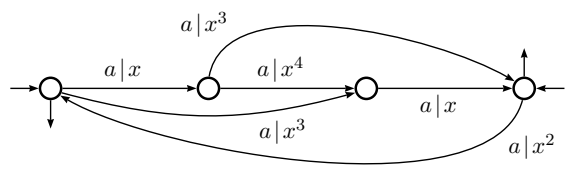


Figure 1.1: Is the relation γ_1 realised by the transducer $\mathcal{G}_1$ functional?

We start by proving that functionality is a decidable property. We then describe some families of remarkable rational functions which we will study in more detail and use in the rest of the chapter.

1.1 Deciding functionality

The first question which arises is thus to know whether, given a rational relation, defined by a (finite) transducer which realises it, or, which comes to the same thing, by a rational expression which denotes its graph, we can *effectively* tell that it is a function, and thereby exploit the functional properties which we shall prove. We have just said that the answer is affirmative; in fact it is not hard to obtain, in two senses: on one hand the proof of this assertion is relatively simple, and on the other the complexity of the algorithm which the proof describes is reasonable, in fact quadratic in the size of the transducer under consideration.

A relation $\theta \colon A^* \to B^*$ is a *function* if the image $f\theta$ of a word f in A^*, $f\theta = \{g \in B^* \mid (f, g) \in \widehat{\theta}\}$, has *at most* one element; that is, if for all f in A^*, *cf. Note 0.1, p. 9*

$$\|f\theta\| \leqslant 1 \ .$$

The domain of the function θ is $\mathsf{Dom}\,\theta = \{f \mid \|f\theta\| = 1\}$. We have thus implicitly defined a '$\mathbb{B}$-function' since implicitly we are interested in '$\mathbb{B}$-relations'. Now we could also be in the world of $\mathbb{K}$-relations (usually we will have $\mathbb{K} = \mathbb{N}$): a $\mathbb{K}$-relation $\theta\colon A^* \to B^*$ is a $\mathbb{K}$-*function* if its *support* is a function from A^* to B^*; that is, if the image $f\theta$ of a word f in A^*,

cf. Def. IV.2.4, p. 552

$$f\theta = \langle\widehat{\theta}, f\rangle = \sum_{g\in B^*} \langle\widehat{\theta}, (f,g)\rangle g \quad ,$$

is a *monomial*; that is, if, for every f, *at most one* coefficient $\langle\widehat{\theta}, (f,g)\rangle$ is non-zero. We extend the notation $\|f\theta\|$ to series in a consistent manner: $\|f\theta\|$ is the number of elements of the *support* of $f\theta$ (that is, $\|f\theta\| = \|\mathsf{supp}\, f\theta\|$), and we thus find that θ is a ($\mathbb{K}$-)function if, for all f in A^*, $\|f\theta\| \leqslant 1$.

Functions are a special sort of relation and a ($\mathbb{B}$-)function from A^* to B^* is *rational unambiguous* if its characteristic series is an $\mathbb{N}$-rational series of $\mathbb{N}\langle\langle A^*\times B^*\rangle\rangle$. Contrary to rational relations in general, every rational function is unambiguous: this will be one of the principal results of this chapter; but we should not get ahead of ourselves.

cf. Def. IV.3.2, p. 558

cf. Cor. 2.1, p. 665

The rational ($\mathbb{B}$-)functions, and more generally the rational $\mathbb{K}$-functions when $\mathbb{K}$ is a positive semiring, are characterised by their representations:

Property 1.1 *Let (λ, μ, ν) be an arbitrary* trim *representation, of dimension Q, of a rational relation $\theta\colon A^* \to B^*$. If θ is a function, then all the non-zero entries of the matrices $f\mu$, for all f in A^*, (and of λ and ν) are monomials; that is,*

$$\forall p, q \in Q\,,\ \forall f \in A^* \qquad \|\lambda_q\| \leqslant 1\,, \quad \|\nu_p\| \leqslant 1\,, \quad \|f\mu_{p,q}\| \leqslant 1 \quad . \tag{1.1}$$

Proof. For all p and q in Q, there exist i and t in Q such that λ_i and ν_t are non-zero, and there exist g and h in A^* such that $g\mu_{i,p}$ and $h\mu_{q,t}$ are non-zero: $g\,f\,h$ belongs to the domain of θ. If $f\mu_{p,q}$ contained two distinct words, the image of $g\,f\,h$ would also contain (at least) two and θ would not be a function. ∎

However, since the property must hold for all matrices $f\mu$, this characterisation is not in itself effective and does not enable us to decide the functionality of a relation, which we shall now do.

Convention For the rest of this chapter we shall consider only *real-time* transducers, which are the automaton equivalent of matrix representations of relations, and we shall call them simply *transducers*.

1.1.1 An effective characterisation of functionality

A transducer $\mathcal{T} = \langle Q, A, B^*, E, I, T\rangle$ is *not functional* if and only if there exist in $\mathcal{T}$ two *distinct* successful computations

$$\begin{aligned} c' &:= q'_0 \xrightarrow[\mathcal{T}]{a_1|u'_1} q'_1 \xrightarrow[\mathcal{T}]{a_2|u'_2} \cdots \xrightarrow[\mathcal{T}]{a_n|u'_n} q'_n \\ \text{and} \qquad c'' &:= q''_0 \xrightarrow[\mathcal{T}]{a_1|u''_1} q''_1 \xrightarrow[\mathcal{T}]{a_2|u''_2} \cdots \xrightarrow[\mathcal{T}]{a_n|u''_n} q''_n \ , \end{aligned} \tag{1.2}$$

with $u'_1u'_2\cdots u'_n \neq u''_1u''_2\cdots u''_n$. There exists at least one j such that $u'_j \neq u''_j$ and, as we suppose, by the above property, that the representation of $\mathcal{T}$ has monomial entries, at least one j such that $q'_j \neq q''_j$. This implies in particular that, by ignoring the second component of labels, the automaton $\mathcal{A}$ underlying $\mathcal{T}$ is *ambiguous*. We shall reuse and extend the method of the *cartesian square of an automaton* which enables us to decide whether an automaton is ambiguous. *cf. § I.1.4.1, p. 74*

By definition, the *(cartesian) product* of $\mathcal{T}$ by itself (the square of $\mathcal{T}$) is the transducer $\mathcal{T}\times\mathcal{T}$ from A^* to $B^*\times B^*$:

$$\mathcal{T}\times\mathcal{T} = \langle\, Q\times Q, A, B^*\times B^*, F, I\times I, T\times T\,\rangle \ ,$$

whose transition set F is defined by

$$F = \{\big((p,r),(a,(u',u''))\big),(q,s)\big) \mid (p,(a,u'),q) \quad \text{and} \quad (r,(a,u''),s)\in E\} \ .$$

In particular, the underlying input automaton of $\mathcal{T}\times\mathcal{T}$ is the cartesian square of the underlying input automaton $\mathcal{A}$ of $\mathcal{T}$. If $\mathcal{A}$ is unambiguous, $\mathcal{T}$ is clearly functional and we have seen that in this case the trim part of $\mathcal{A}\times\mathcal{A}$, or of $\mathcal{T}\times\mathcal{T}$, is reduced to its diagonal. *cf. Lem. I.1.16, p. 75*

To decide whether $\mathcal{T}$ is functional when $\mathcal{A}$ is ambiguous, we shall extend $\mathcal{T}\times\mathcal{T}$ with a *valuation* by forming its product with a suitable action. The valuation describes the conditions under which two words such that $u'_1u'_2\cdots u'_n$ and $u''_1u''_2\cdots u''_n$ are equal or not and, more precisely, what minimal information we must preserve at each step i to be able to reach a conclusion at the final step n.

Let B^* be the output monoid of $\mathcal{T}$, and H_B the set defined by *cf. Exer. II.2.7, p. 237 and Exer. II.6.16, p. 323*

$$H_B = (B^*\times 1_{B^*}) \cup (1_{B^*}\times B^*) \cup \{\mathbf{0}\} \ ,$$

where $\mathbf{0}$ is a new symbol. We shall define an action of $B^*\times B^*$ on H_B in the following manner. We write $u \leqslant v$ if u is a prefix of v, and start by defining a map ψ from $B^*\times B^*$ to H_B by

$$\forall u,v\in B^* \qquad (u,v)\psi = \begin{cases} (v^{-1}u, 1_{B^*}) & \text{if} \quad v\leqslant u\,, \\ (1_{B^*}, u^{-1}v) & \text{if} \quad u\leqslant v\,, \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Intuitively, $(u,v)\psi$ is either the 'lead' $(v^{-1}u, 1_{B^*})$ of the first component u over the second, or its 'delay' $(1_{B^*}, u^{-1}v)$, or $\mathbf{0}$, the observation that u and v are not prefixes of the same word and cannot therefore be 'completed' to give the same output. We then verify:

Lemma 1.1 *The map ω_B from $H_B\times(B^*\times B^*)$ to*[1] *H_B, defined by*

$$\forall (f,g)\in H_B\setminus\mathbf{0}\,,\ \forall (u,v)\in B^*\times B^* \qquad ((f,g),(u,v))\omega_B = (f\,u, g\,v)\psi$$
$$\textit{and} \qquad (\mathbf{0},(u,v))\omega_B = \mathbf{0}$$

is an action (of $B^*\times B^*$ on H_B) which we shall call the Lead or Delay action *(relative to B^*) and which we shall write from now on as a dot.* ■

[1] We shall not forget that if $(f,g)\in H_B\setminus\mathbf{0}$, at least one of f and g is equal to 1_{B^*}.

We verify in particular that $(u,v)\psi = (1_{B^*}, 1_{B^*})$ if and only if $u = v$; that is:

Property 1.2 *The stabiliser of* $(1_{B^*}, 1_{B^*})$ *for* ω_B *is the diagonal of* $B^* \times B^*$ *:*

$$(1_{B^*}, 1_{B^*}) \cdot (u,v) = (1_{B^*}, 1_{B^*}) \quad \Longleftrightarrow \quad u = v \ . \tag{1.3}$$

cf. Exer. II.2.7, p. 237

We also verify that the transition monoid of this action is isomorphic to $B^* \times B^*$ if B has at least two letters, to $\mathbb{Z}$ if there is only one (to avoid confusion with 0, the identity element of $\mathbb{Z}$, we write $\mathbf{0}$ for the 'zero' of H_B).

In what follows, we implicitly consider ω_B also to be an action of $A^* \times (B^* \times B^*)$ on H_B by setting $h \cdot (a,(u,v)) = h \cdot (u,v)$ for all a; that is, by ignoring the first component. Recall that we say an action δ is a *valuation* of an automaton $\mathcal{A}$ if the product $\mathcal{A} \times \delta$ is in bijection with $\mathcal{A}$. We can now state the promised characterisation.

cf. Rem. II.3.7, p. 268

Proposition 1.2 *A transducer* $\mathcal{T}$ *from* A^* *to* B^* *is functional if and only if the Lead or Delay action* ω_B *is a* valuation *of the trim part* $\mathcal{U}$ *of* $\mathcal{T} \times \mathcal{T}$ *such that the value of each final state (of* $\mathcal{U} \times \omega_B$ *) is* $(1_{B^*}, 1_{B^*})$ *.*

Example 1.2 (continued) Figure 1.2 shows the product of the cartesian square of the transducer $\mathcal{G}_1$ from Figure 1.1 by the Lead or Delay action relative to x^*. In this case we can identify $H_{\{x\}}$ with $\mathbb{Z}$ and label the states of the square by the corresponding integers. The figure is already crowded and we could not easily add the labels of the transitions without losing legibility. The input label is always a and remains implicit. Instead of an output label of the form (x^n, x^m) we just indicate the quantity $n - m$, symbolised by the style of the arrow: 0 by a dotted arrow, 1 by a normal arrow, 2 by a bold arrow, 3 by a double arrow, and negative values by the same types of arrow, but dashed. As the product is a valuation and every final state has value 0, γ_1 is functional. □

Proof of Proposition 1.2. The condition is sufficient. Write ν for the valuation defined on $\mathcal{U}$ by the Lead or Delay action: for each state (p,q) in $\mathcal{U}$, $(p,q)\nu$ is an element of H_B, and let, as above, c' and c'' be successful computations of $\mathcal{T}$:

$$\begin{aligned} c' &:= q'_0 \xrightarrow[\mathcal{T}]{a_1|u'_1} q'_1 \xrightarrow[\mathcal{T}]{a_2|u'_2} \cdots \xrightarrow[\mathcal{T}]{a_n|u'_n} q'_n \\ \text{and} \qquad c'' &:= q''_0 \xrightarrow[\mathcal{T}]{a_1|u''_1} q''_1 \xrightarrow[\mathcal{T}]{a_2|u''_2} \cdots \xrightarrow[\mathcal{T}]{a_n|u''_n} q''_n \ . \end{aligned}$$

We have $(q'_0, q''_0)\nu = (1_{B^*}, 1_{B^*})$, $(q'_0, q''_0)\nu \cdot (u'_1 \cdots u'_i, u''_1 \cdots u''_i) = (q'_i, q''_i)\nu$ for each i in $[n]$ and hence in particular:

$$(q'_0, q''_0)\nu \cdot (u'_1 \cdots u'_n, u''_1 \cdots u''_n) = (q'_n, q''_n)\nu = (1_{B^*}, 1_{B^*}) \ ,$$

since (q'_n, q''_n) is a final state of $\mathcal{T} \times \mathcal{T}$. Thus, by (1.3), $u'_1 \cdots u'_n = u''_1 \cdots u''_n$ and $\mathcal{T}$ is functional.

The condition is necessary. There are two possible cases.

(a) The product of $\mathcal{U}$ by ω_B is a valuation but there exists a final state (r', r'') of $\mathcal{U}$ whose valuation is different from $(1_{B^*}, 1_{B^*})$. Therefore there is a successful computation:

$$(i', i'') \xrightarrow[\mathcal{T}\times\mathcal{T}]{f|(u',u'')} (r', r'')$$

and we have: $(1_{B^*}, 1_{B^*}) \cdot (u', u'') \neq (1_{B^*}, 1_{B^*})$. Still by (1.3), $u' \neq u''$ and $\mathcal{T}$ is not functional.

(b) The product of $\mathcal{U}$ by ω_B is not a valuation. There must then exist two successful computations of the form:

$$(i', i'') \xrightarrow[\mathcal{T}\times\mathcal{T}]{f_1|(u_1',u_1'')} (p', p'') \xrightarrow[\mathcal{T}\times\mathcal{T}]{f_2|(u_2',u_2'')} (r', r'')$$

and $$(j', j'') \xrightarrow[\mathcal{T}\times\mathcal{T}]{g_1|(v_1',v_1'')} (p', p'') \xrightarrow[\mathcal{T}\times\mathcal{T}]{f_2|(u_2',u_2'')} (r', r'') \ ,$$

with $(1_{B^*}, 1_{B^*}) \cdot (u_1', u_1'') \neq (1_{B^*}, 1_{B^*}) \cdot (v_1', v_1'')$. We cannot simultaneously have $u_1' u_2' = u_1'' u_2''$ and $v_1' u_2' = v_1'' u_2''$, and $\mathcal{T}$ is not functional. ■

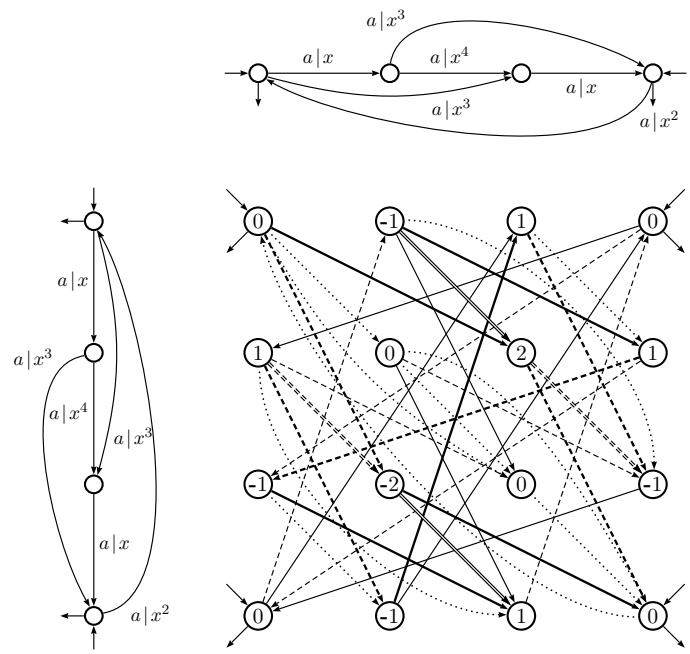


Figure 1.2: The square of $\mathcal{G}_1$ is valued by ω_B (with $B = \{x\}$) and every final state has value 0: γ_1 is functional

The construction of Figure 1.2 is thus in itself a proof that the transducer $\mathcal{G}_1$ is functional. More generally, the computations of the square $\mathcal{T}\times\mathcal{T}$, of its trim part $\mathcal{U}$, and of the product of $\mathcal{U}$ and ω_B are effective: their complexity is proportional to the square of the number of transitions of $\mathcal{T}$, as long as the computations on the words are counted as having a fixed cost, independent of the length of the words. We can therefore state:

Theorem 1.1 *It is decidable whether a transducer $\mathcal{T}$ (finite, with m transitions) realises a functional relation. The complexity of the decision procedure is $O(m^2)$.* ■

Remark 1.1 If $\mathcal{T}$ is a transducer (*real-time*, do not forget) from A^* to B^* which realises a relation α, the automaton obtained from $\mathcal{T}\times\mathcal{T}$ by ignoring the first component is an automaton over $B^*\times B^*$ whose behaviour is equal to $(A^*)\alpha\odot\alpha$, and hence is the graph of the relation $\alpha^{-1}\alpha$. The relation α is a function if and only if $\alpha^{-1}\alpha$ is the restriction of the identity to $\mathsf{Im}\,\alpha$ and the verification of this last property comes down exactly to the proof of Proposition 1.2. □

cf. Prop. IV.2.5, p. 554

cf. Exer. 0.1.6, p. 14

1.1.2 Equivalence of rational functions

Given two functions $\theta\colon A^*\to B^*$ and $\sigma\colon A^*\to B^*$, we say that θ *is contained* in σ ($\theta\subseteq\sigma$), if, as for relations, the graph of θ is included in that of σ ($\widehat{\theta}\subseteq\widehat{\sigma}$). Since σ is a function, this is the same as saying that, for all f in A^*, if $f\theta$ is defined, then $f\theta=f\sigma$. Recall also that $\theta\cup\sigma$ is the relation whose graph is the union of the graphs of θ and σ; that is, $\widehat{\theta\cup\sigma}=\widehat{\theta}\cup\widehat{\sigma}$. It is obvious that $\theta\subseteq\sigma$ if $\mathsf{Dom}\,\theta\subseteq\mathsf{Dom}\,\sigma$ *and* if $\theta\cup\sigma$ is a function. Theorem 1.1 therefore implies:

And Th. I.3.1, p. 110

Corollary 1.3 *The inclusion and hence the equivalence of rational functions is decidable.* ■

Note, by contrast, that the formulation of Post's Theorem which we have used is the same as saying that it is undecidable in general whether the intersection of the graphs of two rational functions θ and σ is empty; that is, whether there exists a word f in A^* such that $f\theta=f\sigma$.

cf. Th. 0.8.2, p. 37 and p. 568

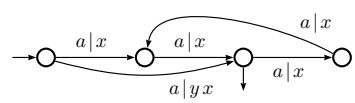
The construction underlying the proof of the corollary, as in the example γ_1 of Figure 1.1, brings us back to the distinction between *functionality* and *unambiguity*. A relation can be not functional and yet remain unambiguous: every rational substitution, hence every inverse rational substitution, every recognisable relation, and every synchronous relation is unambiguous, or rather, can be made unambiguous. Conversely, a rational function θ can well be ambiguous; that is, some pairs $(f,f\theta)$ can be the label of several computations in the transducer that defines θ: for instance, (a^3,x^6) is the label of five computations in $\mathcal{G}_1$, so, if $\theta\cup\sigma$ is a function, every word f in $\mathsf{Dom}\,\theta\cap\mathsf{Dom}\,\sigma$ will have a single image under $\theta\cup\sigma$; but this image will have multiplicity 2 in the transducer associated with $\theta\cup\sigma$ from the transducers which realise θ and σ (supposing that they are unambiguous).

This necessary distinction gives its meaning and importance to the result we shall soon see that ensures that every rational function is *equivalent* (as a $\mathbb{B}$-relation) to an unambiguous rational function.

cf. Cor. 2.1, p. 665

Exercises

- 1.1 Describe the function γ_1 from Example 1.2.
- 1.2 Describe the relation realised by the transducer $\mathcal{G}_2$ in the adjacent figure.



• 1.3 Let $\theta\colon A^* \to B^*$ be a rational relation realised by a trim representation μ of dimension Q. Show directly (that is, without explicitly using the cartesian square of the transducer which corresponds to μ) that if θ is not a function, then there exists a word f in A^*, of length less than $1 + 2\,\|Q\|\,(\|Q\| - 1)$, such that the coefficients of $f\mu$ are not all monomial.

Conclude and comment.

1.4 Show that it is decidable whether a rational relation reduces to:

(a) an intersection with a rational set; (b) a morphism.

• 1.5 Show that it is decidable whether a rational function is recognisable.

1.6 Show that it is decidable whether a rational function is synchronous.

1.2 Sequential functions

The most important family of rational functions, which is so because it corresponds to the most natural model of a physical computer as much as because it is theoretically important, is that of functions realised by a *deterministic* finite automaton producing an output at each transition. Theoretic efficiency calls for a concept that differs slightly from this down-to-earth intuition, *and from common usage*.

We shall content ourselves here and in the following subsection with giving some definitions and examples which illustrate them, and some elementary properties which we can derive directly. This will allow us to describe their role in the set of rational functions. Their study as such is left until Section 4. *cf. p. 692*

1.2.1 Some unconventional terminology

Definition 1.1 We shall say that a rational function is *sequential* if it has a *row-monomial* representation. □

More explicitly, $\alpha\colon A^* \to B^*$ is sequential if there exists a representation (λ, μ, ν) of α such that:

(i) the row vector λ has a single non-zero entry;

(ii) for each a in A, the matrix $a\mu$ is row monomial; that is, each row of $a\mu$ has at most one non-zero entry;

(iii) the non-zero entries of λ, $a\mu$ and ν are words of B^* (or monomials of $\mathbb{K}\langle B^*\rangle$).

In other words, $\alpha\colon A^* \to B^*$ is sequential if it is realised by a *real-time* transducer whose underlying input automaton is *deterministic* and whose outputs, and initial and final functions, take values in B^*.

Remark 1.2 I am aware that this terminology risks creating confusion, particularly for the reader already familiar with the subject, since I call *sequential* that which is commonly called (following Schützenberger[2]) *sub-sequential* and what I call *pure*

[2] In [226]. Schützenberger defines more precisely *left* and *right* functions; for the left–right distinction, see below.

sequential is usually called *sequential*. The developments that follow in this chapter will show, I think, that the *fundamental object* is indeed that which is called here a *sequential function*, and that it thus merits the basic term (see also the bibliographic notes). □

Remark 1.3 In the preceding definition, the output monoid B^* plays no part and could be replaced by any monoid M, *not necessarily free*.

On the other hand, the nature of the output monoid M is essential if we want to know *whether it is decidable* if a relation $\alpha\colon A^* \to M$ is a sequential function. □

cf. Exam. IV.1.6, p. 530

Example 1.3 Division by 3. We saw in the previous chapter that the transducer $\mathcal{Q}_2$, shown again in Figure 1.3, realises the division by 3 of integers (divisible by 3). This transducer is sequential, and its matrix representation is

$$\lambda_1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, \quad a\mu_1 = \begin{pmatrix} a & 0 & 0 \\ 0 & 0 & a \\ 0 & b & 0 \end{pmatrix}, \quad b\mu_1 = \begin{pmatrix} 0 & a & 0 \\ b & 0 & 0 \\ 0 & 0 & b \end{pmatrix}, \quad \nu_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}. \qquad \square$$

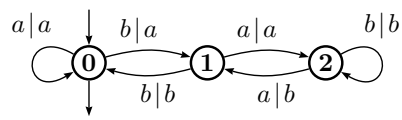


Figure 1.3: The transducer $\mathcal{Q}_2$, divider by 3 (of multiples of 3)

cf. p. 607

Example 1.4 Let us return to the transducer $\mathcal{D}_2$ of Figure IV.6.3 (a), shown again in Figure 1.4. The corresponding matrix representation is row monomial:

$$\lambda_2 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, \quad a\mu_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & a & 0 \\ 0 & ab & 0 \end{pmatrix}, \quad b\mu_2 = \begin{pmatrix} b & 0 & 0 \\ 0 & 0 & 1 \\ 0 & ba & 0 \end{pmatrix}, \quad \nu_2 = \begin{pmatrix} 1 \\ a \\ ab \end{pmatrix}.$$

This automaton realises the function from $\{a,b\}^*$ to itself which replaces the factors abb of each word of A^* by factors baa and which achieves this substitution[3] by reading the words *from left to right*, the last a of a substituted factor being able to combine with a succeeding factor bb to give rise to a new substitution. Thus, $(ababbbb)|\mathcal{D}_2| = abbabaa$. □

cf. Exer. IV.6.7, p. 608

Example 1.5 A morphism (from A^* to B^*) and the intersection with a rational of A^* are sequential functions: a representation of dimension 1 is obviously monomial, and if (λ, μ, ν) is the boolean representation of a *deterministic* automaton which recognises K, (λ, μ', ν), with $a\mu' = (a\mu)a$ (that is, the non-zero entries of $a\mu'$ are equal to a and $a\mu'$ and $a\mu$ have the same support) is a (row-monomial) representation of ι_K. □

cf. Exam. IV.1.3, p. 527

[3]Called *Fibonacci substitution*, as it corresponds to numerical equivalence in the number system defined by the Fibonacci sequence (when a is interpreted as the digit 0 and b as the digit 1).

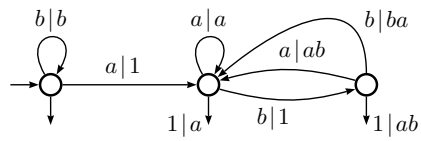


Figure 1.4: The transducer $\mathcal{D}_2$

The matrix representation is certainly the 'right object' for defining sequential functions. However, just as the functional point of view is well suited to dealing with deterministic automata, the definition from a deterministic automaton with output, also a functional definition, is often more expressive, and leads to simpler notation in some cases. *cf. Sec. I.3, p. 101*

Definition 1.2 A *sequential transducer* (from A^* to B^*) is a septuple

$$\mathcal{T} = \langle Q, A, B^*, \delta, \eta, i, T \rangle \ , \qquad \text{where:}$$

(i) the underlying input automaton $\langle Q, A, \delta, i, T \rangle$ is a deterministic automaton over A^*. The transition function δ is a partial function from $Q \times A$ to Q and we naturally write $p \cdot a$ for $(p, a)\delta$;

(ii) the *output function* η is a partial function from $Q \times A$ to B^*, which has the *same domain* as δ. Usually we write $p * a$ for $(p, a)\eta$;

(iii) the *final function* T is a function from Q to B^*. We consider[4] that there is also an initial function I, from Q to B^*, whose domain is the unique initial state i (*cf.* Equation (1.6)). □

In the same way that the function δ is extended to a function from $Q \times A^*$ to Q by $q \cdot fa = (q \cdot f) \cdot a$, η is extended to a function from $Q \times A^*$ to B^* by

$$q * 1_{A^*} = 1_{B^*} \qquad \text{and} \qquad q * fa = (q * f)\,[(q \cdot f) * a] \ , \tag{1.4}$$

from which we deduce, by induction on the length of g, that

$$\forall f, g \in A^* \qquad q * f\,g = (q * f)\,[(q \cdot f) * g] \ . \tag{1.5}$$

The image of a word f under $\mathcal{T}$ is then

$$f|\mathcal{T}| = I_i\,(i * f)\,T_{i \cdot f} \ . \tag{1.6}$$

We verify easily:

Property 1.3 *A function is sequential if and only if it is realised by a sequential transducer.* ■

Remark 1.4 The fact that the initial vector λ of a monomial representation (λ, μ, ν) takes values in B^*, and is not simply a boolean vector, or, which comes to the same thing, that a sequential transducer should have an *initial function* and

[4] A slight abuse of the notation since I does not appear in the definition of $\mathcal{T}$.

not simply an *initial state*, is a slight extension of the usual definition.[5] It is justified by the simplification of certain constructions and above all in the *minimisation* process of sequential transducers (see Section 4). □

cf. Exer. 1.10

Remark 1.5 Following Remark 1.3, we can define sequential transducers from A^* to an arbitrary monoid M with identical notation:

$$\mathcal{T} = \langle\, Q, A, M, \delta, \eta, i, T \,\rangle \ . \qquad \square$$

1.2.2 Dual definitions

Definition 1.3 A rational function is *co-sequential* if it has a *column-monomial* representation. □

More explicitly, $\alpha\colon A^* \to B^*$ is co-sequential[6] if there exists a representation (λ, μ, ν) of α such that:

(i) the row vector ν has a single non-zero entry;

(ii) for each a in A, the matrix $a\mu$ is column monomial;

(iii) the non-zero entries of λ, $a\mu$ and ν are words of B^* (or monomials of $\mathbb{K}\langle B^*\rangle$).

In other words, $\alpha\colon A^* \to B^*$ is co-sequential if it is realised by a *real-time* left transducer whose underlying input automaton is *co-deterministic*, or, which is equivalent, by a *real-time* right transducer whose underlying input automaton is *deterministic*, and whose outputs and the initial and final functions take values in B^*.

Remark 1.6 I continue to flout established terminology, and here is why: as I said in the remark following the very first proposition of this work, I propose to reserve the qualifiers 'left' and 'right' *for automata* (and hence for transducers) which model physical machines, and according to as they read words *from left to right* or *from right to left* respectively. Functions are neither left nor right, but are as they are, and are realised by transducers which can be left or right, with dual properties. □

cf. Rem. I.1.5, p. 58

Example 1.6 Addition in base 2. The addition of two numbers is most obviously performed by a machine with three tapes: one for each of the operands and one for the result. In fact, we can first add the operands digit by digit – this operation can obviously be performed by a very simple one-state three-tape automaton – which gives a number written in the alphabet $\{0, 1, 2\}$. The operation of addition then comes down to rewriting this word in the alphabet $\{0, 1\}$ while preserving the numerical value.

For example, $1101 + 101$ (that is, $13 + 5 = 18$) is first written 1202 (that is, $18 = 2^3 + 2 \cdot 2^2 + 2 \cdot 2^2$) which must be rewritten as 10010 (that is, $18 = 2^4 + 2^1$).

In Figure 1.5 (b) we show the co-sequential transducer which realises this *normalisation* on the alphabet $\{0, 1, 2\}$. To be clearer, we abandon from now on the encoding

[5] Of sub-sequential functions, of course (*cf.* Remark 1.2).

[6] In the usual terminology: *right sub-sequential function.*

of digits as letters which we have used up to now (for the same reason we will not give the matrix representation of this example). In fact, it is not easy to understand the behaviour of a co-sequential (left) transducer (or even, of a co-deterministic (left) transducer); that is why we prefer to start by constructing a *right* sequential transducer – shown in Figure 1.5 (a) – which therefore reads and writes words *from right to left*.[7] It is then enough to reverse all the arrows, including the initial and final arrows, to obtain the transducer we want. □

cf. Note I.2, p. 52

(a) the right transducer (b) the left transducer

Figure 1.5: Two transducers for addition in base 2

Operations on the written forms of numbers, addition and various normalisations, multiplication and division by fixed integers, are an almost inexhaustible field of examples of sequential and co-sequential functions; we shall see some of them in the exercises.

Example 1.4 (continued) Let us look again at the substitution which consists of replacing factors abb by baa. In itself it is not directed, but by proceeding *from right to left*, the last b of a factor substituted can combine with a succeeding factor ab to give rise to another substitution. Thus, $ababbbb$ gives $baaaabb$.

The co-sequential transducer $\mathcal{E}_2$ which performs this transformation is shown in Figure 1.6 (b). To construct it following the above method, we first construct the *right* sequential transducer (Figure 1.6 (a)) which realises the substitution 'bba gives aab' since it reads the words from right to left, then we reverse the arrows. We note that we also transpose the labels (an operation that is noticeable only in the transition $q \xrightarrow{a|ba} r$).

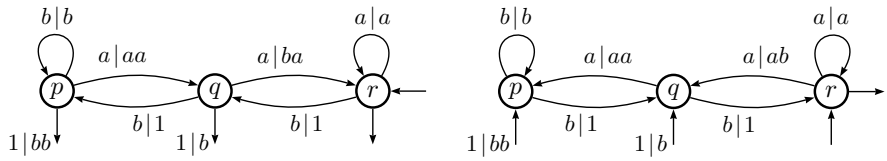
cf. Exer. III.1.25, p. 544

The matrix representation of $\mathcal{E}_2$ is

$$\eta_2 = \begin{pmatrix} bb & b & 1 \end{pmatrix}, \quad a\kappa_2 = \begin{pmatrix} 0 & 0 & 0 \\ aa & 0 & 0 \\ 0 & ab & a \end{pmatrix}, \quad b\kappa_2 = \begin{pmatrix} b & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \xi_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \quad \square$$

Example 1.5 (continued) A morphism (from A^* to B^*) and the intersection with a rational of A^* are co-sequential functions: a representation of dimension 1 is obviously column as well as row monomial, and if (λ, μ, ν) is the boolean representation of a *co-deterministic* automaton which recognises K, then (λ, μ', ν), with $a\mu' = (a\mu)a$, is a column-monomial representation of ι_K. □

[7] The reader should not be unduly disconcerted by this change in the direction of reading, since this is the way in which one is accustomed to perform addition.



(a) the right transducer (b) the left transducer $\mathcal{E}_2$

Figure 1.6: The substitution '$a\,b\,b$ becomes $b\,a\,a$' proceeding from right to left

1.2.3 Composition

Since the composition product of two row-monomial (resp. column-monomial) matrices is obviously a row-monomial (resp. column-monomial) matrix, and a representation of the composition of two relations is given by the composition product of the representations, it follows that:

cf. Th. IV.3.6, p. 567

Proposition 1.4 *The composition of two sequential (resp. co-sequential) functions is a sequential (resp. co-sequential) function.* ■

We will see a number of examples in the exercises. The problem of deciding whether a rational function is sequential (mentioned in Remark 1.3) will be solved in Section 4, after the proof of some characteristic properties of sequential functions.

Exercises

• 1.7 **Normalised division by 3.** Transform the transducer $\mathcal{Q}_2$ of Figure 1.3 so that it both computes the integer quotient by 3, even for numbers that are not multiples of 3, and writes the answer conventionally: no leading a (=0), unless the quotient is zero, in which case the result should be a and not 1_{A^*}.

1.8 **Division by n.** Let k be a fixed base in which integers are written. As in Section IV.4.1, we write $\mathbf{k} = \{0, 1, \ldots, k-1\}$. Show that division by a fixed integer n is a sequential function from $\mathbf{k}^*$ to itself.

• 1.9 **Sequentiality and recognisability.** Prove:

Proposition 1.5 *A recognisable function is sequential (and co-sequential).*

• 1.10 Let $\alpha\colon A^* \to B^*$ be a sequential function and u a word in B^*. Show that the function α' defined by $f\alpha' = u\,(f\alpha)$ is sequential.

• 1.11 **Codes with finite deciphering delay.** (a) Let $\beta_1\colon \{x, y\}^* \to A^*$ be the morphism defined by $x\beta_1 = a$ and $y\beta_1 = a\,b\,a$. Show that β_1^{-1} is a sequential function.

(b) Idem for $\beta_2\colon \{x, y, z\}^* \to A^*$ defined by $x\beta_2 = a\,b$, $y\beta_2 = a\,b\,b$ and $z\beta_2 = b\,a\,a\,b$.

(c) Generalisation. Let $\beta\colon B^* \to A^*$ be a morphism. By definition, $X = B\beta$ is a *code* if β is injective. The code X is said to have *deciphering delay* d if when a word $f = x_1\,x_2 \cdots x_{d+1}$ of X^{d+1} is a *prefix* of a word $g = y_1\,y_2 \cdots y_r$ of X^* then $x_1 = y_1$. (In particular, a code is prefix if its deciphering delay is 0.) Prove:

Proposition 1.6 *Let $\beta\colon B^* \to A^*$ be a morphism. The set $X = B\beta$ is a code with finite deciphering delay if and only if β^{-1} is a sequential function.*

1.12 Verify Equation (1.5).

1.13 Verify Property 1.3.

• 1.14 In imitation of Example 1.6, give a co-sequential (left) transducer for addition in base 3.

• 1.15 **Multiplication, base 2.** Starting from Example 1.3, show that multiplication by 3 of numbers written in base 2 is a co-sequential function.

• 1.16 **Multiplication, base 3.** Give a (co-sequential) transducer which realises multiplication by 2 of numbers written in base 3. Deduce a transducer which realises multiplication by 4.

• 1.17 **Normalisation.** Let k again be a fixed base, and h another integer, greater than k. Logically, we write $\mathbf{h} = \{0, 1, \ldots, h-1\}$. We make every word f in $\mathbf{h}^*$ correspond to the *numerical value* of f in base k, again written $\overline{f}$, and defined by

$$\overline{f} = \sum_{i=0}^{n} d_i\, k^i \qquad \text{if} \qquad f = d_n \cdots d_1 d_0 \quad \text{and} \quad d_i \in \mathbf{h} \ .$$

Normalisation in base k is the function

$$\nu_{h,k} \colon \mathbf{h}^* \quad \longrightarrow \quad \mathbf{k}^* \ ,$$

which preserves the numerical value; that is, $\overline{f\nu_{h,k}} = \overline{f}$ for all f in $\mathbf{h}^*$.

(a) Show that $\nu_{h,k}$ is a co-sequential function.

(b) Evaluate (as a function of k and h) the minimum number of states of a co-sequential transducer which realises $\nu_{h,k}$.

(c) Explain why this solves the problem of addition as well as that of multiplication by a fixed integer, in every base k.

• 1.18 **Avizienis addition.** In general the *Avizienis system*[8] is the process which consists of writing numbers in a base k using positive *and negative* figures taken from an alphabet[9] $\mathfrak{k} = \{\bar{h}, \ldots, \bar{1}, 0, 1, \ldots, h\}$, where h is the smallest integer greater than or equal to $(k+1)/2$, a representation which is not in general unique.

For example, in base 2, $11 = 10\bar{1} = 1\bar{1}1$; in base 3, $21 = \bar{1}02$; in base 10, $27 = 3\bar{3}$.

Normalisation is still the function

$$\nu_k \colon \mathfrak{k}^* \quad \longrightarrow \quad \mathbf{k}^* \ ,$$

which preserves numerical value.

(a) Show that this normalisation is a co-sequential function. Give the corresponding right sequential transducers for $k = 2$ and $k = 3$.

(b) Show that it is possible to choose an inverse normalisation function, a 'translation' from the standard system to the Avizienis system, which is sequential. Give the corresponding sequential transducers for $k = 2$ and $k = 3$.

---

[8] See for example [174]. This system and its remarkable properties for performing arithmetic computations had already been described by Cauchy in two notes with suggestive titles: 'On ways to avoid error in numerical calculations', and 'On ways to verify or simplify diverse operations in decimal arithmetic', *C. R. Acad. Sci.*, t. XI (1840), *cf.* [53, § 105 and 106].

[9] A lower case gothic 'k'.

(c) In the Avizienis system, addition can take the form of a function α_k from $\mathfrak{k}'^*$ to $\mathfrak{k}^*$ where

$$\mathfrak{k}' = \{\overline{2h}, \ldots, \bar{1}, 0, 1, \ldots, 2h\} \ ,$$

(with h as above) and where α_k preserves the numerical value.

Show that α_k can be chosen in such a way that it is a *sequential*[10] function. Give the corresponding sequential transducers for $k = 2$ and $k = 3$.

• 1.19 Show that multiplication (of numbers written in base k) is not a rational function from $\mathbf{k}^* \times \mathbf{k}^*$ to $\mathbf{k}^*$. Show that squaring is not a rational function from $\{0,1\}^*$ to $\{0,1\}^*$.

1.3 Pure sequential functions

We now define a (strict) subfamily of the sequential functions (that is, the functions which *we* call sequential). Intuitively, these are the functions realised by sequential transducers which are always 'up to date' in their output, without needing to know whether the word read has ended or not; that is, transducers all of whose states are final and which do not need to 'add' anything when they have finished reading a word.

Definition 1.4 We say that a *sequential function* is *pure* if it has a row-monomial representation (λ, μ, ν) such that the entries of ν are all equal to 1_{B^*}.

We say that a *co-sequential function* is *pure* if it has a column-monomial representation (λ, μ, ν) such that the entries of λ are all equal to 1_{B^*}. □

In other words, $\alpha\colon A^* \to B^*$ is pure sequential if it is realised by a sequential transducer $\mathcal{T} = \langle Q, A, B^*, \delta, \eta, i, T \rangle$ where the final function T takes the value 1_{B^*} on every state (which we shall therefore call a *pure sequential transducer*).[11] The function realised by $\mathcal{T}$ then simplifies to

cf. Equat. (1.6)

$$f|\mathcal{T}| = I_i\,(i * f) \ , \tag{1.7}$$

and usually, when the unique non-zero entry of I is 1_{B^*}, to

$$f|\mathcal{T}| = (i * f) \ . \tag{1.8}$$

Example 1.3 (continued) **Integer division by** 3. Integer division by 3 of numbers written in binary (with our usual conventions) is a pure sequential function: it is realised by the transducer from Figure 1.3 in which we have made all the states final (Figure 1.7). □

[10] And not *co-sequential*! This is a first wonderful property of the Avizienis system.

[11] Final warning: pure sequential transducers are what – until now! – the literature has called *sequential transducers*, *generalised sequential machines* or even *gsms*; pure sequential functions are what are called *sequential functions* or *gsm-mappings* (see Remark 1.2 and the bibliographic notes).

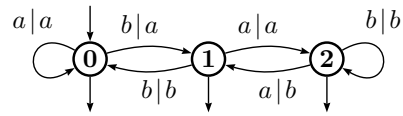


Figure 1.7: Transducer for integer division by 3

Example 1.5 (continued) A morphism (from A^* to B^*) is a pure sequential function. In general, intersection with a rational K of A^* is *not* a pure sequential function. More precisely, ι_K is a pure sequential function if and only if K is recognised by a deterministic automaton all of whose states are final, hence if and only if K is *prefix-closed* (that is, every prefix of a word of K is a word of K). □

The formulas for the composition product of representations: *cf. p. 567*

$$(\zeta, \pi, \omega) = (\lambda, \mu, \nu) \circ (\eta, \kappa, \chi) \ ,$$

and for the initial and final vectors:

$$\forall q \in Q \qquad \zeta_{(q,R)} = \eta \cdot (\lambda_q)\kappa \qquad \text{and} \qquad \omega_{(q,R)} = (\nu_q)\mu \cdot \chi \ , \tag{IV.3.5}$$

imply a specialisation of Proposition 1.4: *cf. Exer. IV.3.5, p. 568*

Proposition 1.7 *The (composition) product of two pure sequential (resp. pure co-sequential) functions is a pure sequential (resp. pure co-sequential) function.* ■

The resemblance to the terminology used for deterministic relations (deterministic and pure deterministic relations) is not accidental. Sequential transducers are (equivalent to) pure sequential transducers over an alphabet with an endmarker; or, alternatively, the final function of a transducer is the trace of an endmarker which has been removed. More formally, we have:

Lemma 1.8 *Let $\theta\colon A^* \to B^*$ be a function, \$ a symbol which does not belong to A, and $\tau\colon A^*_{\$} \to B^*$ the function defined by*

$$\forall f \in A^* \qquad (f)\theta = (f\$)\tau \ .$$

Then θ is sequential if and only if τ is pure sequential.

Proof. The condition is necessary: if τ is represented by (λ, μ, ν), then θ is represented by $(\lambda, \mu, (\$)\mu \cdot \nu)$. The condition is sufficient by the following construction (shown in Figure 1.8): to a sequential transducer $\mathcal{T}_\theta$ which recognises θ, we add a new state z and for each state q in $\mathcal{T}_\theta$ replace the final function T_q by the transition $q \xrightarrow{\$|T_q} z$. Verifying that this new transducer $\mathcal{T}_\tau$ has the required behaviour is immediate.

This construction can be expressed just as well on the matrix representation – we should not forget that sequential functions and pure sequential functions are defined

in that way. If (λ, μ, ν) is a monomial representation of θ, then τ is represented by $(\lambda', \upsilon', \nu')$ with

$$\lambda' = \begin{pmatrix} \boxed{\lambda} & 0 \end{pmatrix},$$

$$a\mu' = \begin{pmatrix} \boxed{a\mu} & \boxed{0} \\ \boxed{0} & 0 \end{pmatrix}, \quad \$\mu' = \begin{pmatrix} \boxed{0} & \boxed{\nu} \\ \boxed{0} & 0 \end{pmatrix} \quad \text{and} \quad \nu' = \begin{pmatrix} \boxed{1} \\ 1 \end{pmatrix}. \qquad \blacksquare$$

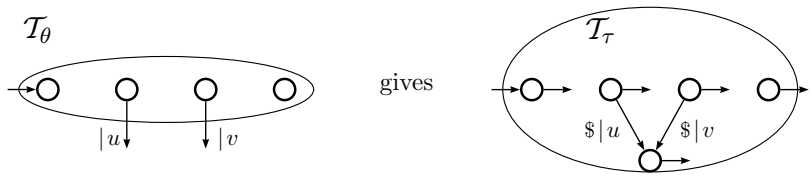


Figure 1.8: From a sequential transducer to a pure sequential transducer

Example 1.5 above suggests the following definition.

Definition 1.5 A function $\alpha\colon A^* \to B^*$ is *prefix-closed*[12] if, for all g and h in A^* such that $(gh)\alpha$ is defined, then $g\alpha$ is defined and is a prefix of $(gh)\alpha$; that is, with the usual notation for prefixes,

$$\forall f, g \in A^* \qquad g \leqslant f \quad \Longrightarrow \quad g\alpha \leqslant f\alpha \ . \tag{1.9}$$

Note that here too the monoid B^* plays no part and that we could say that a function α from A^* to an arbitrary monoid M is prefix-closed if (1.9) is verified, noting that $n \leqslant m$ in M if there exists n' such that $nn' = m$. The definition is therefore consistent with that of a prefix-closed set since $X \subseteq A^*$ is a prefix-closed if and only if χ_X, the characteristic function of X, is prefix-closed. Furthermore, the domain of a prefix-closed function is a prefix-closed set.

Clearly, *a pure sequential function is prefix-closed*. It is interesting to note that the converse is true, in the family of sequential[13] functions, which we shall show in Section 4. Now is also the time to note that the underlying input automaton of a pure sequential transducer is not necessarily *complete*; on the contrary, it is complete only if the function realised is defined on all of A^*.

cf. Prop. 4.11, p. 717

Exercise

- 1.20 Let (n, m) be a pair of integers written in base k. Suppose that the shorter number is padded on the left with zeros so that they are the same length. Verify that the function $\mathsf{max}(n, m)$ is a pure sequential function from $(\mathbf{k} \times \mathbf{k})^*$ to $\mathbf{k}^*$.

[12] The 'usual' terminology varies from *initial segment preserving* [80] to *preserves left factors* [27] to *prefix-preserving* [40].

[13] *Note added in translation:* But not in the family of rational functions.

1.4 Local functions

We could say that rational functions are functions that are computed in a finite amount of memory, apart from the input. This notion of the amount of memory necessary for the computation is clearer for sequential functions: we can imagine a device that receives or reads in sequence the input symbols; and emits or writes, again in sequence, the output symbols, the rules for turning one into the other being entirely contained in the finite set of states of the transducer, the *memory*.

An even more restricted family of functions is that where the only memory is a fixed-size part of the input. The usual way of representing this is to imagine a window of fixed size d which slides over the tape on which the input is written. To each 'configuration' of the window (that is, each length d factor of the input) there corresponds an output, a possibly empty word. The result of the function on a given word is the concatenation of the outputs when the window slides over the word from one end to the other. To take into account the information that corresponds to the start and to the finish of the word, the word is padded at each end with a sufficient number of special symbols. Such a device is illustrated in Figure 1.9.

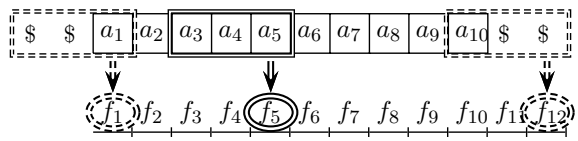


Figure 1.9: Sliding window of length 3

A function realised in this manner is called *local* and the integer d is its *degree*. The simplest example of such functions are morphisms, which are local functions of degree 1.

More formally, we start by defining the result of a *scan* of a word with a window of length d. As before, let \$ be a symbol which, by convention, belongs to no alphabet and which we call the *endmarker*. For each alphabet A, we write $A_\$$ for the set $A \cup \{\$\}$. The scan is a map σ_d from A^* to $\left[A_\$^d\right]^*$, which it is more sensible here to write $\mathrm{V}(A_\$^d)$, defined by

cf. Sec. IV.5.1, p. 577 and p. 659.

cf. Sec. II.5.1, p. 293 and Note 0.10, p. 18

$$
\begin{aligned}
(a_1\, a_2 \cdots a_n)\sigma_d &= \prod_{i=-d+2}^{i=n} (a_i\, a_{i+1} \cdots a_{i+d-1}) \qquad (1.10)\\
&= (a_{-d+1}\, a_{-d+2} \cdots a_1, a_{-d+2}\, a_{-d+3} \cdots a_3, \ldots, a_n\, a_{n+1} \cdots a_{n+d}) \quad ,
\end{aligned}
$$

with the convention that $a_j = \$$ if j is negative or greater than n. For example, by taking $d = 1$ then $d = 2$ and $d = 3$, we obtain

$$
\begin{aligned}
(a\,b\,a\,a\,b)\sigma_1 &= (a, b, a, a, b) \quad ,\\
(a\,b\,a\,a\,b)\sigma_2 &= (\$\,a, a\,b, b\,a, a\,a, a\,b, b\,\$) \quad ,\\
(a\,b\,a\,a\,b)\sigma_3 &= (\$\,\$\,a, \$\,a\,b, a\,b\,a, b\,a\,a, a\,a\,b, a\,b\,\$, b\,\$\,\$) \quad .
\end{aligned}
$$

Definition 1.6 A function $\alpha\colon A^* \to B^*$ is *local, of degree* d, if there exists a morphism $\overleftrightarrow{\alpha}\colon \mathrm{V}(A^d_{\$}) \to B^*$ such that α is the product of σ_d by $\overleftrightarrow{\alpha}$:

$$\alpha = \sigma_d \circ \overleftrightarrow{\alpha} \ . \qquad \square$$

It should be obvious that a local function of degree d is local of degree d' for any d' greater than d. A function is called simply *local* if there exists an integer d such that it is local of degree d.

Once more, the structure of the output monoid, the free monoid B^* in this case, plays no part in this definition and we can define local functions from a free monoid to any monoid. For example, the characteristic function of a local rational language is a local function of degree 2.

Remark 1.7 The definition of the scan of a word by a window can seem pointlessly cumbersome at first. It has the advantage of relating the definition of a local function to that of a *morphism* and avoiding the need to treat specially words that are shorter than or equal to the degree, which to be rigorous would otherwise have been necessary.

In practice, we can content ourselves with padding each word with a single symbol \$ on the left and right instead of specially aligning d. This is done implicitly in Figure 1.10. $\square$

Another way of interpreting the map σ_d is to view $(f)\sigma_d$ as the computation labelled f, which starts in the state which is the prefix of length $d-1$ of f, in the *automaton of factors* (of length $d-1$). From this we deduce the following proposition, shown in Figure 1.10.

Proposition 1.9 *A function $\alpha\colon A^* \to B^*$ is local (of degree d) if and only if it can be realised by a sequential transducer whose underlying input automaton is the* automaton of factors *of length $d-1$.* $\blacksquare$

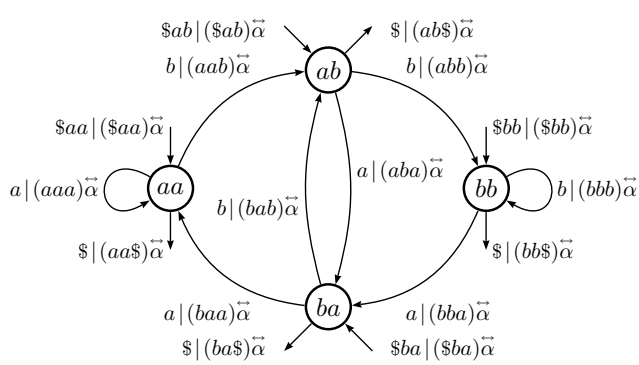


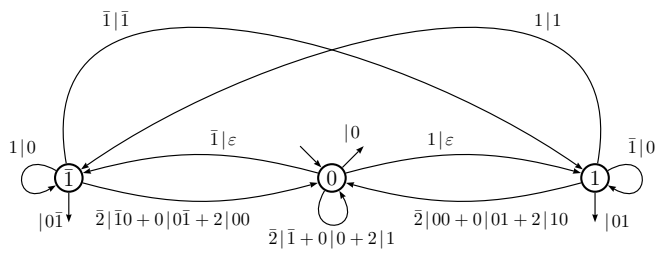
Figure 1.10: Transducer constructed on the automaton of factors of length 2

Recall that an automaton over A^* is *local* (of degree k) if the destination of each computation of length k depends only on its label, and not on its source. From Proposition 1.9, we then deduce:

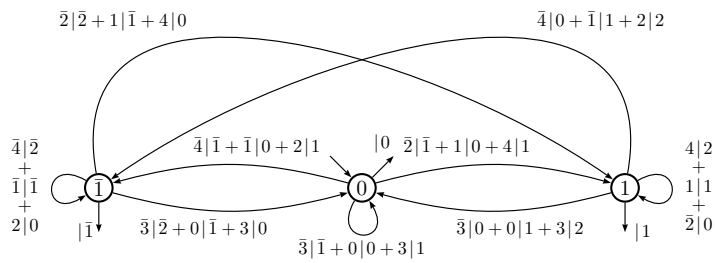
Corollary 1.10 *A function is local if and only if it is realised by a transducer whose underlying input automaton is local.* ■

The practical interest of local functions appears in two different forms, according to whether we interest ourselves in arbitrarily long inputs or on the contrary in fixed-length inputs. In the first case, locality ensures a sort of robustness: an error in the input, which is supposed to be sent via some channel of communication before being fed to the function, cannot propagate indefinitely along the output computed. In the second case, the fixed size of the input allows us to imagine a physical device which can calculate the image of each factor in parallel (especially if all these images have the same length). Certain arithmetic operations can be embedded in computers in this way.

Example 1.7 Avizienis adders. Let us reconsider the sequential transducers which implement addition in the Avizienis system; see Figure 1.11. We observe that the underlying input automata of these transducers are local, respectively of degree 2 for the base 3 adder and of degree 3 for the base 2 adder: these additions can be performed in parallel. □ *cf. Exer. 1.18*



(a) in base 2



(b) in base 3

Figure 1.11: Avizienis adders

We shall say that a local function α from A^* to B^* is *proper* if the restriction of the morphism $\overleftrightarrow{\alpha}$ to A^* is *continuous*; that is, if the image of every element of A^d under $\overleftrightarrow{\alpha}$ is different from 1_{B^*}. We easily verify:

Proposition 1.11 *The composition of a proper local function of degree d from A^* to B^* by a local function of degree d' from B^* to M is a local function of degree at most $d + d'$ from A^* to M.* ■

A local function is both sequential and co-sequential, but the converse is false: intersection with a rational X is sequential and co-sequential but is not local unless X is strictly locally testable.[14] In Section 4, we shall characterise the sequential functions that are local.

Exercises

• 1.21 (a) Prove Proposition 1.11.

(b) Show, with a simple example, that the assumption that the local function α is proper is necessary for the composition $\alpha \circ \beta$ with a local function β to be a local function.

• 1.22 **Soft addition.**[15] We choose a base $b \geqslant 3$ and write the numbers using the alphabet of digits $\mathfrak{b}'' = \{0, 1, \ldots, b-1, b, b+1\}$.

(a) Show that addition is a local function of degree 2 from $(\mathfrak{b}'' \times \mathfrak{b}'')^*$ to $\mathfrak{b}''^*$.

(b) Give a local transducer for the case of base 3.

(c) Can we generalise the result to base 2?

2 Uniformisation of rational relations

By its title, this section seems to refer to the last chapter; however, we are really dealing with functions here. If θ is an arbitrary relation from a set E to another set F, a function τ from E to F *uniformises*[16] θ (or is *a uniformisation* of θ) if, for every e in the domain of θ, $e\tau$ is an element of $e\theta$; that is,

$$\mathsf{Dom}\,\tau = \mathsf{Dom}\,\theta \qquad \text{and} \qquad \forall e \in \mathsf{Dom}\,\theta \qquad e\tau \in e\theta \ .$$

The uniformisation τ of a relation θ therefore consists of a *choice function*, repeated for each of the elements of the domain of θ, in the *selection* of an element $e\tau$ from each of the sets $e\theta$ (*cf.* Figure 2.1).

Examples 2.1 (i) If θ is a relation from A^* (in fact, from an arbitrary set E) to B^*, the *radix uniformisation*[17] is the function θ_{rad} which associates each element e of the domain of θ with the smallest element, in the radix order, of $e\theta$.

---

[14] *Note added in translation:* That is, in brief, recognised by a sub-automaton of the automaton of factors of length d for some d.

[15] (In French, 'addition molle'.) This terminology is attributed by Jean Berstel [28] to Jean Vuillemin.

[16] This terminology comes from logic (descriptive set theory), *cf.* [16] for example.

[17] 'Military uniformisation' is another (tempting) term in English. In French, 'uniformisation militaire' is obviously better than 'uniformisation radicielle'; this is the reason for a different symbol in the French and English editions – the only example of such a discrepancy. See also Exercise 1.3, p. 528.

Along the same lines, we can define the *lexicographic selection* θ_{lex}: it is the function which associates each element e of the domain of θ with the smallest element, in the lexicographic order, of $e\theta$, if it exists. As the lexicographic order *is not a well ordering*, this minimum element does not necessarily exist and θ_{lex} is not always a uniformisation.

(ii) Let σ be a substitution from A^* to B^*. By choosing, for each a in A, an element $a\tau$ in $a\sigma$, we define a morphism $\tau\colon A^* \to B^*$ which is a uniformisation of σ.

Note that even if we choose for $a\tau$ the smallest element, in the radix order, of $a\sigma$, the morphism τ is not necessarily the radix uniformisation of σ.

(iii) Let θ be an *equivalence relation* on A^* and T a set of representatives of classes modulo θ; that is, a *cross-section* of θ. The function which makes each word correspond to its representative in T is a uniformisation of θ. On the other hand, a uniformisation τ of θ is not necessarily in this form since the images under τ of two elements that are equivalent modulo θ can be not equal. □

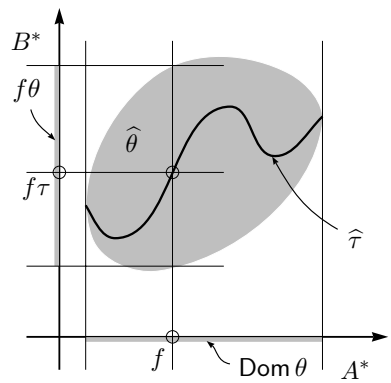


Figure 2.1: A uniformisation τ of a relation θ

If θ belongs to a certain family of relations, we are interested in knowing whether we can construct a uniformisation τ which belongs to it too, or to another given family of functions. The problem is solved for the family of rational relations by the following result:

Theorem 2.1 [**Rational Uniformisation Theorem**] *Every rational relation from A^* to B^* is uniformised by an unambiguous rational function.*

Since the only possible uniformisation of a function is the function itself, we can state the result that has already been promised:

Corollary 2.1 *Every rational function from A^* to B^* is an unambiguous rational relation.* ■

Scholium 2.1 The demonstration of Theorem 2.1 will make clear that the nature of the output, or 'image', monoid, in this case the free monoid B^*, plays no part, and

it can be replaced by an arbitrary monoid; that is, we shall prove the proposition: *every rational relation from A^* to a monoid M is uniformised by an unambiguous rational function from A^* to M.*

In particular, we can (rationally) uniformise a (rational) relation from A^* to $B^* \times C^*$. We should note that if the relations from A^* to $B^* \times C^*$ and those from $A^* \times B^*$ to C^* are indistinguishable from the point of view of rationality, *it is not true* that we can rationally uniformise a rational relation from $A^* \times B^*$ to C^*. □

cf. Exer. 3.6, p. 685

We shall prove Theorem 2.1 as the consequence of the construction of the Schützenberger covering, first directly on the transducers, then on the matrix representations, which enables further developments.

Exercise

• 2.1 Show that every recognisable relation from A^* to B^* is uniformised by a recognisable function, and that every recognisable equivalence relation has a rational cross-section. Can we generalise these results to recognisable relations between arbitrary monoids?

2.1 Proof of Theorem 2.1 (transducer version)

Recall that if $\mathcal{A}$ is an automaton over A^* and $\mathcal{A}_{\mathsf{det}}$ is its determinisation, the accessible part $\mathcal{S}$ of $\mathcal{A}_{\mathsf{det}} \times \mathcal{A}$ is a covering of $\mathcal{A}$, called the *Schützenberger covering* or *S-covering* of $\mathcal{A}$, and that the projection $\pi_{\mathcal{A}_{\mathsf{det}}}$ of $\mathcal{S}$ onto $\mathcal{A}_{\mathsf{det}}$ is a locally co-surjective morphism; and as a result we can, by eliminating some transitions, construct a sub-automaton $\mathcal{T}$ of $\mathcal{S}$, called an *S-immersion* of $\mathcal{A}$, which is unambiguous and equivalent to $\mathcal{A}$.

Th. II.3.2, p. 270

Cor. II.3.14, p. 272

Let $\mathcal{C}$ be a real-time transducer which realises a relation θ (corresponding to a representation (λ, μ, ν) of θ), $\mathcal{A}$ its underlying input automaton, and $\mathcal{T}$ an S-immersion in $\mathcal{A}$. Since $\mathcal{T}$ is an immersion in $\mathcal{A}$ each transition (r, a, s) of $\mathcal{T}$ corresponds to a *unique transition* (p, a, q) in $\mathcal{A}$ and hence to a unique transition $\big(p, (a, a\mu_{p,q}), q\big)$ in $\mathcal{C}$. If we choose, *arbitrarily*, a word w in $a\mu_{p,q}$, we construct a transducer $\mathcal{U}$ by replacing each transition (r, a, s) in $\mathcal{T}$ by $\big(r, (a, w), s\big)$. Since $\mathcal{T}$ is unambiguous, the relation τ realised by $\mathcal{U}$ is an unambiguous function, and since $\mathcal{T}$ is equivalent to $\mathcal{A}$ it has the same domain as θ, and its graph is contained in that of θ, by the choice of w. ■

This proof leads us to give the name *S-uniformisation*[18] *of* θ to a uniformisation obtained as above by selection of the transitions (for the S-immersion), then of the outputs in the S-covering of a transducer which realises θ.

Example 2.2 Let θ_2 be the relation from $\{a, b\}^*$ to itself which in each word replaces one of its factors ab by the set b^+a (and which is undefined for words which do not contain such a factor). Figure 2.2 shows a transducer $\mathcal{E}_2$ which realises θ_2

[18] In truth this name, although convenient and natural after 'S-covering' and 'S-immersion', is not entirely logical since 'uniformisation' refers to a relation θ and the prefix 'S-' to a transducer which realises it.

and whose underlying input automaton is $\mathcal{A}_1$ (on the left, vertically), its determinisation $\mathcal{A}_{1\mathsf{det}}$ (horizontally, at the top) and the result of the construction described in the proof above. □ *cf. Fig. II.3.8, p. 273*

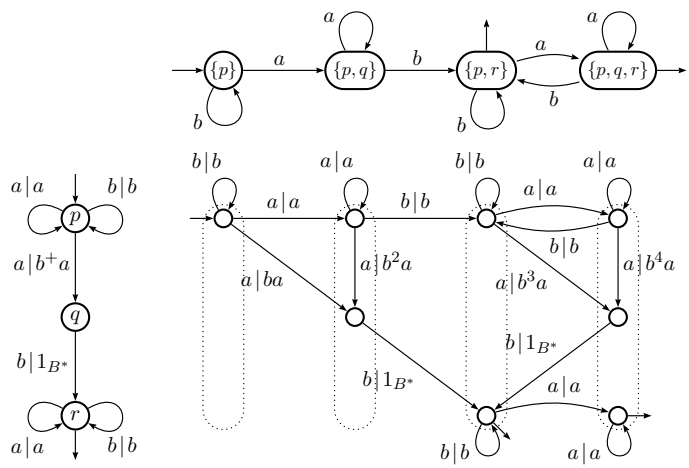


Figure 2.2: The transducer $\mathcal{E}_2$ and an S-uniformisation of θ_2

2.2 Proof of Theorem 2.1 (representation version)

The construction of the S-covering of a transducer $\mathcal{T}$ followed by its S-immersions can be re-expressed in terms of matrix representations. This rewriting adds nothing in itself, but is better suited to the applications that will follow.

2.2.1 Representation of the S-immersions of an automaton

To describe the matrix representation of the S-covering of $\mathcal{A}$, the *accessible part* of $\mathcal{A}_{\mathsf{det}}\times\mathcal{A}$, we shall need some new notation. Let α be an arbitrary matrix of size $Q\times R$ (over an arbitrary semiring) and let P be an arbitrary subset of Q. We write $\alpha^{\overline{P}}$ for the matrix whose *rows* are equal to those of α if their index is in P and are zero otherwise:

$$\forall q \in R \qquad (\alpha^{\overline{P}})_{p,q} = \begin{cases} \alpha_{p,q} & \text{if} \quad p \in P \ , \\ 0 & \text{otherwise.} \end{cases}$$

Dually (to describe the matrix representation of the co-covering), and if S is an arbitrary subset of R, we write $\alpha^{|S|}$ for the matrix whose *columns* are equal to those of α if their index is in S and are zero otherwise:[19]

$$\forall p \in Q \qquad (\alpha^{|S|})_{p,q} = \begin{cases} \alpha_{p,q} & \text{if} \quad q \in S \ , \\ 0 & \text{otherwise.} \end{cases}$$

[19]Here we have another reason to write the cardinal of a set S $\|S\|$ and not $|S|$.

For example, let $Q = R = \{1,2,3\}$ and $P = S = \{1,3\}$;

$$\text{if} \quad \alpha = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{then} \quad \alpha^{\overline{P}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \alpha^{|S|} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} .$$

We start by considering automata over a free monoid. Let (λ, μ, ν) be the (boolean) representation of $\mathcal{A} = \langle Q, A, E, I, T \rangle$ and (η, κ, ξ) the representation of $\mathcal{A}_{\mathsf{det}} = \langle \mathfrak{P}(Q), A, F, \{I\}, U \rangle$; that is,

rewriting, in terms of representations, Equat. (I.3.9) and Equat. (I.3.10), p. 105

$$\begin{aligned} \forall P, S \in \mathfrak{P}(Q)\,,\ \forall a \in A \qquad a\kappa_{P,S} = 1 \quad &\Longleftrightarrow \quad S = \{q \mid \exists p \in P \quad a\mu_{p,q} = 1\}\,, \\ \eta_S = 1 \quad &\Longleftrightarrow \quad S = \{q \mid \lambda_q = 1\} = I\,, \\ \text{and} \qquad \xi_P = 1 \quad &\Longleftrightarrow \quad \exists p \in P \quad \nu_p = 1\,. \end{aligned}$$

By definition (that is, since $\mathcal{A}_{\mathsf{det}}$ is deterministic), the matrices $a\kappa$ are *row monomial*, as is the vector η.

cf. Prop. III.3.5, p. 428

The representation of $\mathcal{A}_{\mathsf{det}} \times \mathcal{A}$ is $(\eta, \kappa, \xi) \otimes (\lambda, \mu, \nu)$. Each of the matrices $(f)\kappa \otimes \mu$ is a $\mathfrak{P}(Q) \times \mathfrak{P}(Q)$ matrix of blocks of size $Q \times Q$. The (boolean) representation of the Schützenberger covering of $\mathcal{A}$ is given by the following easily verified proposition.

Proposition 2.2 *With the above notation, the representation (ζ, σ, ω) of the Schützenberger covering of the automaton $\mathcal{A}$ is defined by*

$$\forall P, S \in \mathfrak{P}(Q)\,,\ \forall a \in A \qquad a\sigma_{(P,Q),(S,Q)} = \begin{cases} a\mu^{\overline{P}} & \text{if} \quad a\kappa_{P,S} = 1 \\ 0 & \text{otherwise,} \end{cases}$$

and by

$$\forall P, S \in \mathfrak{P}(Q) \qquad \zeta_{(S,Q)} = \begin{cases} \lambda & \text{if} \quad \eta_S = 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \omega_{(P,Q)} = \nu^{\overline{P}} \ . \quad \blacksquare$$

In other words, for each a in A, the matrix $a\sigma$ is the $\mathfrak{P}(Q) \times \mathfrak{P}(Q)$ matrix of $Q \times Q$-blocks obtained by replacing *the* non-zero entry of each row $P \in \mathfrak{P}(Q)$ of $a\kappa$ by the ($Q \times Q$-)matrix $a\mu$ of which we have kept only the rows belonging to P (replacing the others by 0).

cf. Exam. II.2.3, p. 240

Example 2.2 (continued) The representation of $\mathcal{A}_1$ is

$$\lambda_1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \ , \quad a\mu_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \ , \quad b\mu_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \ , \quad \nu_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \ .$$

And that of $\mathcal{A}_{1\mathsf{det}}$

$$\eta_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \ , \quad a\kappa_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \ , \quad b\kappa_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \ , \quad \xi_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix} \ .$$

The representation $(\zeta_1, \sigma_1, \omega_1)$ of the Schützenberger covering $\mathcal{S}_1$ of $\mathcal{A}_1$ is thus (the 'large 0's' represent 3×3 blocks of zeros)

$$\zeta_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$a\sigma_1 = \begin{pmatrix} 0 & \begin{matrix} 1&1&0\\0&0&0\\0&0&0 \end{matrix} & 0 & 0 \\ 0 & \begin{matrix} 1&1&0\\0&0&0\\0&0&0 \end{matrix} & 0 & 0 \\ 0 & 0 & 0 & \begin{matrix} 1&1&0\\0&0&0\\0&0&1 \end{matrix} \\ 0 & 0 & 0 & \begin{matrix} 1&1&0\\0&0&0\\0&0&1 \end{matrix} \end{pmatrix}, \quad b\sigma_1 = \begin{pmatrix} \begin{matrix} 1&0&0\\0&0&0\\0&0&0 \end{matrix} & 0 & 0 & 0 \\ 0 & 0 & \begin{matrix} 1&0&0\\0&0&1\\0&0&0 \end{matrix} & 0 \\ 0 & 0 & \begin{matrix} 1&0&0\\0&0&0\\0&0&1 \end{matrix} & 0 \\ 0 & 0 & \begin{matrix} 1&0&0\\0&0&1\\0&0&1 \end{matrix} & 0 \end{pmatrix}, \quad \omega_1 = \begin{pmatrix} 0\\0\\0\\0\\0\\0\\0\\0\\1\\0\\0\\1 \end{pmatrix}.$$

We verify easily that it is indeed the representation of the automaton of Figure II.3.7, which is reproduced in Figure 2.3, except that, in order to make the blocks of the same size, we imagine that the dotted cartouches each contain 3 states, the added states being neither the source nor destination of any transition. □ *cf. p. 271*

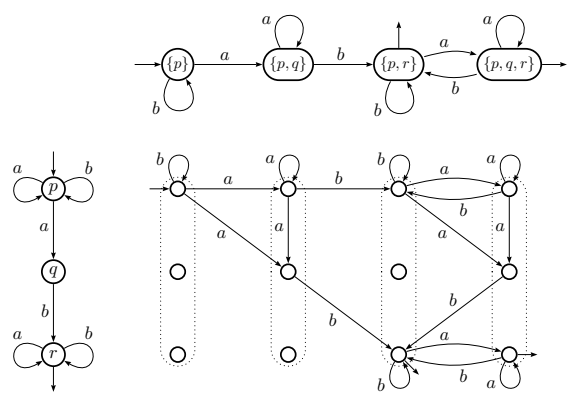


Figure 2.3: The S-covering of $\mathcal{A}_1$, obtained from its matrix representation

The definition of a (boolean) representation of an S-immersion in $\mathcal{A}$ from the representation (ζ, σ, ω) of the Schützenberger covering of $\mathcal{A}$ is immediate. For each a in A, each non-zero block $a\sigma_{(P,Q),(S,Q)}$ in $a\sigma$ is replaced by a *column-monomial* block which has the same non-zero columns as the original block. In other words, each non-zero non-monomial column of each $Q\times Q$-block of $a\sigma$ is made monomial, *but non-zero*, by arbitrarily zeroing some of its entries. The same operation is applied to the $Q\times 1$-blocks of ω.

The simplest way to show that this representation is that of an automaton equivalent to $\mathcal{A}$ is probably to come back to automata and their morphisms (an In-surjective morphism is made In-bijective).

Example 2.2 (continued) There are two S-immersions $\mathcal{T}_1'$ and $\mathcal{T}_1''$ in $\mathcal{A}_1$, of representations $(\zeta_1, \sigma_1', \omega_1')$ and $(\zeta_1, \sigma_1'', \omega_1'')$ respectively. Obviously, $a\sigma_1' = a\sigma_1'' = a\sigma_1$ and $\omega_1' = \omega_1'' = \omega_1$. The construction only really affects the matrices $b\sigma_1'$ and $b\sigma_1''$:

$$b\sigma_1' = \begin{pmatrix} \begin{smallmatrix} 1&0&0\\0&0&0\\0&0&0 \end{smallmatrix} & 0 & 0 & 0 \\ 0 & 0 & \begin{smallmatrix} 1&0&0\\0&0&1\\0&0&0 \end{smallmatrix} & 0 \\ 0 & 0 & \begin{smallmatrix} 1&0&0\\0&0&0\\0&0&1 \end{smallmatrix} & 0 \\ 0 & 0 & \begin{smallmatrix} 1&0&0\\0&0&\boxed{0}\\0&0&1 \end{smallmatrix} & 0 \end{pmatrix}, \qquad b\sigma_1'' = \begin{pmatrix} \begin{smallmatrix} 1&0&0\\0&0&0\\0&0&0 \end{smallmatrix} & 0 & 0 & 0 \\ 0 & 0 & \begin{smallmatrix} 1&0&0\\0&0&1\\0&0&0 \end{smallmatrix} & 0 \\ 0 & 0 & \begin{smallmatrix} 1&0&0\\0&0&0\\0&0&1 \end{smallmatrix} & 0 \\ 0 & 0 & \begin{smallmatrix} 1&0&0\\0&0&1\\0&0&\boxed{0} \end{smallmatrix} & 0 \end{pmatrix}.$$

We have drawn boxes around the zeroed entries to make them easier to spot. □

We clearly see in this example that it is the *blocks* of the matrix $a\sigma$ (or $b\sigma$) which are transformed into column-monomial blocks and not the whole matrix. This leads us to give a name to matrices of this form, which will be essential in the next subsection.

cf. Prop. 2.7, p. 674

Remark 2.1 This matrix construction allows us to understand the true nature of the Schützenberger covering $\mathcal{S}$ of an automaton $\mathcal{A}$. It is because $\mathcal{S}$ is the *accessible part* of $\mathcal{A}_{\mathsf{det}} \times \mathcal{A}$, and *not the whole of* $\mathcal{A}_{\mathsf{det}} \times \mathcal{A}$ (which is also a covering of $\mathcal{A}$, or, which is equivalent, because we replace the 1's of $a\kappa$ by the blocks $a\mu^{\overline{P}}$, and not by the blocks $a\mu$) that we can obtain an immersion by *arbitrarily* zeroing entries in these blocks to make them column monomial. □

2.2.2 Semi-monomial matrices

Definition 2.1 A matrix m is called *(row-column) semi-monomial* if there is a block decomposition of m such that m is *row* monomial as a matrix of blocks and each block is column monomial. □

Naturally, we say that a *representation* is (row-column) semi-monomial if all the matrices λ, ν, and $a\mu$, for all a, are semi-monomial matrices, with respect to the *same* block decomposition. Just as naturally, we have the dual notions of the *column-row* semi-monomial matrix and representation: column-monomial matrices of row-monomial blocks. The compatibility of matrix multiplication with their block decompositions implies:

cf. Exer. 0.6.9, p. 32

Property 2.1 *The product of two semi-monomial matrices (of congruent size and block decomposition) is a semi-monomial matrix.* ■

Everything we have just said applies implicitly to matrices with entries in $\mathfrak{P}(B^*)$, that is $\mathbb{B}\langle\langle B^*\rangle\rangle$, but also holds if we take weights into account and consider matrices with entries in $\mathbb{N}\langle\langle B^*\rangle\rangle$ or even $\mathbb{K}\langle\langle B^*\rangle\rangle$. In fact, we can be more precise in this case. Logically, we say that a monomial of $\mathbb{K}\langle\langle M\rangle\rangle$ is *characteristic* if its coefficient is $1_{\mathbb{K}}$;

and that a (row or column) monomial or semi-monomial matrix is characteristic if its non-zero entries are characteristic monomials. It then follows, in the same way as for Property 2.1 and since in the product of two such matrices we never have to add two non-zero coefficients:

Property 2.2 *The product of two characteristic monomial matrices is a characteristic monomial matrix. The product of two characteristic semi-monomial matrices is a characteristic semi-monomial matrix.* ■

It follows that if the matrices $a\mu$ (and the vectors λ and ν) are characteristic, then so are the matrices $f\mu$ (and $\lambda \cdot f\mu \cdot \nu$), for all words f, and we can state (which is obviously what we were aiming at):

Property 2.3 *A rational relation realised by a characteristic semi-monomial representation is an unambiguous rational function.* ■

2.2.3 Representation of S-uniformisations

We start, as in Subsection 2.1, with a relation θ, realised by a transducer $\mathcal{C}$ whose representation[20] is (λ, μ, ν) (of dimension Q). Let $\mathcal{A}$ be its underlying input automaton and (η, κ, ξ) the (boolean) representation of dimension $\mathfrak{P}(Q)$ of $\mathcal{A}_{\mathsf{det}}$. *A transducer is an automaton* and we construct a covering of $\mathcal{C}$ exactly as we constructed that of $\mathcal{A}$ earlier.[21]

Proposition 2.3 *With the above notations, the representation* (ζ, σ, ω) *of the Schützenberger covering of the transducer* $\mathcal{C}$ *is defined by:*

$$\forall P, S \in \mathfrak{P}(Q)\,,\ \forall a \in A \qquad a\sigma_{(P,Q),(S,Q)} = \begin{cases} a\mu^{\overline{P}} & \textit{if} \quad a\kappa_{P,S} = 1 \\ 0 & \textit{otherwise} \end{cases}$$

and by:

$$\forall P, S \in \mathfrak{P}(Q) \qquad \zeta_{(S,Q)} = \begin{cases} \lambda & \textit{if} \quad \eta_S = 1 \\ 0 & \textit{otherwise} \end{cases} \qquad \textit{and} \qquad \omega_{(P,Q)} = \nu^{\overline{P}}\ . \quad ■$$

Now we repeat, almost verbatim, the construction of a representation of an S-immersion in $\mathcal{A}$. For each letter a in A, each non-zero block $a\sigma_{(P,Q),(S,Q)}$ in $a\sigma$ is replaced by a *column-monomial* block which has the same non-zero columns as the original block. In other words, each non-zero non-monomial column of each $Q \times Q$-block of $a\sigma$ is made monomial, *but non-zero*, by replacing each non-zero entry (a

[20] Do not forget that 'transducer' in this chapter means 'real-time transducer'.

[21] Which is the same as saying that an automaton over A^* can be seen as a transducer from A^* to B^* all of whose transitions have the output 1_{B^*} and which justifies our having written (λ, μ, ν) for the representation of $\mathcal{C}$, like that of $\mathcal{A}$ in § 2.2.1.

subset of B^*) *either by* 0 *(the empty set), or by a word arbitrarily chosen from this entry.* The same operation is applied to the $Q\times 1$-blocks of ω.

We thus obtain a semi-monomial representation. It realises a uniformisation of θ since it is a *function* such that, by construction:

(i) its graph is included in that of θ;

(ii) its domain is the language recognised by an S-immersion in the underlying initial automaton $\mathcal{A}$, hence equal to the language recognised by $\mathcal{A}$ which is the domain of θ.

This construction is hence a proof (the same as before, but differently explained) of Theorem 2.1. ■

Example 2.2 (continued) We again write $(\lambda_1, \mu_1, \nu_1)$ for the representation of $\mathcal{E}_2$:

$$\lambda_1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \ , \quad a\mu_1 = \begin{pmatrix} a & b^+a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{pmatrix} \ , \quad b\mu_1 = \begin{pmatrix} b & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & b \end{pmatrix} \ , \quad \nu_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \ .$$

The representation $(\zeta_1, \sigma_1, \omega_1)$ of the Schützenberger covering of $\mathcal{E}_2$ is then

$$\zeta_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \ ,$$

$$a\sigma_1 = \begin{pmatrix} 0 & \begin{matrix} a & b^+a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} & 0 & 0 \\ 0 & \begin{matrix} a & b^+a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} & 0 & 0 \\ 0 & 0 & 0 & \begin{matrix} a & b^+a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{matrix} \\ 0 & 0 & 0 & \begin{matrix} a & b^+a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{matrix} \end{pmatrix} \ , \quad b\sigma_1 = \begin{pmatrix} \begin{matrix} b & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} & 0 & 0 & 0 \\ 0 & 0 & \begin{matrix} b & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{matrix} & 0 \\ 0 & 0 & \begin{matrix} b & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & b \end{matrix} & 0 \\ 0 & 0 & \begin{matrix} b & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & b \end{matrix} & 0 \end{pmatrix} \ , \quad \omega_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \ .$$

The representation $(\zeta_1, \sigma_1', \omega_1')$ of the uniformisation of Figure 2.2 is given by

$$a\sigma_1' = \begin{pmatrix} 0 & \begin{matrix} a & ba & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} & 0 & 0 \\ 0 & \begin{matrix} a & b^2a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} & 0 & 0 \\ 0 & 0 & 0 & \begin{matrix} a & b^3a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{matrix} \\ 0 & 0 & 0 & \begin{matrix} a & b^4a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{matrix} \end{pmatrix} \ , \qquad b\sigma_1' = \begin{pmatrix} \begin{matrix} b & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} & 0 & 0 & 0 \\ 0 & 0 & \begin{matrix} b & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{matrix} & 0 \\ 0 & 0 & \begin{matrix} b & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & b \end{matrix} & 0 \\ 0 & 0 & \begin{matrix} b & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{matrix} & 0 \end{pmatrix}$$

and $\omega_1' = \omega_1$. □

This construction on the representations also allows us to clarify Corollary 2.1 in a way that is worth stating explicitly (and which implies Corollary 2.1 even more directly).

Proposition 2.4 *Every rational function has a characteristic semi-monomial representation.* ■

Convention In the rest of the chapter we shall say that a matrix with entries in $\mathbb{N}\langle\langle M\rangle\rangle$, of a representation of A^* in $\mathbb{N}\langle\langle M\rangle\rangle$, is *monomial* (resp. *semi-monomial*) to mean that it is *characteristic* monomial (resp. characteristic semi-monomial).

Exercise

• 2.2 (a) Describe the construction of the matrix representation of the Schützenberger co-covering of a transducer, then of that of an S-co-immersion.

(b) Apply the above construction to the transducer $\mathcal{E}_2$ of Example 2.2.

2.3 Decomposition of rational functions

The very same construction with which we just proved the Uniformisation Theorem yields a structure theorem for rational functions, and this is not its least merit.

Theorem 2.2 *Every rational function from one free monoid to another is the composition of a sequential function with a co-sequential function.*

Since the transpose of a rational function is a rational function we also have the dual proposition, that *every rational function (from one free monoid to another) is the composition of a co-sequential function with a sequential function.* In fact, this result can be refined into a stronger decomposition theorem:[22]

Theorem 2.3 [**Elgot and Mezei**] *Every rational function from one free monoid to another is the composition of a pure sequential function with a pure co-sequential function.*

2.3.1 The 'Weak' Decomposition Theorem

From Schützenberger's construction, expressed in terms of representations, we deduced that *every rational function has a semi-monomial representation.* However, almost by the definition of semi-monomial representations, the composition product of a row-monomial representation and a column-monomial representation is a row-column semi-monomial representation. The Theorem of Composition of Representations therefore gives the two following propositions: *cf. Th. IV.3.6, p. 567*

Corollary 2.5 *The composition product of a sequential function by a co-sequential function has a row-column semi-monomial representation.* ■

Corollary 2.6 *The composition product of a co-sequential function by a sequential function has a column-row semi-monomial representation.* ■

[22] *Note added in translation:* Theorem 2.2 – which already states the core of the property – is no less Elgot and Mezei's than Theorem 2.3. It is our method of proof that proceeds in two steps.

Example 1.4 (continued) The composition product of $(\lambda_2, \mu_2, \nu_2)$ by $(\eta_2, \kappa_2, \xi_2)$ gives a semi-monomial representation $(\chi_2, \pi_2, \psi_2) = (\lambda_2, \mu_2, \nu_2) \circ (\eta_2, \kappa_2, \xi_2)$:

$$\chi_2 = \begin{pmatrix} bb & b & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} ,$$

$$a\pi_2 = \begin{pmatrix} 0 & \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix} & 0 \\ 0 & \begin{matrix} 0 & 0 & 0 \\ aa & 0 & 0 \\ 0 & ab & a \end{matrix} & 0 \\ 0 & \begin{matrix} 0 & 0 & 0 \\ aab & aa & 0 \\ 0 & 0 & ab \end{matrix} & 0 \end{pmatrix}, \quad b\pi_2 = \begin{pmatrix} \begin{matrix} b & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{matrix} & 0 & 0 \\ 0 & 0 & \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix} \\ 0 & \begin{matrix} aa & 0 & 0 \\ 0 & ab & a \\ 0 & 0 & 0 \end{matrix} & 0 \end{pmatrix}, \quad \psi_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ a \\ 0 \\ 0 \\ ab \end{pmatrix},$$

cf. p. 544 and p. 632 which we have already seen, as a transducer, in Exercise 1.25. □

In fact, Theorem 2.2 is the converse of Corollary 2.5:

Proposition 2.7 *Let $\tau\colon A^* \to B^*$ be a rational function with a semi-monomial representation (χ, π, ψ) . Then there exists:*

(i) *an alphabet Z;*

(ii) *a sequential function $\theta\colon A^* \to Z^*$ with a row-monomial representation (λ, μ, ν);*

(iii) *and a co-sequential function $\sigma\colon Z^* \to B^*$ with a column-monomial representation (η, κ, ξ) ,*

such that $(\chi, \pi, \psi) = (\lambda, \mu, \nu) \circ (\eta, \kappa, \xi)$ (and hence $\tau = \theta\sigma$).

Proof. Let (χ, π, ψ) be a semi-monomial representation of *dimension* $Q \times R$; that is, for all a in A, $a\pi$ is a $Q \times Q$ matrix of row-monomial blocks, each block being a column-monomial $R \times R$ matrix.

Then let $Z = \{Q \times A\} \bigcup \{Q\}$. The representation (λ, μ, ν), where μ is a morphism from A^* to $(\mathrm{Rat}\, Z^*)^{Q \times Q}$, is defined by

$$\forall p, q \in Q\,,\ \forall a \in A \qquad a\mu_{p,q} = \begin{cases} (p, a) & \text{if the block } a\pi_{(p,R),(q,R)} \text{ is non-zero,} \\ 0 & \text{otherwise,} \end{cases}$$

(since $a\pi$ is semi-monomial, $a\mu$ is row monomial), and by

$$\forall p, q \in Q \qquad \lambda_q = \begin{cases} 1 & \text{if the block } \chi_{(q,R)} \text{ is non-zero,} \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad \nu_p = p \ .$$

The morphism κ of the representation (η, κ, ξ) is defined, for elements of Z of the form (p, a), by

$$\forall p \in Q\,,\ \forall a \in A \qquad (p, a)\kappa = a\pi_{(p,R),(q,R)} \ ,$$

where q is the unique state such that $a\mu_{p,q}$ is non-zero. Since $a\pi$ is semi-monomial, $(p, a)\kappa$ is column monomial. We then choose an arbitrary element t of R and, for

all p in Q, define $p\kappa$ as the $R\times R$ matrix whose tth column is equal to $\psi_{(p,R)}$ and all of whose other columns are zero. Since ψ is semi-monomial, $p\kappa$ is column monomial. The two vectors η and ξ are defined by

$$\eta = \chi_{(q,R)} \qquad \text{where } \chi_{(q,R)} \text{ is the non-zero block of } \chi, \text{ and}$$

$$\forall s \in R \qquad \xi_s = \begin{cases} 1 & \text{if} \quad s = t\,, \\ 0 & \text{otherwise.} \end{cases}$$

It follows from Equations (III.3.4) and (III.3.5) that $(\chi,\pi,\psi) = (\lambda,\mu,\nu)\circ(\eta,\kappa,\xi)$ *cf. p. 566 and p. 567* and hence, by Theorem IV.3.6, that $\tau = \theta\,\sigma$. ■

Propositions 2.4 and 2.7 prove Theorem 2.2. ■

Example 1.4 (continued) Let us decompose the semi-monomial representation (χ_2,π_2,ψ_2) of dimension $(Q\times R)\times(Q\times R)$ with $Q=\{p,q,r\}$ and $R=\{s,t,u\}$, into a product $(\chi_2,\pi_2,\psi_2) = (\lambda'_2,\mu'_2,\nu'_2)\circ(\eta'_2,\kappa'_2,\xi'_2)$. We can obviously take $(\lambda'_2,\mu'_2,\nu'_2) = (\lambda_2,\mu_2,\nu_2)$ and $(\eta'_2,\kappa'_2,\xi'_2) = (\eta_2,\kappa_2,\xi_2)$. But if we follow the construction of the previous proof, we obtain

$$\lambda'_2 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix},$$

$$a\mu'_2 = \begin{pmatrix} 0 & (p,a) & 0 \\ 0 & (q,a) & 0 \\ 0 & (r,a) & 0 \end{pmatrix}, \quad b\mu'_2 = \begin{pmatrix} (p,b) & 0 & 0 \\ 0 & 0 & (q,b) \\ 0 & (r,b) & 0 \end{pmatrix}, \quad \nu'_2 = \begin{pmatrix} p \\ q \\ r \end{pmatrix},$$

then

$$(p,a)\kappa'_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (q,a)\kappa'_2 = \begin{pmatrix} 0 & 0 & 0 \\ aa & 0 & 0 \\ 0 & ab & a \end{pmatrix}, \quad (r,a)\kappa'_2 = \begin{pmatrix} 0 & 0 & 0 \\ aab & aa & 0 \\ 0 & 0 & ab \end{pmatrix},$$

$$(p,b)\kappa'_2 = \begin{pmatrix} b & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad (q,b)\kappa'_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad (r,b)\kappa'_2 = \begin{pmatrix} aa & 0 & 0 \\ 0 & ab & a \\ 0 & 0 & 0 \end{pmatrix},$$

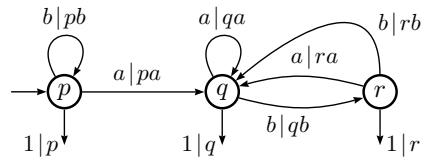
and, if we choose t as the distinguished state of R

$$p\kappa'_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad q\kappa'_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & a & 0 \end{pmatrix}, \quad r\kappa'_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & ab & 0 \end{pmatrix},$$

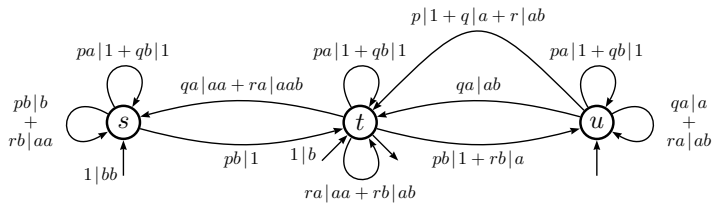
we end up with

$$\eta'_2 = \begin{pmatrix} bb & b & 1 \end{pmatrix}, \qquad \text{and} \qquad \xi'_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

The transducers that correspond to the matrix representations $(\lambda'_2,\mu'_2,\nu'_2)$ and $(\eta'_2,\kappa'_2,\xi'_2)$ are given in Figure 2.4. □



(a) the sequential transducer



(b) the co-sequential transducer

Figure 2.4: Decomposition of the semi-monomial representation (χ_2, π_2, ψ_2)

2.3.2 The 'Strong' Decomposition Theorem

Two improvements in technique take us from the proof of Proposition 2.7 to that of the stronger version of the Decomposition Theorem.

Proof of Theorem 2.3. We reuse the notation and definitions of the last proof, writing i for the unique state of Q such that λ_i is non-zero (and hence equal to 1). The first step consists of changing θ into a *pure sequential function*

$$\theta' \colon A^* \to Z^* \ .$$

By definition, the function θ' realised by the representation (λ, μ, ν'), where ν' is the vector all of whose elements are 1, is pure sequential. If $f = a_1 a_2 \cdots a_n$ is a word in A^*, there exists a unique sequence $p_1, p_2, \ldots, p_n$ of states of Q such that

$$f\theta = (i, a_1, p_1)\,(p_1, a_2, p_2) \cdots (p_{n-1}, a_n, p_n)\, p_n \ ,$$

and hence such that:

$$f\theta' = (i, a_1, p_1)\,(p_1, a_2, p_2) \cdots (p_{n-1}, a_n, p_n) \ .$$

cf. Prop. III.2.14, p. 410

By adapting the standard construction used to make automata proper, we then define *a co-sequential function* $\sigma' \colon Z^* \to B^*$ such that $\theta'\sigma' = \theta\sigma = \tau$. Let t be a new state which does not belong to R and $R' = R \cup \{t\}$. The function σ' is realised by the representation (η', κ', ξ') of dimension R', where the morphism κ' is defined by

$$\forall p, q \in Q\,,\ \forall a \in A \qquad (p, a, q)\kappa' = \begin{pmatrix} \boxed{(p,a,q)\kappa} & \boxed{\cdot} \\ \boxed{0} & 0 \end{pmatrix} \leftarrow \varphi_{(p,a,q)}$$

where $\varphi_{(p,a,q)} = (p,a,q)\kappa \cdot \psi_q$ is a (column-)monomial (column) vector, and where

$$\eta' = \begin{pmatrix} \boxed{\quad \eta \quad} & 0 \end{pmatrix} \qquad \text{and} \qquad \xi' = \begin{pmatrix} \boxed{0} \\ 1 \end{pmatrix} .$$

It is immediate to verify that, for all f in A^*, we have: $\eta \cdot f\kappa \cdot \xi = \eta' \cdot f\kappa' \cdot \xi'$.

The last step – to transform σ' into *a pure co-sequential function* – is even simpler. Let \$ be a symbol that does not belong to Z and $Z' = Z \cup \{\$\}$. The function $\theta'' \colon A^* \to Z'^*$ defined by $f\theta'' = \$ f\theta'$ for all f in A^* (that is, $f\theta'' = \$(i,a_1,p_1)(p_1,a_2,p_2)\cdots(p_{n-1},a_n,p_n)$ if $f = a_1a_2\cdots a_n$) is a pure sequential function. Then, by the dual of Lemma 1.8, we deduce that the function $\sigma'' \colon Z'^* \to B^*$ defined by $(\$ u)\sigma'' = u\sigma'$ is *pure* co-sequential since σ' is co-sequential. ■

cf. Exer. 1.10, p. 656

cf. p. 659

Exercise

• 2.3 Decompose the S-uniformisation of the function θ_2 defined by the transducer in Figure 2.2 (p. 667) into the product of a sequential transducer by a co-sequential transducer.

2.4 The Rational Skimming Theorem II

We saw that if we take weights into account – that is, if we consider an $\mathbb{N}$-rational series s of $\mathbb{N}\mathrm{Rat}\, A^*$ instead of a rational language L of A^* recognised by an automaton $\mathcal{A}$ – then a simple transformation takes us from the construction of an unambiguous S-immersion $\mathcal{T}$ in $\mathcal{A}$ which recognises L to a decomposition of s into $s = \underline{\mathsf{supp}\, s} + t$ where t is also an $\mathbb{N}$-rational series. In the same way, the Rational Uniformisation Theorem will be transformed into a generalisation of the Rational Skimming Theorem for $\mathbb{N}$-rational relations.

cf. § III.4.4.2, p. 465

Theorem 2.4 *Let $\theta \colon A^* \to B^*$ be a rational $\mathbb{N}$-relation. There exist two rational $\mathbb{N}$-relations σ and τ such that:*

(i) *τ is an unambiguous function which uniformises θ;* (ii) $\theta = \sigma + \tau$.

Proof. In fact, Theorem 2.4 does not need a proof, since it has already been proved at the same time as Theorem III.4.16. In the latter theorem's proof, we considered an $\mathbb{N}$-automaton $\mathcal{A}$, which realises an $\mathbb{N}$-rational series s.

cf. p. 465

We replace $\mathcal{A}$ by an $\mathbb{N}$-transducer $\mathcal{E} = \langle I, E, T \rangle$ from A^* to B^*, an automaton over A^* with weights in $\mathbb{N}\mathrm{Rat}\, B^*$; that is, the entries of E are linear combinations, with coefficients in $\mathbb{N}\mathrm{Rat}\, B^*$, of letters of A. We construct, as in the case of Theorem III.4.16, the S-covering $\mathcal{S} = \langle J, F, U \rangle$ of $\mathcal{E}$, and then an S-immersion $\mathcal{T} = \langle J, F_1, U_1 \rangle$. The entries of F_1 are linear combinations of letters of A whose coefficients are *characteristic $\mathbb{N}$-series of words* of B^* (respectively the coefficients of U_1 are *characteristic $\mathbb{N}$-series of words* of B^*).

Since if t belongs to $\mathbb{N}\mathrm{Rat}\, B^*$ and the word f to the support of t, we have that $u = t - \underline{f}$ is an $\mathbb{N}$-rational series of B^*, $F = F_1 + F_2$ (and $U = U_1 + U_2$) and the

cf. Exer. III.4.17, p. 469

entries of F_2 are linear combinations of letters of A whose coefficients are *rational* $\mathbb{N}$-*series* on B^* (respectively the coefficients of U_2 are *rational* $\mathbb{N}$-*series* of B^*).

The proof then continues identically to that of Theorem III.4.16, and the $\mathbb{N}$-transducer $\mathcal{P}$:

$$\mathcal{P} = \left\langle \begin{pmatrix} J & J & 0 \end{pmatrix}, \begin{pmatrix} F_1 & 0 & F_2 \\ 0 & F_1 & 0 \\ 0 & 0 & F \end{pmatrix}, \begin{pmatrix} 0 \\ U_2 \\ U \end{pmatrix} \right\rangle$$

is such that $|\mathcal{E}| = |\mathcal{S}| = |\mathcal{T}| + |\mathcal{P}|$. ■

cf. p. 467 **Example 2.3** We return, with some modifications, to Example III.2.5: we transform the $\mathbb{N}$-automaton $\mathcal{C}_1$ into an $\mathbb{N}$-transducer $\mathcal{E}_3$ from A^* to B^*. This transducer does not realise a particularly special $\mathbb{N}$-relation; it was chosen so that Figure 2.5 could show the construction of the uniformisation of the $\mathbb{N}$-relation $|\mathcal{E}_3|$ and of its complement. □

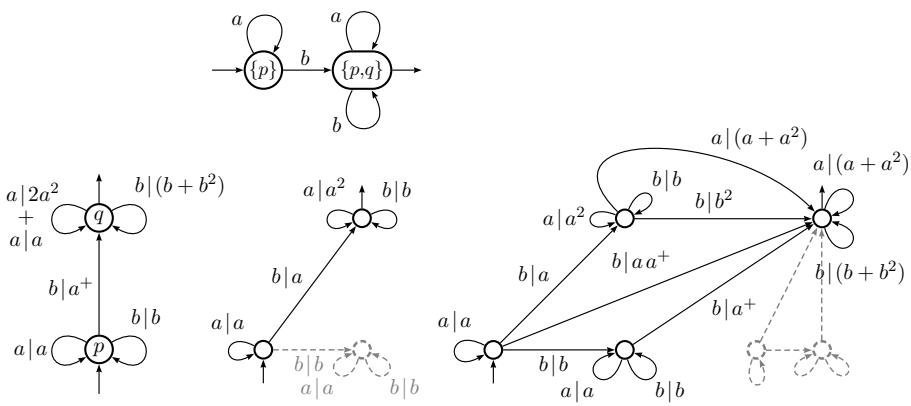


Figure 2.5: An $\mathbb{N}$-transducer $\mathcal{E}_3$ (left), the determinisation of the underlying input automaton of $\mathcal{E}_3$ (top), a uniformisation of $|\mathcal{E}_3|$, and the $\mathbb{N}$-relation which is the complement of this uniformisation.

cf. Fact II.1.1, p. 231 and Rem. II.SE.1.1, p. 350

This theorem allows us to return, as promised long ago, to the proof that, in the monoid $M = \{a,b\}^* \times \{c\}^*$, the subset

$$(a,c)^*(b,1)^* \cup (a,1)^*(b,c)^*$$

is not an unambiguous rational subset. Suppose that it is unambiguous; it can then be seen as the graph of an $\mathbb{N}$-relation θ from $\{a,b\}^*$ to $\{c\}^*$, and we have

$$a^n b^m \theta = \begin{cases} c^n + c^m & \text{if} \quad n \neq m\,, \\ c^n & \text{if} \quad n = m\,. \end{cases}$$

We decompose θ into $\theta = \tau + \sigma$ where τ is a characteristic $\mathbb{N}$-rational function with $\mathsf{Dom}\,\tau = \mathsf{Dom}\,\theta = a^*b^*$. But then $\mathsf{Dom}\,\sigma = \{a^n b^m \mid n \neq m\}$ is not rational and σ cannot be a rational ($\mathbb{N}$-)function; contradiction. In the same spirit, repeated applications of Theorem 2.4 give:

Corollary 2.8 *A rational $\mathbb{N}$-relation θ from A^* to B^* such that, for all f in A^*, the sum of the coefficients of the series $f\theta$ is bounded by a fixed integer k, is the sum of k (unambiguous) rational functions from A^* to B^*.* ■

In fact, we can return to *unweighted* relations and prove – though it is much harder – results in the same vein as those above. Let θ be a relation from A^* to B^*. Recall that we write $\|f\theta\|$ for the number of elements of the image of f under θ and call the *norm* of θ the upper bound of the $\|f\theta\|$ for f in A^*. The *rational relations of finite norm* share with the rational functions some remarkable properties.[23]

Theorem 2.5 [**Weber**] *It is decidable whether a rational relation has finite norm.*

Theorem 2.6 [**Gurari and Ibarra**] *It is decidable whether a rational relation has finite norm k.*

Theorem 2.7 [**Culik and Kahrumäki**] *The equivalence of rational relations of finite norm is decidable.*

Theorem 2.8 [**Weber**] *A rational relation of finite norm k is the sum of k (unambiguous) rational functions.*

3 Cross-section of rational functions

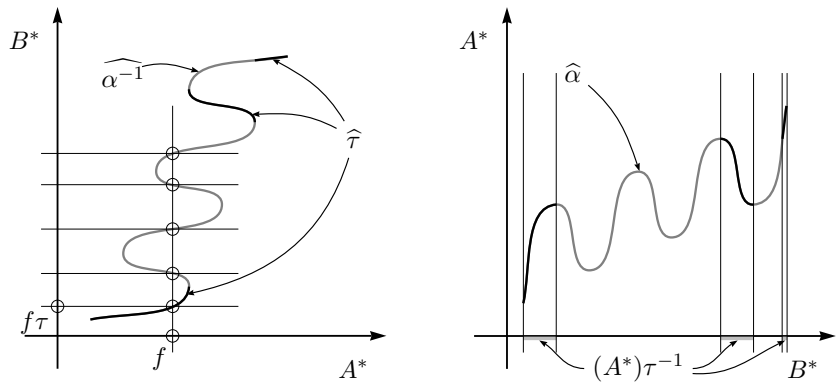
In the case where the relation θ is the *inverse of a function* α, we can give a 'dual' of the Rational Uniformisation Theorem[24] which turns out to be often as useful as the direct version. We see in Figure 3.1 that, from the point of view of the function α, the choice of a uniformisation of $\theta = \alpha^{-1}$ comes down to fixing a representative in each class of the map equivalence of α; that is, to choosing a *cross-section* of $\alpha\alpha^{-1}$, hence to the determination of a subset of the domain of α.

This simple reversal of our point of view will reveal itself to be rich and to allow us to deal with other structures in terms of rationality. In particular, the examination of the free group and of simplification monoids in this way is a first step towards an algebraic treatment of pushdown automata.

Uniformisation or cross-section, in both cases the important thing is the *choice*; we shall in the second subsection take the preceding results and try to make precise the manner in which this choice can be made.

[23] *Note added in translation:* The French edition was missing the important Theorem 2.6 and attributed somewhat incorrectly Theorem 2.5 to Schützenberger. The theory of rational relations of finite norm has been recently rewritten by Rodrigo de Souza and the author (see bibliographic notes).

[24] This is the classic domain–image duality of a function or relation and not the left–right duality with which we are generally concerned.



(a) uniformisation of α^{-1} (b) cross-section of $\alpha\alpha^{-1}$

Figure 3.1: The uniformisation–cross-section duality

3.1 The rational cross-section property

In the following definition, we shall free ourselves from the fact that functions are considered to take values in a monoid, even if this will almost always be the case.[25]

Definition 3.1 Let E be a set. A function $\alpha\colon A^* \to E$ *has the rational cross-section property* if for every rational R of A^* there exists a rational T, contained in R, such that α is injective on T and $T\alpha = R\alpha$ (that is, α induces a bijection between T and $R\alpha$). □

In other words, α has the rational cross-section property if for every rational R of A^* there exists a rational cross-section of $\iota_R\,\alpha\,\alpha^{-1}$; this proposition shows that it is not so much the function α as its *map equivalence* which possesses the rational cross-section property. We also note that we can define this property for functions defined on any monoid and not only on free monoids.

3.1.1 The Rational Cross-Section Theorem

From the Rational Uniformisation Theorem we deduce:

Proposition 3.1 *Every rational function from A^* to B^* has the rational cross-section property.*

Proof. Let $\alpha\colon A^* \to B^*$ be a rational function and R a rational of A^*. The relation $\theta = \alpha^{-1}\iota_R$ from B^* to A^* is rational, of domain $\mathsf{Dom}\,\theta = R\alpha$. It is uniformised by a (rational) function τ, of domain $R\alpha$, and such that for all f in its domain, $((f)\tau)\alpha = f$. The set $T = (R\alpha)\tau$ is rational, contained in $\mathsf{Im}\,\theta = R \cap \mathsf{Dom}\,\alpha \subseteq R$; we have $T\alpha = R\alpha$ and α is injective on T since it is invertible: T is a cross-section of R for α. ■

[25] Always, in this volume; but the definition must still hold when we later use it in the study of pushdown automata.

In particular, we can state:[26]

Theorem 3.1 [**Rational Cross-Section Theorem**] *Every morphism from A^* to B^* has the rational cross-section property.* ■

Scholium 3.1 Theorems 2.1 and 3.1 are truly dual in the sense that while Theorem 3.1 was obtained here as a special case of a simple corollary of Theorem 2.1, we could just as well have proved Theorem 3.1 first, then deduced Proposition 3.1 and Theorem 2.1 immediately. *cf. Exer. 3.1*

Therefore let $\theta\colon A^* \to B^*$ be a rational relation and $\theta = \varphi^{-1}\,\iota_K\,\psi$ its realisation by morphisms given by Theorem IV.1.2, with $\varphi\colon Z^* \to A^*$ and $\psi\colon Z^* \to B^*$ being two morphisms, and K a rational of Z^*. *cf. p. 532*

Let T be a rational cross-section of K for φ. Then $\tau = \varphi^{-1}\,\iota_T\,\psi$ is a uniformisation of θ. Indeed, we have

$$\begin{tikzcd} Z^* \arrow[r, "\iota_K"] \arrow[d, "\varphi"'] & Z^* \arrow[d, "\psi"] \\ A^* \arrow[r, "\theta"'] & B^* \end{tikzcd}$$

$$\mathsf{Dom}\,\tau = T\varphi = K\varphi = \mathsf{Dom}\,\theta \ , \qquad \widehat{\tau} \subseteq \widehat{\theta} \ ,$$

and τ is functional since $\varphi^{-1}\,\iota_T$ is. □

3.1.2 The rational cross-section property for a monoid

We have noted that we can define the rational cross-section property for a function from an arbitrary monoid M to a set E. In fact, it is more interesting, having fixed the monoid M, to consider from this point of view functions from a free monoid *to* M.

Definition 3.2 We shall say that a monoid M *has the rational cross-section property* if every morphism from a free monoid to M has this property. □

With this definition, Theorem 3.1 becomes '*Every free monoid has the rational cross-section property*'. From this result we also deduce the following proposition, on which rest all the later proofs that a given monoid has the rational cross-section property.

Proposition 3.2 *A monoid has the rational cross-section property if and only if there exists a surjective morphism from a free monoid to M which has this property.*

We first show:

Lemma 3.3 *Let M be a monoid such that* $\mathrm{Rat}\,M$ *is closed under intersection and let α be a function from M to a monoid N such that*

(i) $(\mathrm{Rat}\,M)\alpha \subseteq \mathrm{Rat}\,N$ *and* (ii) $(\mathrm{Rat}\,N)\alpha^{-1} \subseteq \mathrm{Rat}\,M$.

Let β be a function from N to a set E. If α and β have the rational cross-section property then so does $\alpha\beta$.

[26] This is the *Rational Cross-Section Theorem* of [80].

Proof. Let R be a rational of M; then $S = R\alpha$ is a rational of N. Let V be a rational cross-section of S for β; then $U = R \cap V\alpha^{-1}$ is a rational of M. Let T be a rational cross-section of U for α; it follows that

$$T\alpha = U\alpha = (R \cap V\alpha^{-1})\alpha = R\alpha \cap V = V\,, \quad \text{and hence} \quad T\alpha\beta = V\beta = S\beta = R\alpha\beta\,.$$

Furthermore, $\alpha\beta$ is injective on T since α is on T and β on $V = T\alpha$: then T is a cross-section of R for $\alpha\beta$, rational by construction. ■

Proof of Proposition 3.2. Let $\alpha\colon A^* \to M$ be an arbitrary morphism and $\sigma\colon B^* \to M$ a surjective morphism which has the rational cross-section property. There exists a morphism $\theta\colon A^* \to B^*$ such that $\alpha = \theta\sigma$, and α has the rational cross-section property by Theorem 3.1 and Lemma 3.3. ■

cf. Lem. 0.4.1, p. 25

The following properties are simple exercises.

Property 3.1 *If M is a monoid with the rational cross-section property, then every rational function from a free monoid to M has this property.* ■

Property 3.2 *If M is a monoid with the rational cross-section property then every rational of M is* unambiguous *(that is,* $\mathrm{Rat}\,M = \mathrm{URat}\,M$ *).* ■

3.1.3 Return to simplification monoids

The last property shows that Definition 3.2 makes sense inasmuch as certain monoids have the rational cross-section property and others do not. (The converse of Property 3.2 is even not true.) The following result ensures that a remarkable family of monoids do.

cf. Exer. 3.3

Theorem 3.2 *The free group, the involutive monoid, and every quotient of a free monoid by a simplification have the rational cross-section property.*

Proof. We reuse the definitions and notation of Section II.6.3 and, more precisely, those of Proposition II.6.18 and of its proof. Let Π be a simplification over A^*. The morphism α_Π and the reduction η_Π have the same map equivalence. The reduction η_Π is written:

cf. Def. II.6.2, p. 315

cf. Prop. II.6.7, p. 309

$$\eta_\Pi = \theta_\Pi \circ \iota_{S_\Pi}$$

where θ_Π is the relation from A^* to itself which makes each word f correspond to the set of words obtained from f by derivation:

$$\begin{tikzcd} A^* \arrow[rr, "\alpha_\Pi"] \arrow[dr, "\eta_\Pi"'] & & T_\Pi \arrow[dl, shift left, "\iota_\Pi"] \\ & A^* \arrow[ur, shift left, "\alpha_\Pi"] & \end{tikzcd}$$

$$\forall f \in A^* \qquad f\theta_\Pi = \{g \in A^* \mid f \underset{\Pi}{\overset{\star}{\longmapsto}} g\}\ .$$

We write U_Π^* for the set of words in A^* which reduce to 1_{A^*} under η_Π. Let R be a rational of A^* recognised by an automaton $\mathcal{A}$ whose matrix representation is (λ, μ, ν). We set

$$H_{\mathcal{A}} = \sum_{w \in U_\Pi^*} w\mu\ ,$$

and the set $R\theta_\Pi$ is recognised by the automaton $\mathcal{A}\theta_\Pi$, whose representation (λ, π, ζ) is defined by *cf. Prop. II.6.19, p. 315*

$$\forall a \in A \qquad a\pi = H_{\mathcal{A}} \cdot a\mu \qquad \text{and} \qquad \zeta = H_{\mathcal{A}} \cdot \nu \ .$$

We shall construct a rational relation φ_R from A^* to itself which has the following properties:

(i) the domain of φ_R is $R\theta_\Pi$;

(ii) the image of φ_R is contained in R;

(iii) and for all g in $R\theta_\Pi$, $g\varphi_R$ consists of words f such that $f \overset{\star}{\underset{\Pi}{\vdash}} g$, hence we have $(g\varphi_R)\eta_\Pi = g\eta_\Pi$.

Theorem 2.1 applied to φ_R gives a rational function τ_R whose domain is $R\theta_\Pi$, and for each g in $R\theta_\Pi$, $g\tau_R$ is a word of R such that $(g\tau_R)\eta_\Pi = g\eta_\Pi$. In particular, $(g\tau_R)\eta_\Pi = g$ if g is in S_Π. The set $T = (S_\Pi)\tau_R$ is rational, contained in R. We have $T\eta_\Pi = S_\Pi \cap \mathsf{Dom}\,\tau_R = R\eta_\Pi$, and η_Π is injective on T since τ_R is a function. It remains to construct the relation φ_R.

We set $a\mu' = a\,a\mu$, for all a in A; the representation (λ, μ', ν) realises ι_R. We write $H'_{\mathcal{A}}$, the image of U_Π^* under μ':

$$H'_{\mathcal{A}} = \sum_{w \in U_\Pi^*} w\mu' \ ,$$

a matrix whose non-zero entries are *non*-rational subsets of A^*. Let ψ_R be the relation from A^* to itself realised by the representation (λ, π', ζ') which is defined by

$$\forall a \in A \qquad a\pi' = H'_{\mathcal{A}} \cdot a\mu' \qquad \text{and} \qquad \zeta' = H'_{\mathcal{A}} \cdot \nu \ .$$

By following the proof of Proposition II.6.19 step by step, it follows that, for all $g = a_1\,a_2 \cdots a_n$ in A^* *cf. p. 315*

$$\begin{aligned}
(a_1\,a_2 \cdots a_n)\psi_R &= \lambda \cdot a_1\pi' \cdot a_2\pi' \cdot \ \cdots \ \cdot a_n\pi' \cdot \zeta' \\
&= \lambda \cdot H'_{\mathcal{A}} \cdot a_1\mu' \cdot H'_{\mathcal{A}} \cdot a_2\mu' \cdot H'_{\mathcal{A}} \cdot \ \cdots \ \cdot H'_{\mathcal{A}} \cdot a_n\mu' \cdot H'_{\mathcal{A}} \cdot \nu \\
&= \left[\sum_{w_0, w_1, \ldots, w_n \in U_\Pi^*} \lambda \cdot w_0\mu \cdot a_1\mu \cdot w_1\mu \cdot \right. \\
&\qquad\qquad \left. a_2\mu \cdot w_2\mu \cdot \ \cdots \ \cdot w_{n-1}\mu \cdot a_n\mu \cdot w_n\mu \cdot \nu \right] \\
&= \{ f \in R \mid \exists w_0, w_1, \ldots, w_n \in U_\Pi^* \\
&\qquad\qquad f = w_0\,a_1\,w_1\,a_2\,w_2 \cdots w_{n-1}a_n\,w_n \} \\
&= \{ f \in R \mid f \overset{\star}{\underset{\Pi}{\vdash}} g \} \ .
\end{aligned}$$

We then define a matrix $G_{\mathcal{A}}$ which has the same support as $H'_{\mathcal{A}}$ and each of whose entries is a word *chosen arbitrarily* from the corresponding entry of $H'_{\mathcal{A}}$, if it is non-zero:

$$\forall p, q \in Q \qquad (G_{\mathcal{A}})_{p,q} \in \left(H'_{\mathcal{A}}\right)_{p,q} \ .$$

From this matrix $G_{\mathcal{A}}$ we define the *rational* relation φ_R, realised by the representation $(\lambda, \pi'', \zeta'')$, with

$$\forall a \in A \qquad a\pi'' = G_{\mathcal{A}} \cdot a\mu' \qquad \text{and} \qquad \zeta'' = G_{\mathcal{A}} \cdot \nu \ .$$

As we have $\quad \mathsf{supp}\, a\pi'' = \mathsf{supp}\, G_{\mathcal{A}} \cdot \mathsf{supp}\, a\mu' = \mathsf{supp}\, H'_{\mathcal{A}} \cdot a\mu = H_{\mathcal{A}} \cdot a\mu = a\pi \quad$ and

$$\mathsf{supp}\, \zeta'' = \mathsf{supp}\, G_{\mathcal{A}} \cdot \nu = \mathsf{supp}\, H'_{\mathcal{A}} \cdot \nu = H_{\mathcal{A}} \cdot \nu = \zeta \ ,$$

the domain of φ_R is the subset of A^* recognised by (λ, π, ζ) ; that is, $R\theta_\Pi$. The graph of φ_R is contained in that of ψ_R, hence in particular the image of φ_R is contained in R and we have, for all g, $(g\varphi_R)\eta_\Pi = g\eta_\Pi$, which completes the proof. ■

Corollary 3.4 *The restriction of the reduction η_Π to every rational on which it is injective is a rational function.* ■

Exercises & Problems

• 3.1 **The return of the Rational Cross-Section Theorem.** The aim of this exercise is to prove Theorem 3.1 directly. Let $A = \{a_1, a_2, \ldots, a_n\}$ be a totally ordered alphabet: $a_1 < \cdots < a_n$. From this order on A, we define the *lexicographic order*, a total order on A^*.

cf. Exer. IV.1.3, p. 528

(a) An elementary projection is the projection of a free monoid A^* onto one of the free monoids $(A \setminus a)^*$ (where a is one of the letters of A). Show that every morphism $\theta\colon A^* \to B^*$ can be decomposed into the product of an injective morphism by elementary projections.

(b) Let π be the (elementary) projection which removes the greatest letter of A. Let $\tau\colon A^* \to A^*$ be the relation which relates to a word f in A^* all the words in A^* that are strictly greater than f in the lexicographic order and that have the same image as f under π. Show that τ is a rational relation.

(c) Let R be a subset of A^*. Show that $R \setminus R\tau$ is a cross-section of R for π.

(d) Show that every elementary projection has the rational cross-section property. Show that every morphism $\theta\colon A^* \to B^*$ has the rational cross-section property.

• 3.2 Prove Properties 3.1 and 3.2.

• 3.3 Show that $\mathbb{N}^2$ does not have the rational cross-section property.

• 3.4 (a) Show that every morphism $\alpha\colon M \to N$ between commutative monoids has the rational cross-section property.

(b) Deduce:

Proposition 3.5 *Every rational relation between commutative monoids is uniformised by an unambiguous rational function.*

• 3.5 (a) Prove the following property, which is a weaker version of the rational cross-section property for the monoids $\mathbb{N}^k$:

Proposition 3.6 *Let $\alpha\colon A^* \to \mathbb{N}^k$ be a morphism. For every rational R of A^*, there exists an integer n and a rational S in R such that $S\alpha = R\alpha$ and each class modulo α contains at most n elements of S.*

(b) Show that if $k \geqslant 3$, the above integer n cannot be chosen independently of R.

(c) Show that if $k = 2$, we can take $n = 2$ in the above proposition, independently of R.

• 3.6 Show that the Rational Uniformisation Theorem holds for relations from $A_1^* \times A_2^* \times \cdots \times A_k^*$ to $B_1^* \times B_2^* \times \cdots \times B_l^*$ if and only if $k = 1$ or $\|A_i\| = \|B_j\| = 1$ for all $i \in [k]$ and all $j \in [l]$.

3.2 Choosing the uniformisation (or the cross-section)

The Rational Uniformisation Theorem, like its dual the Rational Cross-Section Theorem, has an unsatisfactory aspect. They ensure the existence of a rational function (of a rational set) attached to a rational relation (to a morphism) but they say nothing further about this function (on this set). What is even worse, this function, or this set, whose existence is proved constructively, *is not obtained in a manner intrinsic to the relation or morphism*, but depends on the realisation of the relation, and on an arbitrary series of choices (and also on the proof method).

With regard to the results that show the *existence* of rational selection functions, it is reasonable to seek to know whether, or under what conditions, 'natural' selection functions, like radix or lexicographic selection (*cf.* Example 2.1), realise rational uniformisations (or cross-sections). In all generality, the answer to this question is negative and it is not difficult to give an example of a relation whose radix or lexicographic selection is not rational. By means of the duality between uniformisation and cross-section, it is equivalent to showing that the radix (or lexicographic) cross-section of certain rationals for certain functions is not rational.

Example 3.1 Let $A = \{a, b, c\}$ be an ordered alphabet, with $a < b < c$, and let θ be the rational relation from A^* to itself whose graph is

$$\widehat{\theta} = (a, a)^* (b, 1)^* (1, b) \cup (a, 1)^* (b, a)^* (1, c) \ .$$

The domain of θ is $a^* b^*$ and we have $(a^n b^m)\theta = \{a^n b, a^m c\}$, from which we deduce

$$(a^n b^m)\theta_{\mathsf{rad}} = \begin{cases} a^n b & \text{if } n \leqslant m\,, \\ a^n c & \text{otherwise,} \end{cases} \quad \text{and} \quad (a^n b^m)\theta_{\mathsf{lex}} = \begin{cases} a^n b & \text{if } n \geqslant m\,, \\ a^n c & \text{otherwise.} \end{cases}$$

Neither of these uniformisations is rational. □

However, in two cases, those of synchronous relations and of deterministic relations, the 'natural' selections merit further attention.

3.2.1 Uniformisation of synchronous relations

Proposition 3.7 *The lexicographic selection and radix uniformisation of a synchronous relation are synchronous functions.*

Proof. Let $\theta \colon A^* \to B^*$ be a relation and α a (total) order on B^*. Fundamentally, α is a relation: $(u, v) \in \alpha$ if and only if u is less than or equal to v for α. We write α' for the *strict* order relation associated with α; that is, $\alpha' = \alpha \setminus \iota$, where ι is the identity relation.

For all f in A^*, the set of elements of B^* which are strictly greater than an element of $f\theta$ is $f\,[\theta\,\alpha']$, and hence – if it exists – the smallest element of $f\theta$ for α is the unique element of $(f\theta) \setminus (f\,[\theta\,\alpha'])$, if it is non-empty. *Selection with respect to* α of θ – *uniformisation with respect to* α of θ if α is a well ordering – is hence the function θ_α from A^* to B^* defined by

$$\theta_\alpha = \theta \setminus \left[\theta\,\alpha'\right] \quad .$$

cf. Exam. IV.6.4, p. 610

Lexicographic order and radix order are synchronous relations. Since the composition and difference of two synchronous relations are synchronous relations, θ_{lex} and θ_{rad} are rational functions. ■

cf. Prop. IV.6.16, p. 612 and Prop. IV.6.18, p. 613

Since the inverse of a synchronous relation is synchronous and intersection with a rational is also a synchronous relation, we have

Corollary 3.8 *Let* $\theta\colon A^* \to B^*$ *be a synchronous rational function. The radix cross-section and lexicographic selection of a rational of* A^* *for* θ *are rational.* ■

cf. Rem. IV.1.2, p. 533

Note that the radix cross-section T (or the lexicographic selection) of a subset R of A^* *is not* obtained as the image of R under a rational relation from A^* to itself, since it is not obtained as the image of R under an arbitrary relation from A^* to itself: the map that sends R to T *is not additive.*

3.2.2 Uniformisation of deterministic relations

The following example shows that the radix uniformisation of a deterministic relation is not necessarily a rational function.

Example 3.2 Let $A = \{a, b, c\}$ be an ordered alphabet, with $a < b < c$, and let π be the projection $\pi\colon \{a, b, c\}^* \to \{a, c\}^*$. We take $L = a^*(b\,c)^* \cup (a\,b)^*c^* : L\pi = a^*c^*$. For all m and n, we have:

$$(a^m c^n)\pi^{-1} \cap L = \{a^m(b\,c)^n, (a\,b)^m c^n\} \quad .$$

The radix cross-section of L for π (which is the image of the radix uniformisation of the deterministic relation $\iota_L \pi^{-1}$) is

cf. Propty IV.5.5, p. 595

$$T = \{a^m(b\,c)^n \mid m \leqslant n\} \cup \{(a\,b)^m c^n \mid m < n\}$$

and is not rational. □

But all is not lost, and the following result shows that it is not just the properties of the free monoid and of the rational subsets that matter, but also properties of the lexicographic order.

Theorem 3.3 [**Johnson**] *The lexicographic selection of a deterministic rational relation is a deterministic rational function.*

Proof. Let $\mathcal{T} = \langle Q_A, Q_B, A_\$, B_\$, \delta, i, T \rangle$ be a deterministic transducer which realises the relation θ; without loss of generality we can suppose that $T \subseteq Q_A$. We write *cf.* § *IV.5.2.3, p. 582*

$$p \circ (f,u) = \big(p, (f,u)\big)\delta = q \quad \text{or} \quad \emptyset \ ,$$

to mean that q is the state reached by the unique computation of $\mathcal{T}$, if it exists, whose label is (f,u). This 'almost action' is extended to subsets of states by the usual formula:

$$P \circ (f,u) = \bigcup_{p \in P} p \circ (f,u) \ .$$

We shall construct, in three steps, a transducer $\mathcal{M}$, which is a sort of product of $\mathcal{T}$ with an automaton $\mathcal{U}$, and which realises θ_{lex}. The automaton $\mathcal{U}$ is fundamentally an automaton over $A_\$$, even if we extend it to an automaton over $A_\$ \cup B_\$$ by giving the letters of $B_\$$ the identity's role. The idea is that a transition labelled f in $\mathcal{U}$ simulates the *set* of transitions in $\mathcal{T}$ with label (f,u) for all the u in $B_\$^*$. To this end we define the function τ from Q_B to $\mathfrak{P}(Q_A)$ by

$$\forall q \in Q_B \qquad q\tau = \{p \in Q_A \mid \exists v \in B_\$^* \quad q \circ (1,v) = p\} \ . \tag{3.1}$$

This function τ is extended to $Q_A \cup Q_B$ by $p\tau = p$ for all p in Q_A, which is consistent with (3.1) since the only word u of $B_\$^*$ for which $p \circ (1,u)$ is defined is 1_{B^*}. The (deterministic) automaton $\mathcal{U} = \langle \mathfrak{P}(Q_A), A_\$ \cup B_\$, \sigma, \{\emptyset\}, \mathfrak{P}(Q_A \setminus T) \rangle$ is defined by

$$\forall P \subseteq Q_A \,,\ \forall a \in A_\$ \,,\ \forall x \in B_\$ \qquad P \cdot a = \big(P \circ (a,1)\big)\tau \qquad \text{and} \qquad P \cdot b = P \ .$$

At this stage the initial and final states of $\mathcal{U}$ play no part. There follows

$$\forall P \subseteq Q_A \,,\ \forall f \in A_\$^* \qquad P \cdot f = \{q \in Q_A \mid \exists v \in B_\$^* \,,\ \exists p \in P \quad p \circ (f,v) = q\} \ .$$

We then consider the product $\mathcal{T} \times \mathcal{U}$ with the transition function $\zeta = \delta \times \sigma$:

$$\begin{aligned} &\mathcal{T} \times \mathcal{U} = \langle Q_A \times \mathfrak{P}(Q_A), Q_B \times \mathfrak{P}(Q_A), A_\$, B_\$, \zeta, (i, \{\emptyset\}), T \times \mathfrak{P}(Q_A \setminus T) \rangle \ , \\ &\forall r \in Q_A \cup Q_B \,,\ \forall P \subseteq Q_A \,,\ \forall (f,u) \in A_\$^* \times B_\$^* \\ &\qquad\qquad \big((r,P),(f,u)\big)\zeta = \big(r \circ (f,u), P \cdot f\big) \ . \end{aligned}$$

In particular, and since $\{\emptyset\} \in \mathfrak{P}(Q_A \setminus T)$, the two transducers $\mathcal{T}$ and $\mathcal{T} \times \mathcal{U}$ are equivalent: $|\mathcal{T} \times \mathcal{U}| = |\mathcal{T}|$. The transducer $\mathcal{M}$ is obtained from $\mathcal{T} \times \mathcal{U}$ by adding some transitions:

$$\mathcal{M} = \langle Q_A \times \mathfrak{P}(Q_A), Q_B \times \mathfrak{P}(Q_A), A_\$, B_\$, \xi, (i, \{\emptyset\}), T \times \mathfrak{P}(Q_A \setminus T) \rangle \ ,$$

where ξ is given by

$$\begin{aligned} &\forall r \in Q_A \cup Q_B \,,\ \forall P \subseteq Q_A \,, \\ &\qquad \forall a \in A_\$ \qquad \big((r,P),(a,1)\big)\xi = \big(r \circ (a,1), P \cdot a\big) = \big((r,P),(a,1)\big)\zeta \ , \\ &\qquad \forall x \in B_\$ \qquad \big((r,P),(1,x)\big)\xi = \left(r \circ (1,x), P \cdot x \cup \left[\bigcup_{y<x} \big(r \circ (1,y)\big)\tau\right]\right) \ . \end{aligned}$$

Remember the Corrugated Cardboard Lemma, Lem. IV.5.5, p. 581.

On the first component, the computations in $\mathcal{M}$ are identical to those of $\mathcal{T}\times\mathcal{U}$ hence $|\mathcal{M}| \subseteq |\mathcal{T}|$. For each pair (f,u) which it accepts, the transducer $\mathcal{T}$ determines a unique factorisation:

$$(f,u) = (g_1,1)\,(1,v_1)\,(g_2,1)\,(1,v_2)\cdots(g_n,1)\,(1,v_n)\,(g_{n+1},1) \ ,$$

with the g_i in $A_\$^+$ and the v_j in $B_\$^+$, a factorisation such that, among other things, for all j and every factorisation $v_j = w'\,x\,w''$ the computation

$$i \circ \big[(g_1,1)\,(1,v_1)\cdots(g_j,1)\,(1,w'\,x)\big]$$

is defined in $\mathcal{T}$ and hence in $\mathcal{T}\times\mathcal{U}$ and in $\mathcal{M}$. We then find in $\mathcal{M}$ a computation of the form

$$(i,\{\emptyset\}) \xrightarrow[\mathcal{T}\times\mathcal{U}]{(g_1,1)\cdots(g_j,1)(1,w')} (r,\{\emptyset\}) \xrightarrow[\mathcal{M}]{(1,x)} (s,P) \xrightarrow[\mathcal{T}\times\mathcal{U}]{(1,w'')(g_{j+1},1)\cdots(g_{n+1},1)} (p,R)$$

with $P = \big(r\circ(1,y)\big)\tau$ for some $y < x$ in $B_\$$. It follows that

$$(p,R) = \big((i,\{\emptyset\}),(f,u)\big)\xi \ ,$$

if and only if $i\circ(f,u) = p$ holds and R is the set of states q such that there exists a factorisation $u = u'\,x\,u''$, a letter $y < x$ and a word w in $B_\$^*$ such that $q = i\circ(f,u'\,y\,w)$. This amounts to saying that

$$\text{if} \quad (i,\{\emptyset\}) \xrightarrow[\mathcal{M}]{(f,u)} (p,R) \quad \text{then} \quad R = \{q \mid \exists w \in B_\$^* \quad w \not< u \quad q = i\circ(f,w)\} \ ,$$

and hence, given the choice of the final states of $\mathcal{M}$, that (f,u) is accepted by $\mathcal{M}$ if and only if (f,u) is accepted by $\mathcal{T}$ and there exists no word w lexicographically less than u such that (f,w) is accepted by $\mathcal{T}$: this is in fact the definition of lexicographic uniformisation and $|\mathcal{M}| = \theta_{\mathsf{lex}}$. ■

cf. Prop. IV.5.22, p. 592

We deduce in particular from Theorem 3.3:

Corollary 3.9 *Let $\theta\colon A^* \to B^*$ be a continuous morphism. The lexicographic cross-section of a rational R of A^* for θ is rational.* ■

As we said in Example 2.1 (iii), the choice of a cross-section for an equivalence relation θ on A^* comes down to defining a uniformisation of θ (viewed as the relation which associates a word with all those which are equivalent to it modulo θ) but the converse is not true. If θ is rational, Theorem 2.1 ensures the existence of a rational uniformisation τ, but this function τ does not *a priori* have any reason to be constant over each class and thereby to define a cross-section of θ.

The preceding result raises the question again. The lexicographic uniformisation θ_{lex} is constant over each class of θ and defines a cross-section, the *lexicographic cross-section* of θ. Theorem 3.3 implies therefore that the lexicographic cross-section of a deterministic rational equivalence relation is rational. We can construct rational equivalence relations θ whose lexicographic cross-section is not rational, but this does not imply that there is no cross-section of θ which is rational. We are thus led to consider the following problem.

Problem 3.1 [**Johnson**] *Does a rational equivalence relation on a free monoid always have a rational cross-section?*

3.2.3 Theorem 3.3 back on the loom

It is frustrating[27] that the proof of the beautiful Theorem 3.3 consists of an *ad hoc* construction and does not find its natural place with the Rational Uniformisation Theorem and the Weak Decomposition Theorem, as another consequence of the construction of the Schützenberger covering which we have already showed carefully in Chapter II. In fact, the aim of weaving Theorem 3.3 into the cloth of our study is not out of the question: a weaker proposition, but which loses nothing essential, fits in perfectly with the general pattern that we have laid down.

Proposition 3.10 *The lexicographic selection of a deterministic rational relation is an unambiguous rational function.*

A little preparation is necessary. In what follows we assume that the alphabets are (totally) ordered and that if an order is chosen on an alphabet with endmarker $B_\$$, \$ is always the *smallest letter*. We know that the product is almost compatible with the lexicographic order (on $B_\$^*$) in the sense that if $u = f\,x\,h$ and $u' = f\,y\,k$ with f, h and k in $B_\$^*$, x and y in B and $x \not\preccurlyeq y$, then $u\,v \not\preccurlyeq u'\,v'$ for all v and v' in $B_\$^*$. The converse of this assertion gives:

Property 3.3 $\forall u, u', v, v' \in B_\$^*$

$$(u'\,v' \preccurlyeq u\,v \quad \textit{and} \quad u \not\preccurlyeq u') \quad \Longrightarrow \quad u \textit{ is a prefix of } u' \,. \qquad \blacksquare$$

As we shall work with *matrix representations* of relations, it is useful to define the *lexicographic selection* of a matrix:

Definition 3.3 Let M be a $P \times Q$ matrix with entries in $\mathfrak{P}(B^*)$. For each p in P, let w_p be the smallest element in the lexicographic order of $M_{p\bullet} = \sum_{q \in Q} M_{p,q}$, *if it exists*. The *lexicographic selection* of M, written M_{lex}, is the matrix defined by

$$\forall p \in P\,,\ \forall q \in Q \qquad (M_{\mathsf{lex}})_{p,q} = \begin{cases} w_p & \text{if} \quad w_p \in M_{p,q} \ , \\ 0 & \text{otherwise} \ . \end{cases} \qquad \square$$

Note that, by definition, each non-zero entry of M_{lex} is a monomial; that is, is a word of B^*, and that if M is *prefix*, M_{lex} is *row monomial*. If X and Y are matrices of the same dimension over $\mathfrak{P}(B^*)$, we say that X is *contained* in Y if for all indices (p, q), $X_{p,q} \subseteq Y_{p,q}$. There then follows:

Lemma 3.11 *Let M and N be two matrices over $\mathfrak{P}(B^*)$. If M is a* prefix matrix, *then $M_{\mathsf{lex}} N_{\mathsf{lex}}$ is contained in $(M\,N)_{\mathsf{lex}}$.* $\blacksquare$

[27] Perhaps not for the reader, but certainly for the author. *Note added in translation:* the frustration has been alleviated by the recent work of R. de Souza and the author which reunites all these constructions in a common framework; see bibliographic notes.

cf. § IV.5.5.2, p. 592

Proof of Proposition 3.10. Let θ be a deterministic relation from A^* to B^* and let (λ, μ, ν) be a *matrix representation with endmarker* (of dimension Q) which is *prefix* and which realises θ; that is,

$$\forall (f,g) \in A^* \times B^* \qquad (f,g) \in \widehat{\theta} \quad \Longleftrightarrow \quad g\$ \in \lambda \cdot f\mu \cdot \nu \ .$$

Since \$ is the smallest letter of $B^*_{\$}$, it follows that

$$g = f\theta_{\mathsf{lex}} \quad \Longleftrightarrow \quad g\$ \in (\lambda \cdot f\mu \cdot \nu)_{\mathsf{lex}} \ .$$

We reuse the notation used in the solution of Exer. 2.2, p. 673.

Let $\mathcal{C}$ be the transducer defined by (λ, μ, ν) and $\mathcal{A}$ its underlying input automaton. Let (η', κ', ξ') be the (boolean) representation, of dimension $\mathfrak{P}(Q)$, of $\mathcal{A}_{\mathsf{cod}}$. Let, as above, $(\zeta', \sigma', \omega')$ be the representation of the S-co-covering of $\mathcal{C}$ and let (χ'', τ'', ψ'') be the representation of the 'lexicographic S-co-selection' defined as follows:

$$\begin{aligned} \forall S \subseteq Q \qquad & \chi''_{(S,Q)} = \left[\zeta'_{(S,Q)}\right]_{\mathsf{lex}} = \left[\lambda^{|S|}\right]_{\mathsf{lex}} \ , \\ \forall a \in A\,,\ \forall P, S \subseteq Q \qquad & a\tau''_{(P,Q),(S,Q)} = \left[a\sigma'_{(P,Q),(S,Q)}\right]_{\mathsf{lex}} = \left[a\mu^{|S|}\right]_{\mathsf{lex}} \ , \\ \forall P \subseteq Q \qquad & \psi''_{(S,Q)} = \left[\omega'_{(P,Q)}\right]_{\mathsf{lex}} \ . \end{aligned}$$

In other words, the choices which must be made to construct an S-co-selection are no longer arbitrary but are determined by the fact that we take the lexicographic selection of each block in the S-co-covering of $\mathcal{C}$. This procedure for making the blocks row monomial determines at the same time the unambiguous co-immersion in $\mathcal{A}$ and the choice of the words in the entries that are retained.

To conclude, we must deal with the details rather more than for the proof of Theorem 2.1. Let $f = a_1 a_2 \cdots a_n$ be a word of A^* and let

$$P_0 \xrightarrow{a_1} P_1 \xrightarrow{a_2} P_2 \cdots P_{n-1} \xrightarrow{a_n} P_n$$

be the *unique* successful computation with label f in $\mathcal{A}_{\mathsf{cod}}$. We have

$$\begin{aligned} \lambda \cdot f\mu \cdot \nu &= \zeta' \cdot f\sigma' \cdot \omega' \\ &= \zeta'_{(P_0,Q)} \cdot a_1\sigma'_{(P_0,Q),(P_1,Q)} \cdot a_2\sigma'_{(P_1,Q),(P_2,Q)} \cdot \cdots \cdot a_n\sigma'_{(P_{n-1},Q),(P_n,Q)} \cdot \omega'_{(P_n,Q)} \ . \end{aligned}$$

Since each of the matrices on the right hand side of the equation is *prefix* (as a matrix *contained* in a prefix matrix), it follows, by Lemma 3.11, that

$$\begin{aligned} [\lambda \cdot f\mu \cdot \nu]_{\mathsf{lex}} \ni & \left[\zeta'_{(P_0,Q)}\right]_{\mathsf{lex}} \cdot \left[a_1\sigma'_{(P_0,Q),(P_1,Q)}\right]_{\mathsf{lex}} \cdot \cdots \\ & \cdots \cdot \left[a_n\sigma'_{(P_{n-1},Q),(P_n,Q)}\right]_{\mathsf{lex}} \cdot \left[\omega'_{(P_n,Q)}\right]_{\mathsf{lex}} = \chi'' \cdot f\tau'' \cdot \psi'' \ . \end{aligned}$$

Since we know, from the proof of Theorem 2.1, that $\chi'' \cdot f\tau'' \cdot \psi''$ is different from 0, it follows that

$$[\lambda \cdot f\mu \cdot \nu]_{\mathsf{lex}} = \chi'' \cdot f\tau'' \cdot \psi'' \ .$$

Thus, (χ'', τ'', ψ'') is a representation of θ_{lex}; it is unambiguous since it is an S-co-immersion. ∎

Example 2.2 (continued) We reuse the notation of the solution of Exercise 2.2 (b). *cf. p. 728* The S-co-covering of $\mathcal{E}_2$ is $(\zeta_1', \sigma_1', \omega_1')$ and the representation $(\chi_1'', \tau_1'', \psi_1'')$ of the lexicographic uniformisation of θ_2, shown in Figure 3.2, is given by

$$a\tau_1'' = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \begin{matrix} \boxed{a}\,\boxed{0}\,0 \\ 0\;0\;0 \\ 0\;0\;a \end{matrix} & \begin{matrix} a\;0\;0 \\ 0\;0\;0 \\ 0\;0\;a \end{matrix} & \begin{matrix} 0\,\boxed{ba}\,0 \\ 0\;0\;0 \\ 0\;0\;a \end{matrix} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \begin{matrix} 0\;0\;0 \\ 0\;0\;0 \\ 0\;0\;a \end{matrix} \end{pmatrix}$$

and $b\tau_1'' = b\sigma_1'$, $\chi_1'' = \zeta_1'$, $\psi_1'' = \omega_1'$. □

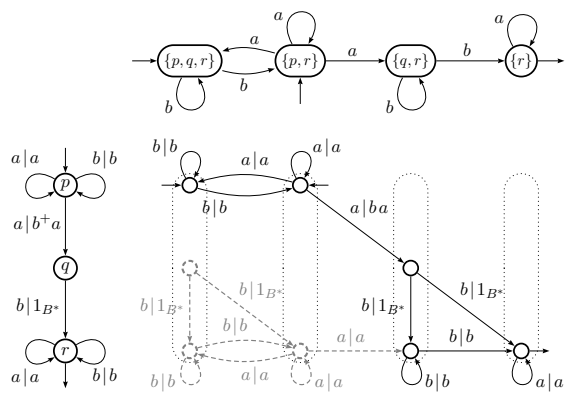


Figure 3.2: The transducer $\mathcal{E}_2$ and the lexicographic uniformisation of θ_2 obtained from an S-co-immersion

Exercises

- 3.7 Let $\alpha\colon \{a,b,c,d,e,f\}^* \to \{x,y,z\}^*$ be the morphism defined by

$$a\alpha = x \;,\quad b\alpha = yxyx \;,\quad c\alpha = xy \;,\quad d\alpha = yz \;,\quad e\alpha = zyzy \;,\qquad \text{and} \qquad f\alpha = z \;.$$

Show that the radix cross-section of $\{a,b,c,d,e,f\}^*$ for α is not rational.

- 3.8 Prove Lemma 3.11.

3.9 Let M and N be two matrices on $\mathfrak{P}(B^*)$ of respective dimensions $P{\times}Q$ and $Q{\times}R$. We say that N is a *continuation* of M if

$$\forall p \in P\,,\ \forall q \in Q \qquad M_{p,q} \neq 0 \quad\Longrightarrow\quad N_{q\,\bullet} \neq 0 \;.$$

Show that if M is prefix and N is a continuation of M, then $M_{\mathsf{lex}}\, N_{\mathsf{lex}} = (M\,N)_{\mathsf{lex}}$.

4 Sequential functions

Rightly named, rightly placed.

Han Fei-Tse

We now return to the study of sequential functions; the previous sections showed how central a place they occupy in the theory of rational functions.

We start by giving two characterisations of sequential functions in the family of rational functions, a functional characterisation and one which we call 'quasi-topological'. Then, working with transducers, we shall show how we can decide whether a rational function is sequential and how we can compute a minimal sequential transducer equivalent to a given sequential transducer. We finish this section by showing that the quasi-topological characterisation holds in a class much larger than the rational functions.

4.1 Two characterisations

4.1.1 Translations of a function

If f and g are two words, we write $f \wedge g$ for the *longest common prefix* of f and g: if $h = f \wedge g$, then either $f = h$ and $g = f\,v$, or $g = h$ and $f = g\,u$, or $f = h\,x\,u$ and $g = h\,y\,v$ and x and y are two distinct letters. If L is a subset of a free monoid B^*, we write

$$\bigwedge_{u \in L} u \ , \qquad \text{or, more simply,} \qquad \bigwedge L \ ,$$

for the *longest common prefix* of all the elements of L. By convention, $\bigwedge \emptyset = \emptyset$, but we must note that if $L \neq \emptyset$, we have $\bigwedge L \neq \emptyset$: at worst, $\bigwedge L = 1_{B^*}$.

To every function $\alpha \colon A^* \to B^*$, we then relate another function[28] from A^* to B^*, written $\overset{\circ}{\alpha}$, and which sends each word f to the longest common prefix of the images under α of all words which start with f:

$$\forall f \in A^* \qquad (f)\overset{\circ}{\alpha} \quad = \bigwedge_{\substack{g \in A^*, \\ fg \in \mathsf{Dom}\,\alpha}} (f\,g)\alpha \ . \tag{4.1}$$

In other words, if we imagine that α is realised by a machine which reads words of A^* letter by letter, then $(f)\overset{\circ}{\alpha}$ is the longest word that one is certain to output having read f – and which we do not know how or whether it will continue – *if* there is any output after reading.

Intuitively, we well understand that if α is a sequential function, $\overset{\circ}{\alpha}$ is 'close' to α – for example, $\overset{\circ}{\chi}_K = \chi_K$ (the characteristic function of the language K) – and that if $\overset{\circ}{\alpha}$ is very different from α, there is little chance that α is sequential; for example,

[28] Which I read 'α top' as it is short, but, in the absence of official terminology, readers may pronounce it as they like.

$(f)\overset{\circ}{\rho} = 1_{A^*}$ for all f, when ρ is the 'mirror' function. We shall formalise this intuition. In particular we deduce from the definition above that

$$\mathsf{Dom}\,\overset{\circ}{\alpha} = \mathsf{Pre}(\mathsf{Dom}\,\alpha) \qquad \text{and} \qquad (1_{A^*})\overset{\circ}{\alpha} = \bigwedge_{f \in \mathsf{Dom}\,\alpha} (f)\alpha = \bigwedge \mathsf{Im}\,\alpha \ .$$

We shall say that a function α is *frank* if $(1_{A^*})\overset{\circ}{\alpha} = 1_{B^*}$; that is, if there is no common non-empty left factor of the images of all the words of the domain of α. In fact, it is not so much the function $\overset{\circ}{\alpha}$ which will interest us as the functions it enables us to define.

Definition 4.1 Let $\alpha\colon A^* \to B^*$ be a function. For all f in A^*, the *(left) translation of α by f* is the function,[29] also from A^* to B^*, written $\alpha/\!\!f$, and defined, for all g to A^*, by

$$(g)\alpha/\!\!f = \left[(f)\overset{\circ}{\alpha}\right]^{-1} \left[(f\,g)\alpha\right] \ . \qquad \square$$

From this we deduce

$$\forall f, g \in A^* \qquad (f\,g)\alpha = \left[(f)\overset{\circ}{\alpha}\right] \left[(g)\alpha/\!\!f\right] \ . \tag{4.2}$$

As a special case of (4.2), we have

$$\forall f \in A^* \qquad f\alpha = \left[(1_{A^*})\overset{\circ}{\alpha}\right] \left[(f)\alpha/\!\!1_{A^*}\right] \tag{4.3}$$

$$= \left[(f)\overset{\circ}{\alpha}\right] \left[(1_{A^*})\alpha/\!\!f\right] \ . \tag{4.4}$$

We thus have $\alpha/\!\!1_{A^*} = \alpha$ if and only if $(1_{A^*})\overset{\circ}{\alpha} = 1_{B^*}$; that is, if and only if α is a frank function. We easily verify:

Property 4.1 $\forall f \in A^* \qquad \mathsf{Dom}\,\alpha/\!\!f = f^{-1}(\mathsf{Dom}\,\alpha)$. ■

Property 4.2 $\forall f, g \in A^* \qquad (f\,g)\overset{\circ}{\alpha} = \left[(f)\overset{\circ}{\alpha}\right] \left[(g)\overset{\circ}{\alpha/\!\!f}\right]$. ■

Property 4.1 shows the relationship between the translation of a function and the (left) quotient of a language. In particular, we have $\chi_K/\!\!f = \chi_{f^{-1}K}$. The operation '(left) quotient of a language' is a (right) action of A^* on $\mathfrak{P}(A^*)$; that is, *cf. Propty I.3.6, p. 112*

$$\forall f, g \in A^*\,,\ \forall X \subseteq A^* \qquad (f\,g)^{-1}X = g^{-1}\left[f^{-1}X\right] \ .$$

Translation, on the other hand, *is not an action*[30] of A^* on the set of functions from A^* to B^*, since the translation by 1_{A^*} is not the identity – but 'almost', since we have:

Lemma 4.1 *For every function $\alpha\colon A^* \to B^*$, and all f and g in A^*, we have*

$$\alpha/\!\!(f\,g) = (\alpha/\!\!f)/\!\!g \ . \qquad ■$$

It follows that every translation is a frank function and that translation is an action of A^* on the set of frank functions.

---

[29] Reutenauer calls it the *derivative of α with respect to f* and writes it, like Eilenberg, $\alpha \cdot f$.

[30] This is why I prefer to write it $\alpha/\!\!f$ rather than $\alpha \cdot f$.

4.1.2 A functional characterisation

The preceding definitions were made with the following result in mind.[31]

Theorem 4.1 [**Raney**] *A function* $\alpha\colon A^* \to B^*$ *is sequential rational if and only if the set of its translations* $\{\alpha/f \mid f \in A^*\}$ *is finite.*

Proof. (i) The condition is sufficient. We associate every function $\alpha\colon A^* \to B^*$ with a sequential transducer $\mathcal{T}_\alpha$ (not necessarily finite), called the *translation transducer*, in the following way. Let $R_\alpha = \{\alpha/f \mid f \in A^*\}$ be the set of translations of α
cf. Def. 1.2 and $\mathcal{T}_\alpha = \langle R_\alpha, A, B^*, \delta_\alpha, \eta_\alpha, J, U \rangle$ the sequential transducer defined by

$$\forall \alpha/f \in R_\alpha\,,\ \forall a \in A \qquad \alpha/f \cdot a = (\alpha/f, a)\delta_\alpha = \alpha/(f\,a)\ , \tag{4.5}$$

and by

$$\forall \alpha/f \in R_\alpha\,,\ \forall a \in A \qquad \alpha/f * a = (\alpha/f, a)\eta_\alpha = (a)\overset{\circ}{\alpha/f} = \left[(f)\overset{\circ}{\alpha}\right]^{-1}\left[(f\,a)\overset{\circ}{\alpha}\right]\ . \tag{4.6}$$

The initial state j is the translation $\alpha/1_{A^*}$, and, more precisely, the *initial function* is

$$\forall q \in R_\alpha \qquad J_q = \begin{cases} (1_{A^*})\overset{\circ}{\alpha} & \text{if} \quad q = \alpha/1_{A^*}\ , \\ 0 & \text{otherwise.} \end{cases} \tag{4.7}$$

and the final function U is defined by

$$\forall \alpha/f \in R_\alpha \qquad U_{\alpha/f} = (1_{A^*})\alpha/f\ . \tag{4.8}$$

Property 4.3 *For every function* $\alpha\colon A^* \to B^*$, $\mathcal{T}_\alpha$ *is a sequential transducer which realises* α *(and is finite if and only if the set of translations of* α *is finite).*

cf. Propty I.3.3, p. 105 *Proof.* Lemma 4.1 ensures that

$$\forall f, g \in A^* \qquad (\alpha/f) \cdot g = (\alpha/f, g)\delta_\alpha = \alpha/(f\,g)\ , \tag{4.9}$$

and hence in particular, for all f, that $i \cdot f = (\alpha/1_{A^*}, f)\delta_\alpha = \alpha/f$, and $\mathcal{T}_\alpha$ is accessible. We show, by induction on the length of g, that for all f and g in A^*,

$$(\alpha/f) * g = (\alpha/f, g)\eta_\alpha = (g)\overset{\circ}{\alpha/f}\ . \tag{4.10}$$

We have

$$\begin{aligned}
(\alpha/f, g\,a)\eta_\alpha &= \left[(\alpha/f, g)\eta_\alpha\right]\left[(\alpha/f \cdot g, a)\eta_\alpha\right] && \text{by definition, that is, by (1.4),} \\
&= \left[(g)\overset{\circ}{\alpha/f}\right]\left[(a)\overset{\circ}{\alpha/(f\,g)}\right] && \text{by induction and by (4.6),} \\
&= (g\,a)\overset{\circ}{\alpha/f} && \text{by Property 4.2.}
\end{aligned}$$

[31] Raney's original proposition applies to the case of letter-to-letter functions, but the substance is identical (*cf.* [192]).

Hence we deduce, for all f in A^*,

$$\begin{aligned}(f)|\mathcal{T}_\alpha| &= J_j\,(j * f)\,U_{j\cdot f} \quad , && \text{by (1.6)}\\ &= \left[(1_{A^*})\overset{\circ}{\alpha}\right]\left[(f)\alpha/\overset{\circ}{1}_{A^*}\right]\left[(1_{A^*})\alpha/f\right]\\ &= \left[(f)\overset{\circ}{\alpha}\right]\left[(1_{A^*})\alpha/f\right] = (f)\alpha \quad . \end{aligned}$$

■

(ii) The condition is necessary. Let $\mathcal{B} = \langle\, Q, A, B^*, \delta, \eta, i, T\,\rangle$ be a sequential transducer[32] and α the function that it realises:

$$\forall f \in A^* \qquad (f)\alpha = (f)|\mathcal{B}| = I_i\,(i * f)\,T_{i\cdot f} \quad .$$

For each q in Q, let $\mathcal{B}_q$ be the transducer deduced from $\mathcal{B}$ by taking q as the initial state, with an initial function equal to 1_{B^*} (note $\mathcal{B}_i$ is not equal to $\mathcal{B}$ unless $I_i = 1_{B^*}$):

$$\forall g \in A^* \qquad g|\mathcal{B}_q| = (q * g)\,T_{q\cdot g} \qquad\qquad (\text{that is, in particular, } g\alpha = I_i\,g|\mathcal{B}_i|) \quad .$$

We shall show that each translation α/f is determined by the state $i\cdot f$; more precisely:

Property 4.4 *For all f in A^*, $\alpha/f = |\mathcal{B}_{i\cdot f}|/1_{A^*}$.*

Proof. For all f and g in A^*, we have

$$(f\,g)\alpha = I_i\,(i * (f\,g))\,T_{i\cdot(fg)} = I_i\,(i * f)\,((i\cdot f) * g)\,T_{(i\cdot f)\cdot g} \quad ,$$

from which we deduce

$$(f)\overset{\circ}{\alpha} = I_i\,(i * f)\,(1_{A^*})\overset{\circ}{\alpha}_{i\cdot f} \quad ,$$

and hence

$$(g)\alpha/f = \left[(f)\overset{\circ}{\alpha}\right]^{-1}\left[(f\,g)\alpha\right] = (g)[\alpha_{i\cdot f}/1_{A^*}] \quad .$$

■

This completes the proof of Theorem 4.1. ■

We should see this result as a generalisation of Proposition I.3.10, a characterisation of rational languages according to the finiteness of the set of quotients, and of Theorem III.4.8, a characterisation of rational series according to the finiteness of their rank. We shall see in § 4.3 that the parallel can be pursued, and that the transducer $\mathcal{T}_\alpha$ that we have just constructed in the proof of Theorem 4.1 is the *minimal transducer* of the function under consideration, a generalisation, or counterpart, of the *minimal automaton* of a language.

cf. p. 114

cf. p. 452

[32] We write it $\mathcal{B}$, and not the seemingly more natural $\mathcal{T}$, since $\mathcal{T}_q$ and T_q would be too similar typographically.

4.1.3 A quasi-topological point of view

We define the *prefix distance* on the words of a free monoid by

$$\forall f, g \in A^* \qquad \mathbf{d}_{\mathrm{p}}(f, g) = |f| + |g| - 2|f \wedge g| \ .$$

In other words, if $f = h\,f'$ and $g = h\,g'$ with $h = f \wedge g$, then

$$\mathbf{d}_{\mathrm{p}}(f, g) = |f'| + |g'| \ .$$

This function $\mathbf{d}_{\mathrm{p}}$ is indeed a distance, but the topology that it defines on A^* is the discrete topology – that is, two distinct words have a minimum distance (equal to 1) – and the only convergent sequences are those that are stationary from a certain rank on. In fact, we shall use the prefix distance not to evaluate how *close* two words become, which is the usual aim of a topology, but rather how *far apart* they become, hence the 'quasi' of the paragraph's title.[33]

Definition 4.2 A function from one free monoid to another, $\alpha\colon A^* \to B^*$, is called *uniformly bounded*[34] if for every integer n there exists an integer N which is an upper bound on the prefix distance of the images under α of two arbitrary words (in the domain of α) whose prefix distance is less than n; that is,

$$\forall n \in \mathbb{N}\,,\ \exists N \in \mathbb{N}\,,\ \forall f, g \in \mathsf{Dom}\,\alpha \qquad \mathbf{d}_{\mathrm{p}}(f, g) \leqslant n \Longrightarrow \mathbf{d}_{\mathrm{p}}(f\alpha, g\alpha) \leqslant N \ . \qquad \square$$

Remark 4.1 Let us consider for a moment the defining equation of the *uniform continuity* of a function α (from one metric space to another):

$$\forall \varepsilon > 0\,,\ \exists \eta > 0\,,\ \forall x, y \in \mathsf{Dom}\,\alpha \qquad \mathbf{d}(x, y) \leqslant \eta \implies \mathbf{d}(x\alpha, y\alpha) \leqslant \varepsilon \ . \tag{4.11}$$

This equation implicitly assumes that we 'make ε tend to 0'. If we are thinking of 'making ε tend to $+\infty$', the universal quantifier over ε no longer makes sense since if we find, for some ε_0, an η_0 which satisfies (4.11), this same η_0 will satisfy the same equation for any ε greater than ε_0. We must therefore in this case exchange the roles of ε and η, and we obtain

$$\forall \eta > 0\,,\ \exists \varepsilon > 0\,,\ \forall x, y \in \mathsf{Dom}\,\alpha \qquad \mathbf{d}(x, y) \leqslant \eta \implies \mathbf{d}(x\alpha, y\alpha) \leqslant \varepsilon \ . \tag{4.12}$$

This last equation, which is that of Definition 4.2, says that the image of a bounded set (a ball of radius η) is included in a bounded set of fixed diameter ($2\,\varepsilon$) dependent on η but independent of the centre of the ball under consideration, in the domain of the definition of the entire function; hence the chosen terminology. When the ratio between η and ε can be bounded independently of η, we come back to a much more standard notion which we shall now see. It is important to note that we are interested in these properties not when the distance η tends to zero, which is not possible with the prefix distance, but when this distance becomes as large as we want. $\square$

[33] In this context, it would obviously be a terrible idea to replace $\mathbf{d}_{\mathrm{p}}$ by a distance that is topologically equivalent and bounded by 1 (*cf.* Rem. III.1.3, p. 387).

[34] Such a function is called 'of bounded variation' in [55, 27]. This phrase is already used with another meaning in other areas of mathematics. We use the terminology of [199].

Definition 4.3 A function from one free monoid to another, $\alpha\colon A^* \to B^*$, is said to be *a Lipschitz function*, or simply to be *Lipschitz*, if there exists an integer k which is an upper bound on ratio between the (prefix) distance of the images under α of two arbitrary words (in the domain of α) and the distance of these two words; that is,

$$\exists k \in \mathbb{N}\,,\ \forall f, g \in \mathsf{Dom}\,\alpha \qquad \mathbf{d}_{\mathrm{p}}\,(f\alpha, g\alpha) \leqslant k\,\mathbf{d}_{\mathrm{p}}\,(f, g) \quad . \qquad \square$$

Clearly, a Lipschitz function is uniformly bounded.

Proposition 4.2 *A sequential function is Lipschitz.*

Proof. Let $\mathcal{T} = \langle Q, A, B^*, \delta, \eta, I, T \rangle$ be a sequential transducer which realises a sequential function α and let ℓ be a majorant of the lengths of the outputs of $\mathcal{T}$:

$$\forall q \in Q\,,\ \forall a \in A \qquad |(q,a)\eta| \leqslant \ell \qquad \text{and} \qquad |T_q| \leqslant \ell \ .$$

Let f and g be in $\mathsf{Dom}\,\alpha$ and $h = f \wedge g$; we set $f = h\,f'$ and $g = h\,g'$. It follows that

$$(f)\alpha = I_i\,(i * h)\,((i \cdot h) * f')\,T_{(i\cdot f)} \qquad \text{and} \qquad (g)\alpha = I_i\,(i * h)\,((i \cdot h) * g')\,T_{(i\cdot g)} \ ,$$

hence

$$\begin{aligned}\mathbf{d}_{\mathrm{p}}\,((f)\alpha, (g)\alpha) &\leqslant |((i \cdot h) * f')\,T_{(i\cdot f)}| + |((i \cdot h) * g')\,T_{(i\cdot g)}| \\ &\leqslant \ell\,(|f'| + |g'| + 2) = \ell\,(\mathbf{d}_{\mathrm{p}}\,(f, g) + 2) \\ &\leqslant (\ell + 2)\,\mathbf{d}_{\mathrm{p}}\,(f, g) \quad . \qquad \blacksquare\end{aligned}$$

The converse of Proposition 4.2 is true for the class of functions whose inverses preserve rationality. We shall prove this remarkable result in all generality in Subsection 4.4. The converse for rational functions, a weaker version which we could call by contrast the Little Sequentiality Theorem, is easier to prove.

Theorem 4.2 *Let* $\alpha\colon A^* \to B^*$ *be a* rational function. *The three following conditions are equivalent:*

(i) α *is sequential;*

(ii) α *is Lipschitz;*

(iii) α *is uniformly bounded.*

Proof. We have just shown with Proposition 4.2 that (i) $\Rightarrow$ (ii); we know that (ii) $\Rightarrow$ (iii); it remains to show that (iii) $\Rightarrow$ (i).

Let (λ, μ, ν) be a trim representation (of dimension Q) of a uniformly bounded rational function α. For all f in A^*, write $(f)\zeta = \lambda \cdot f\mu$, which is a *row vector* of $(B^*)^Q$, and $(f)\overset{\circ}{\zeta}$ for the longest common prefix of all the non-zero entries of $(f)\zeta$:

$$(f)\overset{\circ}{\zeta} = \bigwedge_{p \in Q} \big((f)\zeta\big)_p \ .$$

Write $(f)\zeta^\sharp$ for the *row vector*[35] of $(B^*)^Q$ whose entries are the (left) quotient of the entries of $(f)\zeta$ by $(f)\overset{\circ}{\zeta}$:

$$(f)\zeta^\sharp = \left[(f)\overset{\circ}{\zeta}\right]^{-1}\left[(f)\zeta\right] \qquad \text{and we thus have} \qquad (f)\zeta = \left[(f)\overset{\circ}{\zeta}\right]\left[(f)\zeta^\sharp\right] \ .$$

We note that if $(f)\zeta^\sharp$ has at least one entry different from 1_{B^*}, there is at least one other non-zero entry. For each f in A^*, write β_f for the function from A^* to B^* defined by

$$\forall g \in A^* \qquad (g)\beta_f = (f)\zeta^\sharp \cdot g\mu \cdot \nu \ .$$

It follows that

$$(f\,g)\alpha = \lambda \cdot f\mu \cdot g\mu \cdot \nu = (f)\zeta \cdot g\nu \cdot \nu \quad = \left[(f)\overset{\circ}{\zeta}\right]\left[(f)\zeta^\sharp \cdot g\mu \cdot \nu\right] = (f)\overset{\circ}{\zeta}\,(g)\beta_f \ ,$$

from which we deduce

$$\alpha/f = \beta_f/1_{A^*} \ ,$$

and since $(f)\zeta^\sharp = (f')\zeta^\sharp$ implies that $\beta_f = \beta_{f'}$, it suffices, by Theorem 4.1, to show that there is only a finite number of distinct vectors $(f)\zeta^\sharp$.

Suppose that there is an infinite number of such vectors. As there is only a finite number of possible distinct supports, $(2^Q - 1)$ at most, there exist words h such that $(h)\zeta^\sharp$ has at least one entry which is an arbitrarily long word in B^*. Again since Q is finite, there exists a state p such that there exist words h such that $\left((h)\zeta^\sharp\right)_p$ is an arbitrarily long word in B^*. For each of these h, there exists a q such that $\left((h)\zeta\right)_p \wedge \left((h)\zeta\right)_q = (f)\overset{\circ}{\zeta}$ and hence such that $\left((h)\zeta^\sharp\right)_p \wedge \left((h)\zeta^\sharp\right)_q = 1_{B^*}$; by the same argument of compactness as above, there exists a state r such that there exist some h's such that $\left((h)\zeta^\sharp\right)_p$ is an arbitrarily long word in B^* and that $\left((h)\zeta^\sharp\right)_p \wedge \left((h)\zeta^\sharp\right)_r = 1_{B^*}$.

cf. Exer. 4.2

As (λ, μ, ν) is trim, there exist two fixed words f' and g' in A^* such that

$$u = \iota_p \cdot f'\mu \cdot \nu \qquad \text{and} \qquad v = \iota_q \cdot g'\mu \cdot \nu$$

are non-zero (where ι_p and ι_q are row vectors whose only non-zero entries are p and q respectively). Hence

$$(h\,f')\alpha = \left[(h)\overset{\circ}{\zeta}\right]\left[((h)\zeta^\sharp)_p\right] u \qquad \text{and} \qquad (h\,g')\alpha = \left[(h)\overset{\circ}{\zeta}\right]\left[((h)\zeta^\sharp)_q\right] v \ ,$$

and so $h\,f'$ and $h\,g'$ both belong to $\mathsf{Dom}\,\alpha$. However

$$\mathbf{d}_{\mathrm{p}}\left(h\,f', h\,g'\right) \leqslant |f'| + |g'|$$

is bounded independently of h, and

$$\begin{aligned}\mathbf{d}_{\mathrm{p}}\left((h\,f')\alpha, (h\,g')\alpha\right) &= \mathbf{d}_{\mathrm{p}}\left(\left[\left[(h)\overset{\circ}{\zeta}\right]\left[\left((h)\zeta^\sharp\right)_p\right] u\right], \left[\left[(h)\overset{\circ}{\zeta}\right]\left[\left((h)\zeta^\sharp\right)_q\right] v\right]\right) \\ &= |\left((h)\zeta^\sharp\right)_p| + |\left((h)\zeta^\sharp\right)_q| + |u| + |v| \ ,\end{aligned} \tag{4.13}$$

can be made arbitrarily large: this contradicts the assumption that α is uniformly bounded. ■

[35] N.B. the notation is a little confusing: $(f)\zeta$ and $(f)\zeta^\sharp$ are *vectors* (whose entries are words) and $(f)\overset{\circ}{\zeta}$ is a *word*, an element of B^*.

Exercises

• 4.1 Show, with an example, that a deterministic rational relation can be functional without being sequential. Verify that a sequential function is deterministic.

4.2 Let $L \subset B^*$ and $w = \bigwedge_{u \in L} u$. Show that $\forall u \in L\,,\ \exists v \in L \quad w = u \wedge v$.

• 4.3 Verify Property 4.1.

• 4.4 Verify Property 4.2.

• 4.5 Prove Lemma 4.1.

4.6 (a) Verify that the prefix distance is indeed a distance in the usual sense; that is, satisfies the properties of symmetry, positivity and the triangular inequality. *cf. § III.1.3.1, p. 386*

(b) Interpret the prefix distance on the Cayley graph of the free monoid.

(c) Let x be an element of $\mathrm{F}(A)$. Write $|x|$ for the length of the *reduced word* of $\widetilde{A}^*$ which represents x. Verify that for f and g in A^*, $\mathbf{d}_{\mathrm{p}}(f,g) = |f^{-1}g| = |g^{-1}f|$. *cf. § II.6.2.4, p. 308*

4.7 Let ω_B be the 'Lead or Delay' action of $B^* \times B^*$ on $H_B = (B^* \times 1_{B^*}) \cup (1_{B^*} \times B^*) \cup \{\mathbf{0}\}$. The non-zero elements of H_B have a natural length. *cf. Exer. II.2.7, p. 237, Exer. II.6.16, p. 323 and Lem. 1.1, p. 647*

(a) Verify that if $(1_{B^*}, 1_{B^*}) \cdot (f,g) \neq \mathbf{0}$, then $\left|\left[(1_{B^*}, 1_{B^*}) \cdot (f,g)\right]\right| = \mathbf{d}_{\mathrm{p}}(f,g)$.

(b) Verify that if $(1_{B^*}, 1_{B^*}) \cdot (f,g) = \mathbf{0}$, then, for all (u,v) in $B^* \times B^*$, it holds that $\mathbf{d}_{\mathrm{p}}(f\,u, g\,v) \geqslant |u| + |v|$.

• 4.8 **Uniform boundedness and Lipschitz condition.** (a) Let α be a function whose domain is prefix-closed. Show that α is uniformly bounded if and only if it is Lipschitz.

(b) Give an example of a uniformly bounded function which is not Lipschitz.

• 4.9 Correct the slight imprecision in Equation (4.13).

4.2 Deciding sequentiality

We shall reuse the method which enabled us to characterise the *functionality* of a transducer to characterise the transducers that realise a *sequential function*. As in Section 1, we write ω_B for the action of $B^* \times B^*$ on the set

$$H_B = (B^* \times 1_{B^*}) \cup (1_{B^*} \times B^*) \cup \{\mathbf{0}\}$$

called the 'Lead or Delay' action. Recall also that *a transducer is functional if and only if the product*[36] *of the* trim part $\mathcal{U}$ *of* $\mathcal{T} \times \mathcal{T}$ *with* ω_B *is a* valuation *of* $\mathcal{U}$ *(such that all the final states are valued by* $(1_{B^*}, 1_{B^*})$*).* *cf. Prop. 1.2, p. 648*

Proposition 4.3 *A functional transducer* $\mathcal{T}$ *realises a sequential function if and only if the product of the* accessible part $\mathcal{V}$ *of* $\mathcal{T} \times \mathcal{T}$ *with the 'Lead or Delay' action* ω_B*:*

(i) *is finite;*

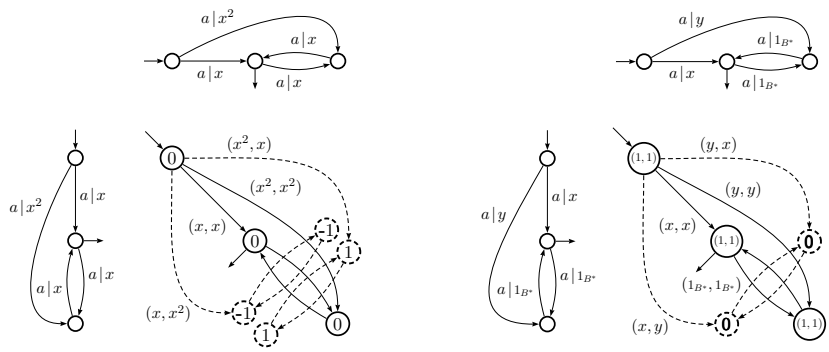
(ii) *is such that if a state whose value in* $\mathcal{V} \times \omega_B$ *is* $\mathbf{0}$ *belongs to a circuit of* $\mathcal{V}$*, all the transitions of this circuit are labelled with* $(1_{B^*}, 1_{B^*})$*.*

[36] The *product* of an automaton with an action is an *accessible* automaton (*cf.* Definition II.3.6, p. 267).

cf. Lem. I.1.16, p. 75 and Propty I.1.7, p. 78

Note the parallel between automata and transducers: we characterise unambiguous (respectively deterministic) automata by a condition on the trim part (respectively on the accessible part) of the cartesian square of the automaton and we characterise functional transducers (respectively which realise a sequential function) by a condition on the product of the trim part (respectively on the accessible part) of the cartesian square[37] of the transducer with ω_B.

Let us illustrate the construction used in Proposition 4.3. Figure 4.1 shows two cases in which the realised function is sequential: in (a), since the accessible part of the product is finite and no state is valued by $\mathbf{0}$; in (b), because the accessible part is also finite and the states valued by $\mathbf{0}$ belong to a cycles all of whose transitions are labelled $(1_{B^*}, 1_{B^*})$. We note, in example (a), that the condition in Proposition 4.3 *is not* that ω_B is a valuation of $\mathcal{V}$, contrary to the condition stated in Proposition 1.2 to characterise functionality.



(a) (b)

Figure 4.1: Two transducers which realise sequential functions

Figure 4.2 shows two cases in which the function realised is not sequential: in (a), because the accessible part of the product is infinite; in (b), because although the accessible part is finite, some states are valued by $\mathbf{0}$ and belong to cycles not all of whose transitions are labelled with $(1_{B^*}, 1_{B^*})$.[38]

With the proof of proposition Proposition 4.3 in mind, we first prove a lemma.

Lemma 4.4 *Let $(1_{B^*}, w)$ be in $H_B \setminus \mathbf{0}$ and (u, v) in $B^* \times B^* \setminus (1_{B^*}, 1_{B^*})$. The set $X = \{(1_{B^*}, w) \cdot (u, v)^n \mid n \in \mathbb{N}\}$ is finite and does not contain $\mathbf{0}$ if and only if u and v are conjugate by a word t; that is, $ut = tv$, and w is equal to $u^k t$ for some k. If this condition is satisfied, then X is in fact a singleton.*

[37] It was with precisely this parallel in mind that we stated Property I.1.7, in Chapter I, which might otherwise have seemed purposeless.

[38] *Note added in translation:* In both figures, 4.1 and 4.2, the conventions used for representing the values given to states are not the same in subfigures (a) as in subfigures (b). In (a), the output alphabet is $B = \{x\}$ and $H_{\{x\}}$ is identified with $\mathbb{Z}$ as in Figure 1.2, p. 649, whereas in (b) $B = \{x, y\}$ and H_B is as recalled above.

Proof. If the condition is satisfied, it follows that

$$(1_{B^*}, w) \cdot (u, v) = (1_{B^*}, u^{-1}(w\,v)) = (1_{B^*}, u^{-1}(u^k t\,v))$$
$$= (1_{B^*}, u^{k-1} t\,v) = (1_{B^*}, u^{k-1} u\,t) = (1_{B^*}, w) \quad ,$$

and X is reduced to the singleton $(1_{B^*}, w)$. Conversely, if X does not contain $\mathbf{0}$, then one of the following conditions is necessarily satisfied:

(i) either $u = 1_{B^*}$;

(ii) or $v = 1_{B^*}$ and w is a prefix of a power of u;

(iii) or w is a prefix of a power of u – that is, $w = u^k t$ where t is a prefix of u – and there exist two integers h and l such that u^k is the conjugate of v^l with t.

It is obvious that X is finite only if it is (iii) which is satisfied, with $h = l$. ■

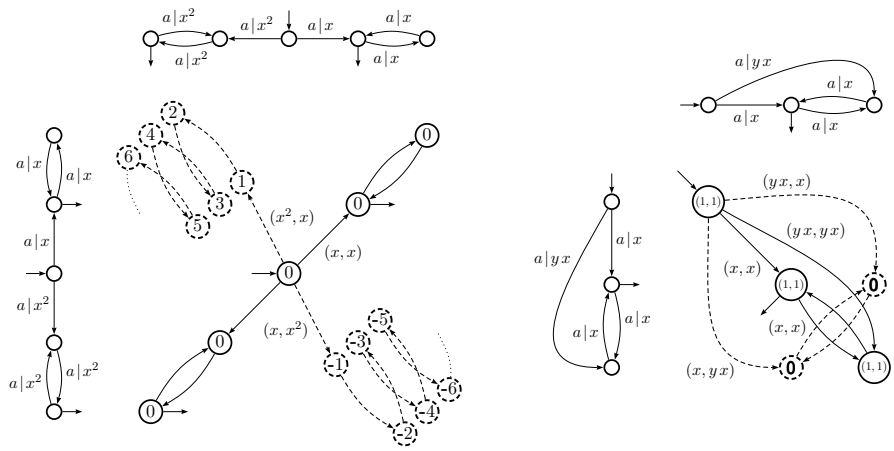


Figure 4.2: Two transducers which realise non-sequential functions

Proof of Proposition 4.3. By virtue of Theorem 4.2, it is enough to show that the conditions stated by the proposition are satisfied if and only if the function realised by $\mathcal{T}$ is uniformly bounded.

(i) The conditions are sufficient. As in Proposition 4.2, we write l for an upper bound on the lengths of the outputs of $\mathcal{T}$ and let m be an upper bound on the lengths of the valuations of the states in the produc $\mathcal{V} \times \omega_B$.

Let f and g be in $\mathsf{Dom}\,\alpha$; we write $h = f \wedge g$, $f = h\,f'$ and $g = h\,g'$. There exist in $\mathcal{T}$ two successful computations

$$i \xrightarrow{h|u} p \xrightarrow{f'|w} t \qquad \text{and} \qquad j \xrightarrow{h|v} q \xrightarrow{g'|z} s \quad ,$$

so that
$$(i, j) \xrightarrow{h|(u,v)} (p, q) \tag{4.14}$$

is a computation in $\mathcal{V}$. First case: $(1_{B^*}, 1_{B^*}) \cdot (u, v) \neq \mathbf{0}$, then

$$\mathbf{d}_{\mathrm{p}}\,(f\alpha, g\alpha) = \mathbf{d}_{\mathrm{p}}\,(u\,w, v\,z)$$
$$\leqslant m + |w| + |z| \leqslant m + l\,(|f'| + |g'|) \;=\; m + l\;\mathbf{d}_{\mathrm{p}}\,(f, g) \quad .$$

Second case: $(1_{B^*}, 1_{B^*}) \cdot (u, v) = \mathbf{0}$, and we factorise h into $h = h_1\, a\, h_2\, h_3$, with a in A, h_1, h_2 and h_3 in A^* (and possibly equal to 1_{A^*}), in such a way that the computation (4.14) factorises into

$$(i, j) \xrightarrow{h_1|(u_1,v_1)} (p_1, q_1) \xrightarrow{a|(x,y)} (p_2, q_2) \xrightarrow{h_2|(1_{B^*},1_{B^*})} (p_3, q_3) \xrightarrow{h_3|(u_2,v_2)} (p, q) \quad ,$$

where the valuation of (p_1, q_1) is different from $\mathbf{0}$, that of (p_2, q_2) is equal to $\mathbf{0}$ and (u_1, v_1) is different from $(1_{B^*}, 1_{B^*})$ if h_3 is different from 1_{A^*}. As all the states that follow (p_2, q_2) in the computation are valued by $\mathbf{0}$, the computation

$$(p_3, q_3) \xrightarrow{h_3|(u_2,v_2)} (p, q)$$

cannot contain a circuit and its length is bounded by $\|Q\|^2 = n^2$. It therefore follows that

$$\begin{aligned}\mathbf{d}_{\mathrm{p}}\,(f\alpha, g\alpha) &= \mathbf{d}_{\mathrm{p}}\,(u_1\, x\, u_3\, w, v_1\, y\, v_3\, z)\\ &\leqslant m + l\,(n^2+1) + l\,(|f'| + |g'|) = m + l\,(n^2+1) + l\ \mathbf{d}_{\mathrm{p}}\,(f, g) \quad .\end{aligned}$$

In both cases, α is a uniformly bounded (rational) function.

(ii) The conditions are necessary. First case: there exists in $\mathcal{V} \times \omega_B$ a circuit whose states are valued by $\mathbf{0}$ and whose label is not equal to $(1_{B^*}, 1_{B^*})$; we thus find in $\mathcal{V}$ a computation

$$(i, j) \xrightarrow{h_1|(u_1,v_1)} (p, q) \xrightarrow{h_2|(u_2,v_2)} (p, q) \quad ,$$

such that $(1, 1) \cdot (u_1, v_1) = \mathbf{0}$. Hence we deduce that

$$\mathbf{d}_{\mathrm{p}}\left((h_1\, h_2^n\, f')\alpha, (h_1\, h_2^n\, g')\alpha\right) = \mathbf{d}_{\mathrm{p}}\,(u_1\, u_2^n\, w, v_1\, v_2^n\, z) \geqslant n\,(|u_2| + |v_2|) + |w| + |z|$$

can be made arbitrarily large. Second case: the product $\mathcal{V} \times \omega_B$ is infinite. The result is that there exists in $\mathcal{V}$ at least one computation

$$(i, j) \xrightarrow{h_1|(u_1,v_1)} (p, q) \xrightarrow{h_2|(u_2,v_2)} (p, q)$$

which becomes an infinite graph in $\mathcal{V} \times \omega_B$. Hence

$$(1_{B^*}, 1_{B^*}) \cdot (u_1, v_1) = z \neq \mathbf{0} \qquad \text{and} \qquad \forall n \in \mathbb{N} \qquad z \cdot (u_2, v_2)^n \neq \mathbf{0} \quad .$$

From Lemma 4.4 it follows that first, $|u_2| \neq |v_2|$ and next, that there exists an n_0 such that

$$\left|z \cdot (u_2, v_2)^n\right| \geqslant (n - n_0)\left(\big||u_2| - |v_2|\big|\right) \quad ,$$

and hence

$$\begin{aligned}\mathbf{d}_{\mathrm{p}}\left((h_1\, h_2^n\, f')\alpha, (h_1\, h_2^n\, g')\alpha\right) &= \mathbf{d}_{\mathrm{p}}\,(u_1\, u_2^n\, w, v_1\, v_2^n\, z)\\ &\geqslant \left((n - n_0)\left(\big||u_2| - |v_2|\big|\right)\right) - \left(\big||w| - |z|\big|\right)\end{aligned}$$

can be made arbitrarily large. In both cases,

$$\mathbf{d}_{\mathrm{p}}\left(h_1\, h_2^n\, f', h_1\, h_2^n\, g'\right) \leqslant |f'| + |g'|$$

is fixed, and α is not uniformly bounded. ■

Not only can the conditions in Proposition 4.3 be effectively tested, but this can be done by an algorithm whose complexity is polynomial in the size of the transducer $\mathcal{T}$. More precisely:

Proposition 4.5 *Let $\mathcal{T}$ be a real-time transducer with n states, m transitions, and outputs whose length is bounded by ℓ. We can decide with a complexity less than $O(\ell^2 n^4 m^2)$ whether $\mathcal{T}$ realises a sequential function.*

The square $\mathcal{T} \times \mathcal{T}$ and its accessible part $\mathcal{V}$ can be computed in $O(m^2)$. The construction of the product $\mathcal{W}$ of $\mathcal{V}$ with ω_B is trickier: as the 'length' of the valuation of a state of $\mathcal{W}$ is bounded by a linear function of ℓ, a direct, but too coarse, estimate would give $\mathcal{W}$ an exponential size with respect to that of $\mathcal{T}$. We are nonetheless assured of a polynomial complexity by the fact that the non-null valuations – which are elements of H_B and hence almost elements of B^* – of states of $\mathcal{V}$ which are really used in the construction of $\mathcal{W}$ are prefixes of each other; there is therefore a linear number of such valuations and we can show that they are computable in polynomial time. The details of the proof, which in this case also contain its essence, are left as exercises. *cf. Exer. 4.12*

Remark 4.2 The fact the we can *decide* whether the function α realised by a transducer $\mathcal{T}$ is a sequential function, *with complexity polynomial* in the size of $\mathcal{T}$, does not mean, in the case where the answer is affirmative, that a sequential transducer which realises α is of *polynomial size* with respect to that of $\mathcal{T}$.

The simplest example of this distinction between *decision* and *construction* is given by the case of characteristic functions of languages. In this very particular case, the complexity of the decision is zero as the answer is always yes, but we know that the size of an equivalent sequential transducer (that is, of a deterministic automaton which recognises the domain of $|\mathcal{T}|$) can be exponential with respect to that of $\mathcal{T}$. □

Remark 4.3 The original proof[39] of Proposition 4.3 is via the definition of the *twinning* property of two states of a transducer. It is not hard to verify, using the definitions introduced here, that two states p and q of a transducer $\mathcal{T}$ are (non-trivially) twinned if:

(i) (p, q) is accessible in $\mathcal{T} \times \mathcal{T}$;

(ii) (p, q) belongs to a circuit in $\mathcal{V}$ whose label is a *pair of conjugate words* distinct from $(1_{B^*}, 1_{B^*})$;

(iii) (p, q) is not valued by $\mathbf{0}$ in the product of $\mathcal{V}$ with ω_B.

We then verify that a transducer realises a sequential function if and only if each pair of states is twinned and this circumstance is decidable. Thus, the conditions expressed by Proposition 4.3 and the twinning property are the same; it is only their formulation which differs and the technical details of the second are in a way hidden in Lemma 4.4. □

Exercises

4.10 Describe the functions realised by the transducers in Figure 4.1 and give the equivalent sequential transducers.

[39] By Ch. Choffrut in [55, Prop. 3.2]; *cf.* also [27, Prop. IV.6.4].

4.11 Describe the functions realised by the transducers in Figure 4.2.

• 4.12 **Decidability of sequentiality.** Let $\mathcal{T}$ be a real time transducer with n states, m transitions, and the lengths of the outputs be bounded by ℓ.

(a) Let $\mathcal{V}$ be the accessible part of $\mathcal{T}\times\mathcal{T}$ and $\mathcal{V}'$ the sub-automaton of $\mathcal{V}$ comprised of the states which are co-accessible to (at least) one circuit whose label is distinct from $(1_{B^*}, 1_{B^*})$. Evaluate the complexity of the computation of $\mathcal{V}'$.

(b) Let $\mathcal{W}$ be the product of $\mathcal{V}$ with ω_B and let $\mathcal{W}'$ be the product of $\mathcal{V}'$ with ω_B. Verify that the conditions of Proposition 4.3 hold for $\mathcal{W}$ if and only if they hold for $\mathcal{W}'$.

(c) Verify that if $((p,q),(1_{B^*},w))$ and $((p,q),(1_{B^*},w'))$ are two states of $\mathcal{W}'$ then w and w' are comparable (for the prefix order) or condition (ii) of Proposition 4.3 is not satisfied.

(d) Let $((p,q),(1_{B^*},w))$ be a state of $\mathcal{W}'$; show that if the conditions of Proposition 4.3 are satisfied then $|w| \leqslant \ell\, n^2$.

(e) Prove Proposition 4.5.

4.3 Minimisation

Just as the set of *quotients* of a language allows us to construct the minimal automaton of a language, the set of *translations* of a sequential function α defines the minimal transducer of α. We shall now prove this, in two steps, but first we give a construction on transducers (and in fact on automata) which will be used in the rest of this section.

4.3.1 Conjugacy

Let M be a monoid. The sequential function $|\mathcal{B}|$ realised by a sequential transducer $\mathcal{B} = \langle\, Q, A, M, \delta, \eta, I, T \,\rangle$ can be realised by the matrix representation (I, μ, T) where $\mu\colon A^* \to M^{Q\times Q}$ is the morphism defined by

$$\forall a \in A \qquad a\mu_{p,q} = \begin{cases} (p,a)\eta & \text{if} \quad q = (p,a)\delta \ , \\ 0 & \text{otherwise.} \end{cases}$$

It then follows, for all f in A^*, that $f|\mathcal{B}| = I \cdot f\mu \cdot T$.

If the monoid M is a group G, or a submonoid of a group G, we say that the transducer $\mathcal{B}' = \langle\, Q, A, G, \delta, \eta', I', T' \,\rangle$ is the *conjugate* of $\mathcal{B}$ by the matrix R, if R is a *diagonal* $Q\times Q$ matrix with entries in G and if the matrix representation (I', μ', T') of $\mathcal{B}'$ can be derived from that of $\mathcal{B}$ with the equations

$$\forall a \in A \qquad a\mu' = R^{-1} \cdot a\mu \cdot R \ , \qquad I' = I \cdot R \ , \qquad \text{and} \qquad T' = R^{-1} \cdot T \ .$$

It could happen, *a priori*, that the transitions of $\mathcal{B}'$ were no longer labelled with elements of M but with elements of G. Nevertheless, since $R \cdot R^{-1}$ is the identity matrix, we deduce immediately from the definition:[40]

[40] *Note added in translation:* Conjugacy, and not only by diagonal matrices, is a very efficient way to describe equivalence of tranducers, and of automata in general, and is so used in recent works; see bibliographic notes.

Property 4.5 *The conjugate of a transducer $\mathcal{B}$ is equivalent to $\mathcal{B}$.* ■

If we return to transducers in automaton form, we have, by writing $R_{p,p} = g_p$ for all p in Q,

$$\forall a \in A \qquad (p,a)\eta' = {g_p}^{-1}\,[(p,a)\eta]\,g_{p\cdot a} \ , \tag{4.15}$$

$$T'_p = {g_p}^{-1}\,T_q \ , \qquad \text{and} \qquad I'_i = I_i\,g_i \ , \quad I'_q = I_q = 0 \qquad \forall q \neq i \ . \tag{4.16}$$

This transformation is shown in Figure 4.3. Note also that two conjugate transducers have the same underlying input automaton.

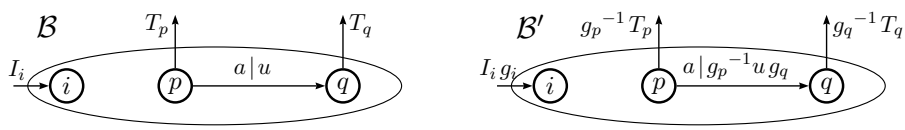


Figure 4.3: Conjugation of a transducer

This construction will be used later in the case where M is the free monoid B^*, viewed as a submonoid of the free group $\mathrm{F}(B)$. To do this, we adopt the implicit convention that the elements of the free group are represented by their reduced word in $\widetilde{B}^*$ and that multiplication is performed in $\mathrm{F}(B)$. This convention implies in particular that B^* is a submonoid of $\mathrm{F}(B)$; it is also consistent with the computation of the quotient: if f and g are words in B^*, then g^{-1} is in $\mathrm{F}(B)$ and the product $g^{-1}f$ in $\mathrm{F}(B)$ is in B^* if and only if the quotient $g^{-1}f$ is defined in B^*. If $g_p \in B^*$ for all p, conjugacy comes down to 'making' $\mathcal{B}$ output 'earlier', or, if $g_p \in \overline{B}^*$, 'later'. *cf. Exer. II.2.15, p. 246*

Remark 4.4 We can see, we probably *should have* seen, the length-preserving transducer resynchronisation algorithm as the conjugacy of a covering. □

Remark 4.5 Conjugacy can be defined just as well for *automata*: if $\mathcal{A} = \langle\, I, E, T \,\rangle$ is an automaton over M, submonoid of a group G and R is a diagonal matrix as above, then

$$\mathcal{A}' = \langle\, I \cdot R,\, R^{-1} \cdot E \cdot R,\, R^{-1} \cdot T \,\rangle$$

is an automaton over G, equivalent to $\mathcal{A}$. □

4.3.2 Blockage of a sequential transducer

As in the proof of Raney's Theorem, let $\mathcal{B} = \langle\, Q, A, B^*, \delta, \eta, I, T \,\rangle$ be a sequential transducer. For each q in Q, let $\mathcal{B}_q$ be the transducer derived from $\mathcal{B}$ by taking q as the initial state, with an initial function equal to 1_{B^*}. *Th. 4.1, p. 694*

We say that a state q is *blocked* if $|\mathcal{B}_q|$ is frank; that is, if $|\mathcal{B}_q| = |\mathcal{B}_q|/1_{A^*}$ or if the longest common prefix of all the outputs of $|\mathcal{B}_q|$ is 1_{B^*}. We shall say that a transducer $\mathcal{B}$ is *blocked* if all its states are.

Example 4.1 From (4.6) and (4.8), we deduce, by applying (4.4), that the 'translation transducer' $\mathcal{T}_\alpha$ of a function α is a blocked transducer. □ *cf. p. 694*

Proposition 4.6 *Every sequential transducer from A^* to B^* is the conjugate of a blocked sequential transducer, which we can compute effectively.*

Proof. Let $\mathcal{B} = \langle Q, A, B^*, \delta, \eta, I, T \rangle$ be a sequential transducer and, for each state q, let r_q be the longest common prefix of all the outputs of $\mathcal{B}_q$:

$$r_q = (1_{A^*})\overset{\circ}{|\mathcal{B}_q|} = \bigwedge \mathsf{Im}\, \mathcal{B}_q = \bigwedge_{g \in A^*} [q * g]\, T_{q \cdot g} \ .$$

cf. Exer. I.3.8, p. 111 and § 4.3.4, p. 708 below

This r_q is obviously effectively computable. Then let

$$\mathcal{B}' = \langle Q, A, \mathrm{F}(B), \delta, \eta', I', T' \rangle \ ,$$

be the sequential transducer conjugate of $\mathcal{B}$ by the matrix R, with $R_{q,q} = r_q$ for all q in Q, which is *a priori* a transducer *from* A^* *to* $\mathrm{F}(B)$. There follows

$$\forall p \in Q\,,\ \forall a \in A \qquad (p,a)\eta' = {r_p}^{-1}\,[(p,a)\eta]\, r_q \qquad \text{with } q = (p,a)\delta \ ,$$

$$\forall p \in Q \qquad T'_p = {r_p}^{-1} T_p \ , \qquad \text{and} \quad I'_i = I_i\, r_i \ , \quad I'_q = I_q = 0 \qquad \forall q \neq i \ .$$

By definition, r_p is a prefix of T_p and T'_p is in B^*. Since r_q is the longest common prefix of all the outputs of $\mathcal{B}_q$, $[(p,a)\eta]\, r_q$ is the longest common prefix of all the $(a\,g)|\mathcal{B}_p|$ and r_p is a prefix of $[(p,a)\eta]\, r_q$. Thus, $\mathcal{B}'$ is indeed a transducer from A^* to B^*.

Not only is $\mathcal{B}'$ equivalent to $\mathcal{B}$ but more precisely, for all q in Q, we have

$$\forall f \in A^* \qquad f|\mathcal{B}'_q| = {r_q}^{-1}\left[f|\mathcal{B}_q|\right] = \left[(1_{A^*})\overset{\circ}{|\mathcal{B}_q|}\right]^{-1}\left[f|\mathcal{B}_q|\right] = (f)|\mathcal{B}_q|/1_{A^*} \ .$$

Thus, $|\mathcal{B}'_q|$ is a frank function and q is blocked (in $\mathcal{B}'$): $\mathcal{B}'$ is a blocked transducer. ■

Recall that $\mathcal{B}$ and $\mathcal{B}'$ have the same underlying input automaton and hence the *same number of states.*

Example 4.2 The transducer $\mathcal{B}_2$ of Figure 4.4(a) realises the function α_2: $(a^{2n}b)\alpha_2 = (x\,y)^n x$. We easily compute (with the notation of Proposition 4.6):

$$r_p = y\,x \ , \qquad r_q = x \ , \qquad \text{and} \qquad r_r = 1_{B^*} \ ,$$

from which we obtain the blocked transducer $\mathcal{B}'_2$ of Figure 4.4(b). □

(a) transducer $\mathcal{B}_2$ (b) blocked transducer $\mathcal{B}'_2$

Figure 4.4: Blockage of $\mathcal{B}_2$

4.3.3 Reduction

We reuse the above notation. Two states p and q of the transducer $\mathcal{B}$ are *separate* if the functions $|\mathcal{B}_p|$ and $|\mathcal{B}_q|$ are not equal; a transducer is *reduced* if its states are pairwise separate.

Example 4.3 By Property 4.4, the 'translation transducer' $\mathcal{T}_\alpha$ of a function α is a reduced transducer. □ *cf. p. 695*

For the two following results, it will be sufficient to consider the transducers from A^* to B^* as automata[41] over $A^*\times B^*$. In particular, if $\mathcal{B}$ is a sequential transducer and $\psi\colon \mathcal{B} \to \mathcal{C}$ is a covering, then, for every state p of $\mathcal{B}$, we have that ψ is a covering of $\mathcal{B}_p$ on $(\mathcal{B}_p)\psi$, and, if $p\psi = q\psi$, that $(\mathcal{B}_p)\psi = (\mathcal{B}_q)\psi$, hence the property:

Property 4.6 *Let $\mathcal{B}$ be a sequential transducer and $\psi\colon \mathcal{B} \to \mathcal{C}$ a covering. Let p and q be two states of $\mathcal{B}$; if $p\psi = q\psi$, then $|\mathcal{B}_p| = |\mathcal{B}_q|$.* ■

Proposition 4.7 *Let α be a sequential function. The translation transducer $\mathcal{T}_\alpha = \langle R_\alpha, A, B^*, \delta_\alpha, \eta_\alpha, J, U \rangle$ is the minimal quotient of every accessible and blocked sequential transducer $\mathcal{B} = \langle Q, A, B^*, \delta, \eta, I, T \rangle$ which realises α.*

Proof. We define $\varphi\colon Q \to R_\alpha$ by

$$p\varphi = \alpha/f \qquad \text{with} \qquad i \cdot f = p \ . \tag{4.17}$$

Since $\mathcal{B}$ is accessible, there exists at least one f in A^* such that $i \cdot f = p$ for each p in Q. Furthermore, Property 4.4 states that

$$\alpha/f = |\mathcal{B}_{i\cdot f}|/1_{A^*} \ , \tag{4.18}$$

and hence (4.17) indeed defines a function from Q to R_α – that is, if f and f' are such that $i \cdot f = i \cdot f' = p$ – then $\alpha/f = |\mathcal{B}_{i\cdot f}|/1_{A^*} = \alpha/f'$. Since $\mathcal{B}$ is blocked, (4.18) becomes

$$\alpha/f = |\mathcal{B}_{i\cdot f}| \ , \tag{4.19}$$

and we must verify that, under this condition, φ is a covering. We first have

$$\forall p \in Q \qquad T_p = 1_{A^*}|\mathcal{B}_p| = (1_{A^*})\alpha/f = U_{\alpha/f} \qquad \text{and} \qquad I_i = (1_{A^*})\mathring{\alpha} = J_{\alpha/1_{A^*}} \ .$$

Then for all a in A and all p in Q, we have (with $i \cdot f = p$)

$$\forall g \in A^* \qquad (a\,g)|\mathcal{B}_p| = (p * a)\ \Big[g|\mathcal{B}_{p\cdot a}|\Big] = (a\,g)\alpha/f = (\alpha/f, a)\eta_\alpha\ \Big[g\alpha/f\,a\Big] \ .$$

Since $|\mathcal{B}_{p\cdot a}| = \alpha/f\,a$, it follows that $(p * a) = (p, a)\eta = (\alpha/f, a)\eta_\alpha$, which completes the verification.

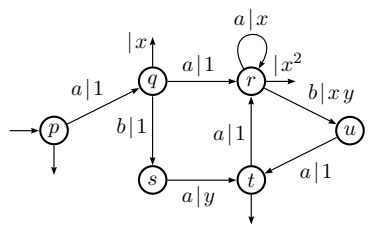
From Property 4.6, we deduce that $\mathcal{T}_\alpha$, since it is reduced, cannot have a proper quotient, and hence that it is the minimal quotient of $\mathcal{B}$. ■

Since every sequential transducer is equivalent to a blocked sequential transducer with the same number of states, we deduce directly from Proposition 4.7 the following characterisation of the translation transducer of a sequential function α, which will from now on be called instead the *minimal transducer* of α.

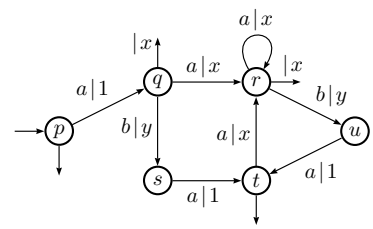
[41] We could have considered transducers from A^* to B^* as automata over A^* with weights in $\mathbb{K} = \mathfrak{P}(B^*)$ (or $\mathbb{K} = \mathbb{N}\langle\langle B^*\rangle\rangle$). Inasmuch as all the transducers considered are sequential (that is, the underlying input automata are deterministic) there is never *addition* of the coefficients of two transitions in a $\mathbb{K}$-covering: the real power of the notion (with respect to that of covering) is not used.

Theorem 4.3 *Let α be a sequential function. The translation transducer $\mathcal{T}_\alpha$ is the unique (up to isomorphism) reduced accessible transducer which realises α. It has the minimum number of states amongst all the sequential transducers which realise α.* ■

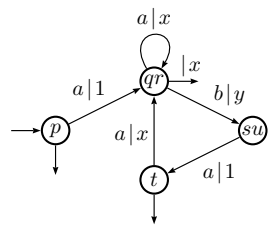
Example 4.4 Figure 4.5 shows the two steps of the minimisation of the transducer shown in (a): blockage in (b) and reduction in (c). □



(a) Transducer to minimise



(b) Blockage



(c) Reduction

Figure 4.5: Minimisation of a transducer

4.3.4 Effective computation

The proof above, which enables the computation of $\mathcal{T}_\alpha$ from an arbitrary sequential transducer $\mathcal{B}$ which realises α by blockage followed by reduction, is effective.

Reduction is realised by a process of successive refinement of partitions, called 'Moore's algorithm', whose complexity is quadratic in a naive implementation. The blockage of $\mathcal{B}$ assumes the computation, for each state p of $\mathcal{B}$, of the longest common prefix r_p of the image of $\mathcal{B}_p$. It would be a terrible idea to compute each r_p independently by the naive algorithm, which assumes a determinisation. The description of a method of blockage of $\mathcal{B}$ with a complexity proportional to the number of transitions of $\mathcal{B}$ (and to the maximum of the lengths of the r_p) is an interesting exercise which is worth elaborating upon.

cf. also § III.2.5.4, p. 421

cf. Exer. I.3.8, p. 111

The blockage of a transducer can be viewed as an algorithm which is applied to the underlying output automaton; that is, to a *generalised* (finite) automaton over a free monoid. So let $\mathcal{B} = \langle Q, B^*, I, E, T \rangle$ be such an automaton. The set E is such that for each pair of states (p, q) there is only a finite number of transitions (p, u, q)

cf. § I.2.4.2, p. 97

with u in B^*. The initial and final functions I and T also take values in B^*: for each state p, I_p is also considered as the label of an incoming transition to p; and T_p as that of a transition leaving p. In addition, the automaton $\mathcal{B}$ is assumed to be trim.

The general idea of the algorithm is to show that we can replace the computation of the r_p, which implies knowledge of the languages recognised by $\mathcal{B}_p$,[42] by a succession of computations of factors of r_p which can be done by examining the labels of the transitions that leave either from p, or from a set of states accessible from p by spontaneous transitions. To describe the algorithm more precisely requires a little terminology.

We write $\mathcal{B}_\varepsilon$ for the sub-automaton[43] of $\mathcal{B}$ formed uniquely from its spontaneous transitions; that is, from the transitions labelled 1_{A^*}. We compute the strongly connected components of $\mathcal{B}_\varepsilon$. The spontaneous transitions which belong to these strongly connected components are called *silent*. The transitions of $\mathcal{B}$, spontaneous or not, that are not silent are called *active transitions*. The strongly connected components of $\mathcal{B}$ can change in the course of the algorithm: they can *grow* and hence some spontaneous transitions can become silent. On the other hand, a silent transition can never become active.

Each step of the algorithm has two phases: the computation of a *valuation* of each state, and the *conjugacy* of these valuations.

Phase 1: valuation. For each state p, let $v'(p)$ be the longest common prefix of all the labels of active transitions that leave from p (which includes, according to our conventions, T_p). If p has no outgoing active transition (it must have at least one outgoing transition, hence silent, since $\mathcal{B}$ is trim), then $v'(p)$ is considered to be 'infinite' or 'indefinite' and is written $\top$. By convention, the longest common prefix of u and $\top$ is u, for all u in B^*.

Still for each state p, then let $v(p)$, called the 'valuation of p', be the longest common prefix of all the $v'(q)$, for all the q in the strongly connected component of p in $\mathcal{B}_\varepsilon$. Thus, $v(p) = v'(p)$ if p is alone in its strongly connected component (or if p is the only state of its strongly connected component whose v' is not infinite).

We write $s(p)$ for the minimal length of the words which are accepted by $\mathcal{B}$ when we take p as the unique initial state (and keep T as the final function).

If $v(p) = 1_{B^*}$, the state p is called 'locally[44] blocked'. The following property forms the heart of the algorithm (using $\mathcal{B}$ to denote both the transducer and its underlying output automaton):

Property 4.7 *A generalised automaton is blocked if and only if all its states are locally blocked.*

[42] We continue to write the automaton thus derived from $\mathcal{B}$, taking the initial function to be the characteristic vector of p.

[43] N.B. This notation is unrelated to the notation $\mathcal{B}_p$.

[44] N.B. This 'locally' is not quite local since the comparisons extend to the entire strongly connected component of p in $\mathcal{B}_\varepsilon$.

Proof. Let p be a state of $\mathcal{B}$ which is not locally blocked; that is, $v(p)$ is not 1_{B^*}. Every successful computation with source p factorises as

$$p \xrightarrow{1_{B^*}} p' \xrightarrow{v(p)\,u} q \xrightarrow{w} r \xrightarrow{T_r} \quad ,$$

where the computation $p \xrightarrow{1_{B^*}} p'$ is formed uniquely from *silent transitions* and $\big(p', v(p')\,u, q\big)$ is the *the first active transition* of the computation. By construction, $v(p) = v(p')$. If p' is a final state, $v(p)$ is also a prefix of $T_{p'}$. Thus, in any case, $v(p)$ is a common prefix of all the words in $|\mathcal{B}_p|$, and $\mathcal{B}$ is not blocked.

Conversely, suppose that $\mathcal{B}$ is not blocked (that is, that there exists a state p which is not blocked) and let r_p be the longest common prefix of all the words in $|\mathcal{B}_p|$. By assumption, r_p is not equal to 1_{B^*} and let x be its first letter.

Let P be the set of states which can be reached from p by a sequence of spontaneous transitions:

$$P = \{q \mid p \xrightarrow[\mathcal{B}]{1_{B^*}} q\} \quad .$$

For all q in P, $r_q = r_p$. There are two possible cases.

(a) There exists a q in P which is not the origin of any spontaneous transition. Then all the labels of transitions leaving q, including T_q if it is not 0, have x as their first letter, and $v(q)$ is a non-empty word whose first letter is x: q is not locally blocked.

(b) In the contrary case, let q be a state in P such that its strongly connected component in $\mathcal{B}_\varepsilon$ coincides with the states that can be reached from q by spontaneous transitions. Such a state exists since Q is finite.

A successful computation with source p which passes through q factorises into

$$p \xrightarrow{1_{B^*}} q \xrightarrow{1_{B^*}} q' \xrightarrow{v(p)\,u} r \xrightarrow{w} s \xrightarrow{T_s} \quad ,$$

or indeed into

$$p \xrightarrow{1_{B^*}} q \xrightarrow{1_{B^*}} q' \xrightarrow{T_{q'}} \quad .$$

All the *active* transitions which leave from an arbitrary state q' of the strongly connected component of q – there is at least one such q' since $\mathcal{B}$ is trim and none of its active transitions can be spontaneous by the choice of this component – have a label whose first letter is x and $v(q)$ is a non-empty word whose first letter is x: q is not locally blocked. ■

We note that a non-blocked state in $\mathcal{B}$ can be locally blocked; it is this that gives the proof above its subtlety and flavour.

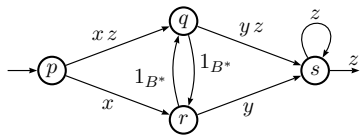
Phase 2: conjugacy. If not all the states of $\mathcal{B}$ are locally blocked, we take the conjugate of $\mathcal{B}$ by the matrix V, with $V_{p,p} = v(p)$ for all p in Q. By means of Equations (4.15) and (4.16), each transition (p, u, q) is replaced by $(p, [v(p)]^{-1}\, u\, [v(q)]\,, q)$ and, in a consistent manner, I_q is replaced by $I_q\,[v(q)]$ and T_p is replaced by $[v(p)]^{-1}\, T_p$.

Since all the states of a strongly connected component of $\mathcal{B}_\varepsilon$ have the same valuation, every silent transition remains silent after this transformation. On the other hand, some transitions can become spontaneous, which implies modifications of the strongly connected components of $\mathcal{B}_\varepsilon$.

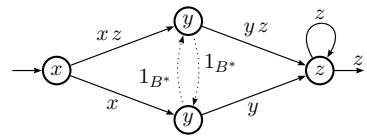
There is at least one state p which was not locally blocked before this second phase, and for which $s(p)$ strictly decreases afterwards. The quantity

$$S = \sum_{p \in Q} s(p)$$

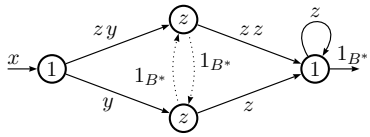
thus decreases at each step of the algorithm, which ensures termination. The series of automata of Figure 4.6 shows a three-step run of this algorithm.



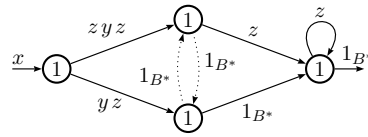
(a) Automaton to block



(b) After the first phase of the first step. The states bear their valuations.



(c) After the first phase of the second step



(d) After the first phase of the third step. The automaton is blocked.

Figure 4.6: Blockage of a transducer, effective algorithm

4.4 The (Great) Sequentiality Theorem

We shall now prove the characterisation of sequential functions announced in the introduction to this section.

Theorem 4.4 [**Choffrut**] *A (partial) function $\alpha\colon A^* \to B^*$ whose inverse preserves rationality (that is, for all S in $\operatorname{Rat} B^*$, $S\alpha^{-1}$ is in $\operatorname{Rat} A^*$) is a sequential function if and only if it is uniformly bounded.*

We already know that a sequential function is uniformly bounded (Theorem 4.2). It is thus the converse that is remarkable, since it states that in a certain context *rationality* is the consequence of a *topological condition*. We shall first prove the result for a restricted class of functions, those whose domain is prefix-closed; that is, if $(f\,g)\alpha$ is defined, $f\alpha$ is too.

4.4.1 Differential of a function

Definition 4.4 Let $\alpha\colon A^* \to B^*$ be a function whose domain is prefix-closed. The *differential of* α is the function $\check{\alpha}$ from A^* to $\mathrm{F}(B)$, the free group generated by B, defined by

$$(1_{A^*})\check{\alpha} = (1_{A^*})\alpha\,, \qquad \forall f \in A^*\,,\ \forall a \in A \qquad (f\,a)\check{\alpha} = [f\alpha]^{-1}\,[(f\,a)\alpha]\ ,$$

and where the left hand side is defined if and only if the right hand side is. □

In other words, the differential of α at $f\,a$ is the element of $\mathrm{F}(B)$ by which $f\alpha$ must be multiplied (on the right) to obtain $(f\,a)\alpha$. The following property flows directly from this interpretation and shows that we can reconstruct a function from its differential.

Property 4.8 *Let $\alpha\colon A^* \to B^*$ be a function whose domain is prefix-closed. For all f in A^*, if $f = a_1 a_2 \cdots a_n$ with the a_i in A, we have*

$$f\alpha = (1_{A^*})\check{\alpha}\,(a_1)\check{\alpha}\,(a_1 a_2)\check{\alpha} \cdots (a_1 a_2 \cdots a_n)\check{\alpha} \quad . \qquad \blacksquare$$

In the equation above we again use the implicit convention that the elements of the free group are represented by their reduced word in $\widetilde{B}^*$ and that multiplication is performed in $\mathrm{F}(B)$. This implies that we can relate an element t of $\mathrm{F}(B)$ to the length[45] of the reduced word that represents it, which we call the *length of* t. Since B is a finite alphabet, for each integer k, there exist only a finite number of elements of $\mathrm{F}(B)$ whose length is less than or equal to k.

Example 4.5 If $\iota\colon A^* \to A^*$ is the identity, then $(f\,a)\check{\iota} = a$.

Exam. 1.1, p. 645 If $\theta_1\colon a^* \to x^*$ is defined by $(a^{2p})\theta_1 = x^{4p}$ and $(a^{2p+1})\theta_1 = x^{2p+1}$, then $(a^{2p})\check{\theta}_1 = x^{2p+1}$ and $(a^{2p+1})\check{\theta}_1 = \overline{x}^{2p-1}$. $\square$

Def. 1.5, p. 660 Recall that we say that a function $\alpha\colon A^* \to B^*$ is *prefix-closed* if for all $f\,g$ in $\mathsf{Dom}\,\alpha$ $f\alpha$ is a prefix of $(f\,g)\alpha$. In particular, the domain of a prefix-closed function is prefix-closed. It then follows that:

Property 4.9 *A function $\alpha\colon A^* \to B^*$ is prefix-closed if and only if the differential $\check{\alpha}$ takes values in B^* (a subset of $\mathrm{F}(B)$).* $\blacksquare$

The differential of a function is naturally linked to the uniform boundedness of the function by the following property.

Property 4.10 *Let $\alpha\colon A^* \to B^*$ be a function whose domain is prefix-closed. The differential $\check{\alpha}$ has a finite image if and only if there exists an integer k such that*

$$\forall f \in A^*,\ \forall a \in A \qquad f\,a \in \mathsf{Dom}\,\alpha \quad \Longrightarrow \quad \mathbf{d}_{\mathrm{p}}\,(f\alpha, (f\,a)\alpha) \leqslant k \quad ,$$

and hence if and only if α is uniformly bounded. $\blacksquare$

We can now state the restricted form of the Sequentiality Theorem.

Theorem 4.5 *Let $\alpha\colon A^* \to B^*$ be a function whose domain is prefix-closed. The following conditions are equivalent.*

(i) *α is a sequential function;*

(ii) *the differential $\check{\alpha}$ is of finite image and the inverse of α preserves rationality;*

(iii) *the differential $\check{\alpha}$ is of finite image and for all t in $\mathrm{F}(B)$, $(t)\check{\alpha}^{-1}$ is a rational subset of A^*.*

[45] This is a length on $\mathrm{F}(B)$, but not a *gradation*, *cf.* Def. III.1.3, p. 383.

It is obvious that (i) implies (ii) by Property 4.10 since a sequential function is uniformly bounded (and its inverse preserves rationality). The proof of the two other non-trivial implications rests on two constructions, one somewhat mechanical, the other full of tricks, which we proceed to address below.

4.4.2 Proof of Theorem 4.5 (iii) $\Rightarrow$ (i)

The hypothesis allows us first to construct an automaton, independently of the 'nature' of the function $\check{\alpha}$.

Lemma 4.8 *Let φ be a function from A^* to a set E such that $\mathsf{Im}\,\varphi$ is finite and, for all e in E, $e\varphi^{-1}$ is a rational of A^*. There then exists a deterministic finite automaton over A^* with a terminal function that takes values in E, $\mathcal{A} = \langle Q, A, \delta, i, T \rangle$ such that, for all f in A^**

$$f\varphi = T_{i\cdot f} \quad .$$

Proof. For each e in $\mathsf{Im}\,\varphi$, let γ_e be the coarsest right regular equivalence on A^* which saturates $e\varphi^{-1}$:

$$\forall f, g \in A^* \qquad f \equiv g \quad [\gamma_e] \quad \Longleftrightarrow \quad \forall h \in A^* \quad (fh)\varphi = e \Leftrightarrow (gh)\varphi = e \quad .$$

As $e\varphi^{-1}$ is a rational of A^*, γ_e is of *finite index* (that is, has a finite number of classes). Let δ be the intersection of all the equivalences γ_e:

$$\delta = \bigcap_{e \in \mathsf{Im}\,\varphi} \gamma_e \quad .$$

It is right regular, and of finite index since $\mathsf{Im}\,\varphi$ is finite. It defines an action of A^* on the set of its classes and hence a finite automaton. We take $Q = A^*/\delta$ to be the set of classes modulo δ and δ induces an action of A^* on Q by

$$\forall f \in A^*\,,\ \forall a \in A \qquad [f]_\delta \cdot a = [fa]_\delta \quad .$$

The initial state is $i = [1_{A^*}]_\delta$ and we take as the final function

$$\forall f \in A^* \qquad T_{[f]_\delta} = f\varphi \quad .$$

Since δ is finer than the map equivalence of φ, the action and the final function are both defined by the equations above. It then follows, for all f in A^*, that

$$f|\mathcal{A}| = T_{i\cdot f} = f\varphi \quad . \qquad \blacksquare$$

In the case of the function $\check{\alpha}$, the automaton $\mathcal{A}$ above transforms, in two steps, into a transducer from A^* to B^* that realises α. We begin by constructing from $\mathcal{A}$ a *pure sequential transducer* from A^* to $\mathrm{F}(B)$ that realises α. Let $\mathcal{U} = \langle Q, A, \mathrm{F}(B), \delta, \eta, J, U \rangle$ be defined by

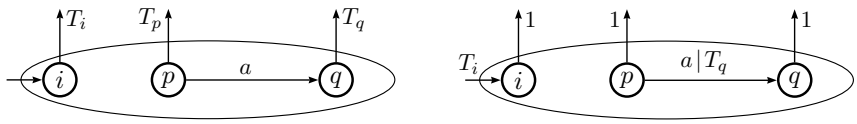
$$\forall p \in Q\,,\ \forall a \in A \qquad (p, a)\eta = p * a = T_{p\cdot a} \quad ,$$

$$J_i = T_i = (1_{A^*})\check{\alpha} = (1_{A^*})\alpha\,, \quad \forall q \neq i \quad J_q = 0\,, \quad \text{and} \quad \forall p \in Q \quad U_p = 1_{\mathrm{F}(B)} \quad ,$$

as shown in Figure 4.7. We immediately verify, by induction on the length of f – or,

which is the same thing, by the following sequence of equalities – that $\mathcal{U}$ realises α: for all f in A^*, if $f = a_1 a_2 \cdots a_n$ with the a_i in A, we have

$$\begin{aligned} f\alpha &= (1_{A^*})\check{\alpha}\,(a_1)\check{\alpha}\,(a_1 a_2)\check{\alpha}\cdots(a_1 a_2\cdots a_n)\check{\alpha} \\ &= J_i\, T_{i\cdot a_1}\, T_{i\cdot a_1 a_2}\cdots T_{i\cdot a_1 a_2\cdots a_n} \\ &= J_i\,(i * a_1)\,((i\cdot a_1) * a_2)\cdots((i\cdot a_1 a_2\cdots a_{n-1}) * a_n)\, U_{i\cdot a_1 a_2\cdots a_n} \\ &= f|\mathcal{U}| \ . \end{aligned}$$



(a) the automaton $\mathcal{A}$ (b) the transducer $\mathcal{U}$

Figure 4.7: Construction of a pure sequential transducer by conjugacy in $\mathrm{F}(B)$

Going from $\mathcal{U}$ to an equivalent sequential transducer from A^* to B^* is a sort of Fatou property which it is worth making into a proposition.

Lemma 4.9 *A sequential transducer from A^* to $\mathrm{F}(B)$ whose domain is prefix-closed and whose image is contained in B^* is conjugate to a sequential transducer from A^* to B^*.*

Proof. Let $\mathcal{U} = \langle Q, A, \mathrm{F}(B), \delta, \eta, J, U \rangle$ be a sequential transducer whose domain is prefix-closed. We first note that $\mathcal{U}$ is conjugate to a pure sequential transducer from A^* to $\mathrm{F}(B)$: by assumption, all the U_q are different from 0 and it is enough to conjugate $\mathcal{U}$ by the matrix R, with $R_{p,p} = {U_p}^{-1}$ for all p in Q to obtain a pure sequential transducer. We now suppose that $\mathcal{U}$ is pure sequential and that $|\mathcal{U}|$, which we write α for simplicity, takes values in B^*. The following computation rule in $\mathrm{F}(B)$ is easily verified:

cf. Lem. II.6.27, p. 320, for example

Property 4.11 *Let w be in B^* and t in $\mathrm{F}(B)$. If $w\,t$ is in B^*, then $t = \overline{u}\,v$ with u and v in B^* and u is a suffix of w.* ■

Since $\mathcal{U}$ is pure we have, for all f and g in A^* and all a in A,

$$(f\,g)|\mathcal{U}| = (f\,g)\alpha = f\alpha\,(p * g) \qquad \text{and} \qquad (f\,a)\alpha = f\alpha\,(p * a)\,, \qquad \text{with } p = i\cdot f\,.$$

Hence, by the property

$$p * a = \overline{u}_{p,a}\, v_{p,a} \qquad \text{and} \qquad p * g = \overline{u}_{p,g}\, v_{p,g} \ ,$$

with $u_{p,a}$, $v_{p,a}$, $u_{p,g}$ and $v_{p,g}$ in B^*. For each p in Q, we fix an f such that $p = i\cdot f$ and, for all the g in A^*, $u_{p,g}$ is a suffix of $f\alpha$. Let r_p be the longest of these suffixes, and let $\mathcal{T} = \langle Q, A, \mathrm{F}(B), \delta', \eta', I, T \rangle$ be the conjugate of $\mathcal{U}$ by the matrix R defined by

$$\forall p \in Q \qquad R_{p,p} = \overline{r}_p = (r_p)^{-1} \ .$$

It remains to show that $\mathcal{T}$ is a transducer from A^* to B^*; that is, for all p in Q and all a in A, we have

$$(p,a)\eta' = r_p\,(p,a)\eta\,\overline{r}_{p\cdot a} = r_p\,\overline{u}_{p,a}\,v_{p,a}\,\overline{r}_{p\cdot a} \in B^* \qquad \text{and also} \qquad I_i = J_i\,\overline{r}_i \in B^* \ .$$

By definition of $r_{p\cdot a}$, there exists a g such that

$$(p\cdot a, g)\eta = \overline{r}_{p\cdot a}\,v_{p\cdot a,g} \ ,$$

and hence we have

$$(p, a\,g)\eta = \overline{u}_{p,a}\,v_{p,a}\,\overline{r}_{p\cdot a}\,v_{p\cdot a,g} = \overline{u}_{p,a\,g}\,v_{p,a\,g} \ ,$$

from which we deduce

$$\overline{u}_{p,a}\,v_{p,a}\,\overline{r}_{p\cdot a} = \overline{u}_{p,a\,g}\,v' \qquad \text{for some } v' \text{ in } B^*.$$

Since, by definition, $u_{p,a\,g}$ is a suffix of r_p, $r_p\,\overline{u}_{p,a}\,v_{p,a}\,\overline{r}_{p\cdot a}$ belongs to B^*. In the same way, there exists a g such that $(i,g)\eta = \overline{r}_i\,v_{i,g}$. Since

$$g\alpha = J_i\,(i,g)\eta = J_i\,\overline{r}_i\,v_{i,g}$$

is in B^*, $J_i\,\overline{r}_i$ is in B^* as well and this last assertion completes the proof of the lemma and of part (iii) $\Rightarrow$ (i) of Theorem 4.5. ■

4.4.3 Proof of Theorem 4.5 (ii) $\Rightarrow$ (iii)

We start with a definition and a lemma, independent of Theorem 4.5.

Definition 4.5 Two functions β and γ from A^* to B^* are called *adjacent* if the (prefix) distance of the images of every word for which β and γ are defined is bounded; that is, if there exists an integer K such that

$$\forall f \in A^* \qquad f \in \mathsf{Dom}\,\beta \cap \mathsf{Dom}\,\gamma \quad \Longrightarrow \quad \mathbf{d}_{\mathrm{p}}\,(f\beta, f\gamma) \leqslant K \ . \tag{4.20}$$

Lemma 4.10 *Let β and γ be adjacent functions from A^* to B^* whose inverses preserve rationality. Then the set of words on which they are equal is rational:*

$$E(\beta,\gamma) = \{f \in A^* \mid f\beta = f\gamma\} \in \mathrm{Rat}\,A^* \ .$$

Proof. Note that if β and γ satisfy the hypotheses, then so do their restrictions to an arbitrary rational of A^*. First suppose that if $f\beta$ and $f\gamma$ are defined, *they have the same length.* By (4.20), it suffices, to prove $f\beta = f\gamma$, either for these two words to be equal and of length less than K, or for them to be of length greater than or equal to K and to have *the same suffix of length* K. Hence, in this case,

$$E(\beta,\gamma) = \bigcup_{|u|<K} \big((u)\beta^{-1} \cap (u)\gamma^{-1}\big) \cup \bigcup_{|u|=K} \big((B^*u)\beta^{-1} \cap (B^*u)\gamma^{-1}\big)$$

is rational by the hypothesis that β^{-1} and γ^{-1} preserve rationals.

The general case reduces to the previous case if we show that the language L of words f for which $f\beta$ and $f\gamma$ have the same length

$$L = \{f \in A^* \mid |f\beta| = |f\gamma|\} \ ,$$

is rational, and we replace β and γ by their restriction to L. Then, again by (4.20), the lengths of $f\beta$ and $f\gamma$ differ by at most K and hence they are equal if and only if they are congruent modulo $K+1$ (this is the 'trick' in the lemma); that is,

$$|f\beta| = |f\gamma| \quad \Longleftrightarrow \quad \exists i \in \{0, 1, \ldots, K\} \qquad |f\beta| \equiv |f\gamma| \equiv i \quad [K+1]$$

and hence $L = \bigcup_{0 \leqslant i \leqslant K} \left[(B^i(B^{K+1})^*)\beta^{-1} \cap (B^i(B^{K+1})^*)\gamma^{-1}\right]$ is rational. ■

We now come to the end of the proof of Theorem 4.5 itself. Let $t = \overline{u}\,v = \overline{f\alpha}\,(f\,a)\alpha$ be in the image of $\overset{\vee}{\alpha}$, where $\overline{u}\,v$ is the reduced word which represents t (u and v having no common proper prefix). As $\overset{\vee}{\alpha}$ is of finite image, there exists an integer k such that $|u| + |v| \leqslant k$.

We choose a letter a in A and let β and γ be the two functions from A^* to B^* defined by

$$f\beta = [f\alpha]\ u^{-1} \qquad \text{and} \qquad f\gamma = [(f\,a)\alpha]\ v^{-1} \ ,$$

where the two right quotients are taken in B^*, and not in $\mathrm{F}(B)$. Since the inverse of α preserves rationality, so do the inverses of β and γ:

$$\begin{aligned} & S\beta^{-1} = \{f \in A^* \mid [f\alpha]\ u^{-1} \in S\} = (S\,u)\alpha^{-1} \\ \text{and} \qquad & S\gamma^{-1} = \{f \in A^* \mid [(f\,a)\alpha]\ v^{-1} \in S\} = \left[(S\,v)\alpha^{-1}\right]\ a^{-1} \ . \end{aligned}$$

Furthermore, β and γ are adjacent since for all f in $\mathsf{Dom}\,\beta \cap \mathsf{Dom}\,\gamma$

$$\begin{aligned} \mathbf{d}_{\mathrm{p}}\,(f\beta, f\gamma) &= \mathbf{d}_{\mathrm{p}}\left([f\alpha]\ u^{-1}, [(f\,a)\alpha]\ v^{-1}\right) \\ &\leqslant \mathbf{d}_{\mathrm{p}}\left([f\alpha]\ u^{-1}, f\alpha\right) + \mathbf{d}_{\mathrm{p}}\,(f\alpha, (f\,a)\alpha) + \mathbf{d}_{\mathrm{p}}\left((f\,a)\alpha, [(f\,a)\alpha]\ v^{-1}\right) \\ &\leqslant 2\,k \ . \end{aligned}$$

Thus

$$(f\,a)\overset{\vee}{\alpha} = t \quad \Longleftrightarrow \quad f\beta = f\gamma \quad \Longleftrightarrow \quad f \in E(\beta, \gamma) \ ,$$

and hence $(t)\overset{\vee}{\alpha}^{-1}$ is rational by Lemma 4.10 and since A is finite, and this completes the proof of Theorem 4.5. ■

4.4.4 Return to the Sequentiality Theorem

We shall now turn the proof of Theorem 4.4 into a proof of Theorem 4.5 by constructing a function that satisfies the hypotheses of the latter.

To this end, let α be a uniformly bounded function whose inverse preserves rationality. In particular, $\mathsf{Dom}\,\alpha = (B^*)\alpha^{-1}$ is a rational of A^*, recognised by a trim deterministic automaton $\mathcal{A} = \langle\, Q, A, \delta, i, T\,\rangle$. For each state q of $\mathcal{A}$, let h_q be a word in A^*, arbitrary but fixed, such that $q \cdot h_q$ is a final state of $\mathcal{A}$, with the additional condition that $h_q = 1_{A^*}$ if q is final.

Let θ be the function from A^* to B^* defined by

$$\forall f \in A^* \qquad f\theta = (f\,h_q)\alpha \qquad \text{with} \qquad q = i \cdot f \ .$$

The condition that $h_q = 1_{A^*}$ if q is a final state of $\mathcal{A}$ implies that the function α is the restriction of θ to $D = \mathsf{Dom}\,\alpha$; that is,

$$\alpha = \iota_D \circ \theta \ . \tag{4.21}$$

Furthermore, $\mathsf{Dom}\,\theta = \mathsf{Pre}(|\mathcal{A}|)$ is prefix-closed.

Let K be an upper bound on the lengths of h_q. The assumption that α is uniformly bounded allows us to write

$$\forall n \in \mathbb{N}\,,\ \exists N(n) \in \mathbb{N} \quad \forall f, g \in \mathsf{Dom}\,\alpha \qquad \mathbf{d}_{\mathrm{p}}\,(f,g) \leqslant n \implies \mathbf{d}_{\mathrm{p}}\,(f\alpha, g\alpha) \leqslant N(n)\,.$$

We deduce that, if f' and g' belong to the domain of θ, we have

$$\mathbf{d}_{\mathrm{p}}\left(f', g'\right) \leqslant n \implies \mathbf{d}_{\mathrm{p}}\left(f'h_q, g'h_r\right) \leqslant n + 2\,K \implies \mathbf{d}_{\mathrm{p}}\left(f'\theta, g'\theta\right) \leqslant N(n + 2\,K) \ ,$$

and θ is uniformly bounded. For every language L in B^*, we have

$$\begin{aligned} L\theta^{-1} &= \{f \in A^* \mid f\theta \in L\} = \bigcup_{q \in Q} \{f \in A^* \mid i \cdot f = q \quad \text{and} \quad (f\,h_q)\alpha \in L\} \\ &= \bigcup_{q \in Q} \left[\{f \in A^* \mid i \cdot f = q\} \cap \left[L\alpha^{-1}\right] h_q^{-1}\right] \ , \end{aligned}$$

where the right quotient $\left[L\alpha^{-1}\right] h_q^{-1}$ is taken in B^*. Thus θ^{-1} preserves rationality if α^{-1} does.

By Theorem 4.5, θ is sequential and, by (4.21), α is too since the product of the composition of two sequential functions is a sequential function. This completes the proof of Theorem 4.4. ■

4.5 Pure sequential functions and local functions

We can now characterise pure sequential functions and local functions in the family of sequential functions.

Proposition 4.11 *A sequential function is pure if and only if it is prefix-closed.*

Proof. Let $\mathcal{T} = \langle Q, A, B^*, \delta, \eta, I, T \rangle$ be a sequential transducer which realises a prefix-closed function α. We can assume that $\mathcal{T}$ is trim: for all q in Q, there exists f such that $i \cdot f = q$. We can even suppose that for all q in Q, $T_q \neq 0$. In the other case ($T_q = 0$) we have, for all f such that $i \cdot f = q$, $f\,\alpha = 0$ and hence, for all g, that $(fg)\alpha = 0$: we can remove the state q (and all the transitions which touch it) without changing the function realised by $\mathcal{T}$.

For all f in A^* and all a in A, we have

$$f\theta = I_i\,(i * f)\,T_{i \cdot f} \quad \text{is a prefix of} \quad (fa)\theta = I_i\,(i * f)\,((i \cdot f) * a)\,T_{i \cdot fa} \ .$$

Hence, for all q in Q, T_q *is a prefix of* $(q * a)\,T_{q \cdot a}$.

If we conjugate $\mathcal{T}$ by the matrix R where $R_{q,q} = T_q$, we obtain the transducer $\mathcal{T}' = \langle Q, A, \mathrm{F}(B), \delta, \eta', I', T' \rangle$ with

$$I'_i = I_i T_i \ , \quad (q,a)\eta' = (T_q)^{-1} [(q,a)\eta] \, T_{q \cdot a} \qquad \text{and} \qquad T'_q = 1 \ ,$$

for all q in Q and all a in A, which is a pure sequential transducer from A^* to B^* (and not to $\mathrm{F}(B)$). Verifying that $\mathcal{T}'$ realises α is immediate by induction on the length of the words. ■

It is obvious that conversely a pure sequential function is prefix-closed.[46]

Theorem 4.6 [**Vaysse**] *A sequential function is local if and only if the underlying input automaton of its minimal transducer is local.*

Proof. The condition is sufficient by Corollary 1.10. Conversely, if $\alpha \colon A^* \to B^*$ is a local function of degree d, whatever f and g are in A^*, and if u is of length $d-1$, then it follows immediately that

$$\alpha/_f u = \alpha/_g u \ .$$

Hence the state of the minimal transducer of α that is the end of a computation of length $d-1$ is entirely determined by the label of the input word: the underlying input automaton of $\mathcal{T}_\alpha$ is local. ■

[46] *Note added in translation:* The function that maps a^n to a^n and $a^n b$ to a^{2n} for every integer n is rational and prefix-closed but not sequential.

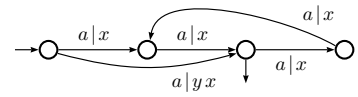
SOLUTIONS TO THE EXERCISES

1. Functionary

1.1 *Describe the function γ_1 from Example 1.2.*

Solution: $a^n\gamma_1 = x^{2n}$ for every integer n.

1.2 *Describe the relation γ_2 realised by the transducer $\mathcal{G}_2$ below.*



Solution: γ_2 is the *function* defined by

$$a^n\gamma_2 = \begin{cases} y\,x^n & \text{if} \quad n \equiv 1 \bmod 3\ ; \\ x^n & \text{if} \quad n \equiv 2 \bmod 3\ . \end{cases}$$

1.3 *Let $\theta\colon A^* \to B^*$ be a rational relation realised by a trim representation μ of dimension Q. Show directly (that is, without explicitly using the cartesian square of the transducer which corresponds to μ) that if θ is not a function, then there exists a word f in A^*, of length less than $1 + 2\,\|Q\|\,(\|Q\| - 1)$, such that the coefficients of $f\mu$ are not all monomial.*[47]

Conclude and comment.

Solution: Let $\mu\colon A^* \to B^{*Q\times Q}$ be a morphism; we set $N = 1 + 2\,\|Q\|\,(\|Q\| - 1)$. Let f be in A^* *of minimum length* such that there exists p and q in Q such that $\|f\mu_{p,q}\| \geqslant 2$. We suppose that $|f| > N$ and write $f = a_1a_2\cdots a_n$ with the a_i in A. These hypotheses imply that there exist, in the transducer $\mathcal{T}$ associated with μ, two *distinct* computations from $p = q_0$ to $q = q_n$:

$$\begin{aligned} & q_0 \xrightarrow{a_1|u'_1} q'_1 \xrightarrow{a_2|u'_2} \cdots \xrightarrow{a_{n-1}|u'_{n-1}} q'_{n-1} \xrightarrow{a_n|u'_n} q_n \\ \text{and} \qquad & q_0 \xrightarrow{a_1|u''_1} q''_1 \xrightarrow{a_2|u''_2} \cdots \xrightarrow{a_{n-1}|u''_{n-1}} q''_{n-1} \xrightarrow{a_n|u''_n} q_n\ , \end{aligned} \tag{SE.1.1}$$

where the u'_i and the u''_i are in B^* and

$$u'_1u'_2\cdots u'_n \neq u''_1u''_2\cdots u''_n\ . \tag{SE.1.2}$$

For all i in $[n]$, $q'_i \neq q''_i$; otherwise the hypotheses, in particular the minimality of the length of f, imply that

$$u'_1u'_2\cdots u'_i = u''_1u''_2\cdots u''_i \qquad \text{and} \qquad u'_{i+1}u'_{i+2}\cdots u'_n = u''_{i+1}u''_{i+2}\cdots u''_n\ ,$$

which contradicts inequality (SE.1.2). We thus have $n-1$ (greater than or equal to N) pairs (q'_i, q''_i) of different states; these pairs cannot all be distinct and there is even (at least) one pair of states which is repeated (at least) three times: there exists three distinct integers $1 \leqslant i < j < k \leqslant n-1$ such that $(q'_i, q''_i) = (q'_j, q''_j) = (q'_k, q''_k) = (r', r'')$. The computations described in (SE.1.1) factorise into

$$\begin{aligned} & q_0 \xrightarrow{g_1|v'_1} r' \xrightarrow{g_2|v'_2} r' \xrightarrow{g_3|v'_3} r' \xrightarrow{g_4|v'_4} q_n \\ & q_0 \xrightarrow{g_1|v''_1} r'' \xrightarrow{g_2|v''_2} r'' \xrightarrow{g_3|v''_3} r'' \xrightarrow{g_4|v''_4} q_n \end{aligned}$$

[47] Result and proof due to Schützenberger as part of the proof of Theorem 1.1 [224, 225].

from which we deduce that we we can observe in $\mathcal{T}$ the following three pairs of computations:

$$q_0 \xrightarrow{g_1|v'_1} r' \xrightarrow{g_4|v'_4} q_n \quad \text{and} \quad q_0 \xrightarrow{g_1|v''_1} r'' \xrightarrow{g_4|v''_4} q_n \ ,$$
$$q_0 \xrightarrow{g_1|v'_1} r' \xrightarrow{g_2|v'_2} r' \xrightarrow{g_4|v'_4} q_n \quad \text{and} \quad q_0 \xrightarrow{g_1|v''_1} r'' \xrightarrow{g_2|v''_2} r'' \xrightarrow{g_4|v''_4} q_n \ ,$$
$$q_0 \xrightarrow{g_1|v'_1} r' \xrightarrow{g_3|v'_3} r' \xrightarrow{g_4|v'_4} q_n \quad \text{and} \quad q_0 \xrightarrow{g_1|v''_1} r'' \xrightarrow{g_3|v''_3} r'' \xrightarrow{g_4|v''_4} q_n \ .$$

As the three words g_1g_4, $g_1g_2g_4$ and $g_1g_3g_4$ are strictly shorter than f, we have, by assumption, $\|(g_1g_4)\mu_{p,q}\| = \|(g_1g_2g_4)\mu_{p,q}\| = \|(g_1g_3g_4)\mu_{p,q}\| = 1$ and hence

$$v'_1\,v'_4 = v''_1\,v''_4 \ , \quad v'_1\,v'_2\,v'_4 = v''_1\,v''_2\,v''_4 \quad \text{and} \quad v'_1\,v'_3\,v'_4 = v''_1\,v''_3\,v''_4 \ . \tag{SE.1.3}$$

cf. Exer. 0.4.8, p. 27 and p. 42

We easily verify that (SE.1.3) implies that

$$u'_1\,u'_2\cdots u'_n = v'_1\,v'_2\,v'_3\,v'_4 = v''_1\,v''_2\,v''_3\,v''_4 = u''_1\,u''_2\cdots u''_n \ , \tag{SE.1.4}$$

which contradicts (SE.1.2). The result is that if there exists an f such that $\|f\mu_{p,q}\| \geqslant 2$, there exists one whose length is less than N.

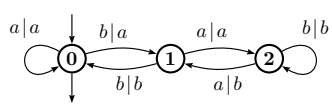
This result is in itself a proof of the decidability of the functionality of a real-time transducer. Though its proof is conceptually simpler than that of Theorem 1.1 given in the text, it corresponds to a decision procedure exponentially more complex. This result is part of the family of iteration lemmas, and the same remarks apply to it as to the rest of them.

cf. Rem. I.1.11, p. 72

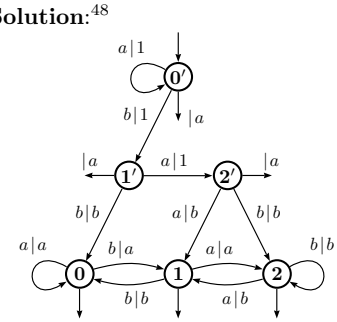
1.5 *Show that it is decidable whether a rational function is recognisable.*

Solution: A rational function is recognisable if and only if it is of finite image (Propty IV.1.10, p. 540). This last property is obviously decidable (Prop. I.1.11, p. 70).

1.7 ***Normalised division by*** 3. *Transform the transducer $\mathcal{Q}_2$ of Figure 1.3 (and shown below) so that it both computes the integer quotient by 3, even for numbers that are not multiples of 3, and writes the answer conventionally: no leading a (=0), unless the quotient is zero, in which case the result should be an a and not 1_{A^*}.*



Solution:[48]



Prop. 1.5, p. 656

1.9 *Prove that every recognisable function is sequential (and co-sequential).*

Solution: The graph of a recognisable function α can be written $\widehat{\alpha} = \bigcup_{j\in J} R_j \times w_j$ where J is finite and the rationals R_j are pairwise disjoint.

cf. for instance Lem. 4.8, p. 713

Let $\mathcal{A} = \langle Q, A, \delta, i, T\rangle$ be a deterministic automaton *which recognises all the* R_j; that is, T is a (partial) function from Q to J and for all f in A^*, $T_{i\cdot f} = j \Leftrightarrow f \in R_j$. If (λ, μ, ν) is the boolean representation of $\mathcal{A}$, then (λ, μ, ν'), with $\nu'_p = w_j$ when $T_p = j$, is a monomial representation of α.

1.10 *Let $\alpha\colon A^* \to B^*$ be a sequential function and u a word in B^*. Show that the function α' defined by $f\alpha' = u(f\alpha)$ is sequential.*

[48] *Cf.* [197].

Solution: If (λ, μ, ν) is a monomial representation of α, then (λ', μ, ν), with $\lambda' = u\,\lambda$ (that is, the unique non-zero component of λ' is the product of u by the unique non-zero component of λ) is a monomial representation of α'.

1.11 ***Codes with finite deciphering delay.*** (a) *Let* $\beta_1 \colon \{x, y\}^* \to A^*$ *be the morphism defined by* $x\beta_1 = a$ *and* $y\beta_1 = a\,b\,a$ *. Show that* β_1^{-1} *is a sequential function.*

(b) *Idem for* $\beta_2 \colon \{x, y, z\}^* \to A^*$ *defined by* $x\beta_2 = a\,b$ *,* $y\beta_2 = a\,b\,b$ *and* $z\beta_2 = b\,a\,a\,b$ *.*

(c) *Generalisation. Let* $\beta \colon B^* \to A^*$ *be a morphism. By definition,* $X = B\beta$ *is a* code *if* β *is injective. The code* X *is said to have* deciphering delay d *if, when a word* $f = x_1\,x_2 \cdots x_{d+1}$ *of* X^{d+1} *is a* prefix *of a word* $g = y_1\,y_2 \cdots y_r$ *of* X^**, then* $x_1 = y_1$ *. (In particular, a code is prefix if its deciphering delay is* 0*.)*

Prove that set $X = B\beta$ *is a code with finite deciphering delay if and only if* β^{-1} *is a sequential function (Proposition 1.6).*[50] *cf. p. 656*

Solution: (a) β_1^{-1} is realised by the transducer of Figure SE.1.1(a).

(b) β_2^{-1} is realised by the transducer of Figure SE.1.1(b).

(c) Suppose $X = B\beta$ has deciphering delay d. Let Q be the set of *proper* prefixes of X^{d+1}. The representation (λ, μ, ν) of size Q, defined by

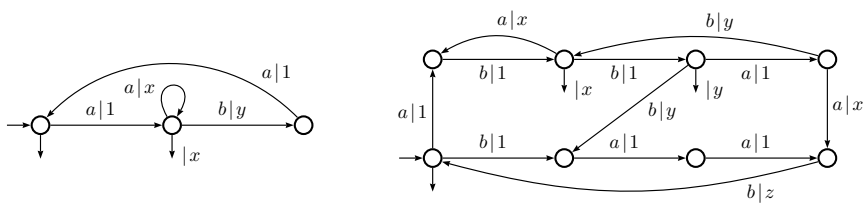
$$a\mu_{p,q} = \begin{cases} 1 & \text{if} \quad q = p\,a \in Q\,, \\ x_1 & \text{if} \quad p\,a \in X^{d+1}\,,\ p\,a = x_1\,x_2 \cdots x_{d+1} \ \text{and}\ q = x_2 \cdots x_{d+1}\,, \\ 0 & \text{otherwise;} \end{cases}$$

$$\lambda_q = \begin{cases} 1 & \text{if} \quad q = 1_{B^*}\,, \\ 0 & \text{otherwise;} \end{cases} \qquad \nu_p = \begin{cases} p & \text{if} \quad p \in X^{\leqslant d}\,, \\ 0 & \text{otherwise,} \end{cases}$$

is row monomial, and realises β^{-1}.

Conversely, suppose that $X = B\beta$ is a code and that β^{-1} is a sequential function, realised by $\mathcal{T} = \langle\, Q, A, B, \delta, \eta, i, T \,\rangle$. Let d be the maximum of the lengths of the T_q.

Let f be in X^{d+1}. The length of a word $u = i * f$ in B^* is at least 1 since the length of $f|\mathcal{T}| = (i * f)\,T_{i\cdot f}$ is $d+1$ and that of $T_{i\cdot f}$ is at most d. Thus u, and hence its first letter, is a prefix of the image of every word of which f is a prefix.

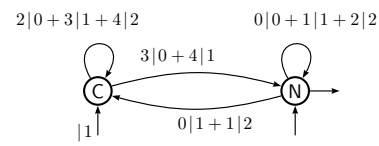


(a) decoder β_1^{-1} (b) decoder β_2^{-1}

Figure SE.1.1: Sequential decoders

1.14 *In imitation of Example 1.6, give a co-sequential (left) transducer for addition in base* 3*.*

Solution: The transducer in the adjacent figure.



[50] Result from [56].

1.15 ***Multiplication, base* 2.** *Starting from Example 1.3, show that multiplication by* 3 *of numbers written in base* 2 *is a co-sequential function.*

cf. p. 652, or solution of Exercise 1.7 above

Solution: If in the transducer $\mathcal{Q}_2$ of Figure 1.3, we *swap* the input and output – see Figure SE.1.2 (a) – we obtain a function which realises multiplication by 3 since the input is the quotient by 3 of the output, so the output is equal to the input multiplied by 3.

This transducer realises a co-sequential function, since its matrix representation turns out to be column monomial:

$$\lambda_3 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \ ,$$

$$a\mu_3 = \begin{pmatrix} a & b & 0 \\ 0 & 0 & a \\ 0 & 0 & 0 \end{pmatrix} , \qquad b\mu_3 = \begin{pmatrix} 0 & 0 & 0 \\ b & 0 & 0 \\ 0 & a & b \end{pmatrix} , \qquad \nu_3 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} .$$

Note that the domain of the function realised by this transducer is $[a + a\,(a\,b^+)^*b]^*$, not the whole of $\{a, b\}^*$: all the numbers are accepted as input (happily: every number is *multipliable* by 3), but not all their written forms. The transducer of Figure SE.1.2 (b) makes up for this deficiency.

(a) input and output swapped in the divider by 3 (b) likewise, but for all written forms

Figure SE.1.2: Multiplier by 3, in base 2

1.16 ***Multiplication, base* 3.** *Give a (co-sequential) transducer which realises multiplication by* 2 *for numbers written in base* 3. *Deduce a transducer which realises multiplication by* 4.

Solution: From the multiplication table in base 3,

$$\begin{array}{rr} 2 \times 2 = 11 & 0 \times 2 + 1 = 1 \\ 1 \times 2 + 1 = 10 & 2 \times 2 + 1 = 12 \ , \end{array}$$

we obtain a right sequential transducer which realises multiplication by 2 (Figure SE.1.3 (a)) whose transitions we reverse to obtain the left co-sequential transducer which multiplies numbers by 2 (Figure SE.1.3 (b)).

Multiplication by 4 can be obtained by two consecutive multiplications by 2; that is, by the composition of the function realised by the previous transducer with itself. The method for the composition of letter-by-letter transducers leads to the transducer of Figure SE.1.3 (c) (the conscientious reader will notice that the computation was already performed in Exercise IV.6.14).

cf. § IV.6.2.3, p. 607

1.17 ***Normalisation.*** *Let k again be a fixed base, and h another integer, greater than k. Logically, we write* $\mathbf{h} = \{0, 1, \ldots, h-1\}$. *We make every word f in $\mathbf{h}^*$ correspond to* the numerical value *of f in base k, again written $\overline{f}$, and defined by*

$$\overline{f} = \sum_{i=0}^{n} d_i\, k^i \qquad \text{if} \qquad f = d_n \cdots d_1 d_0 \quad \text{and} \quad d_i \in \mathbf{h} \ .$$

Normalisation in base k *is the function*

$$\nu_{h,k} \colon \mathbf{h}^* \quad \longrightarrow \quad \mathbf{k}^* \ ,$$

which preserves the numerical value; that is, $\overline{f\nu_{h,k}} = \overline{f}$, *for all f in $\mathbf{h}^*$.*

(a) *Show that* $\nu_{h,k}$ *is a co-sequential function.*

(b) *Evaluate (as a function of* k *and* h*) the minimum number of states of a co-sequential transducer which realises* $\nu_{h,k}$.

(c) *Explain why this solves the problem of addition as well as that of multiplication by a fixed integer, in every base* k.

Solution: (a) It is handy to construct a right sequential transducer *with an infinite number of states* which answers the question for all h, and then to verify that, for a given input alphabet $\mathbf{h}$, the accessible part of this transducer is finite. Hence let

$$\mathcal{N}_k = \langle\, \mathbb{N}, \mathbb{N}, \mathbf{k}, \delta, \eta, \{0\}, T \,\rangle \ ,$$

be the *right* transducer defined by the transitions

$$\forall p, n \in \mathbb{N} \qquad p \xrightarrow{n|d} q \qquad \text{with} \qquad p + n = q\,k + d \ , \quad 0 \leqslant d < k \ , \tag{SE.1.5}$$

and by the final function

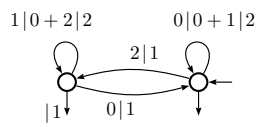
$$\forall p \in \mathbb{N} \qquad T_p = \langle\, p \,\rangle_k \qquad \text{where } \langle\, p \,\rangle_k \text{ is the base } k \text{ representation of the integer } p.$$

If we suppose that $n \leqslant h-1$, we obtain from (SE.1.5) that $p \leqslant h-1$ implies that $q \leqslant h-1$, since $k \geqslant 2$. Thus, when the inputs are smaller than h, the same is true of the number of accessible states and $\mathcal{N}_{h,k}$, the accessible part of the corresponding restriction of $\mathcal{N}_k$, is finite.

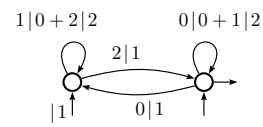
(b) More precisely it is shown, by induction on h, that the number of states of $\mathcal{N}_{h,k}$ is[51]

$$\left\lceil \frac{h-k}{k-1} \right\rceil + 1 \ .$$

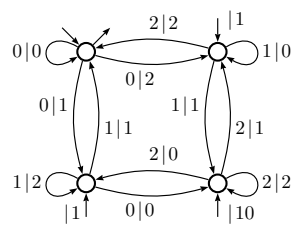
(c) For addition, we first realise digit-by-digit addition, which gives a number written in the alphabet $\mathbf{h}$ with $h \leqslant 2\,k-1$; for multiplication by n, we first multiply each digit by n, which gives a number written in alphabet $\mathbf{h}$ with $h \leqslant n\,(k-1)+1$.



(a) multiplication by 2, right transducer



(b) multiplication by 2, left transducer



(c) multiplication by 4, left transducer

Figure SE.1.3: Multiplications by 2 and 4 in base 3

[51] $\lceil x \rceil$ is the smallest integer greater than or equal to x.

1.18 ***Avizienis Addition.*** *We choose a base k and set* $\mathfrak{k} = \{\bar{h}, \ldots, \bar{1}, 0, 1, \ldots, h\}$ *, where* $h = \lfloor (k+1)/2 \rfloor$ *.* Normalisation *is the function*

$$\nu_k : \mathfrak{k}^* \quad \longrightarrow \quad \mathbf{k}^*$$

which preserves the numerical value.

(a) *Show that this normalisation is a co-sequential function. Give the corresponding right sequential transducers for $k = 2$ and $k = 3$.*

(b) *Show that it is possible to choose an inverse normalisation function – that is a 'translation'*[52] *from the standard system to the Avizienis system – that is sequential. Give the corresponding sequential transducers for $k = 2$ and $k = 3$.*

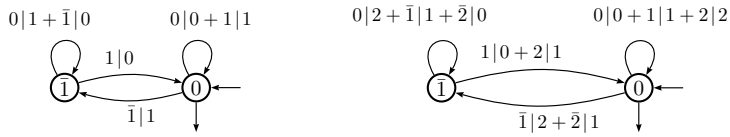
(c) *In the Avizienis system, addition can take the form of a function α_k from $\mathfrak{k}'^*$ to $\mathfrak{k}^*$ where*

$$\mathfrak{k}' = \{\overline{2h}, \ldots, \bar{1}, 0, 1, \ldots, 2h\} \ ,$$

(with h as above) and where α_k preserves the numerical value.

Show that α_k can be chosen in such a way that it is a sequential *function. Give the corresponding sequential transducers for $k = 2$ and $k = 3$.*

Solution: (a) *Cf.* Figure SE.1.4[53]



(a) in base 2 (b) in base 3

Figure SE.1.4: Avizienis normalisation

(b) and (c) The first case is included in the second. The construction is greatly simplified by the convention of taking transducers whose outputs systematically have one more digit (possibly a 0) that the input (*cf.* Figure SE.1.5).

1.19 *Show that multiplication (of numbers written in base k) is not a rational function from $\mathbf{k}^* \times \mathbf{k}^*$ to $\mathbf{k}^*$. Show that squaring is not a rational function from $\{0,1\}^*$ to $\{0,1\}^*$.*

Solution: We compute

$$1\,0^n\,1 \cdot 1^m = \begin{cases} 1\,0^{n+1}\,1^{m-n-2}\,0\,1^{n+1} & \text{if} \quad n \leqslant m-2\,, \\ 1^m\,0^{n-m+1}\,1^m & \text{if} \quad n \geqslant m-1\,, \end{cases}$$

if $k = 2$. We have an analogous result if $k \geqslant 3$ and the image of the recognisable subset $1\,0^*\,1 \times 1^*$ is not rational.[54] From the fact that in base 2 $(1^n)^2 = 1^{n-1}\,0^n\,1$, we also deduce that squaring is not a rational function.[55]

1.20 *Let (n, m) be a pair of integers written in base k. Suppose that the shorter number is padded on the left with zeros so that they are the same length. Verify that the function* $\mathsf{max}(n, m)$ *is a pure sequential function from* $(\mathbf{k} \times \mathbf{k})^*$ *to* $\mathbf{k}^*$.[56]

---

[52] *Note added in translation:* In French, 'traduction', which is a word with no already established mathematical meanings.

[53] We give the transducers for $k = 2$ and $k = 3$; for co-sequentiality in the general case, see for example [174].

[54] *Cf.* [27, Exer. III.5.3]; Berstel attributes the proof to Messerschmidt.

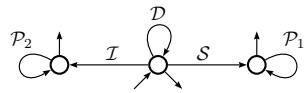
[55] *Cf.* [159, Prop. 7.1.6].

[56] Exercise borrowed from [80, XI,4].

Solution: We define the following subsets of $(\mathbf{k}\times\mathbf{k})^*\times\mathbf{k}^*$:

$$\mathcal{D} = \{(d,d) \mid d \mid d \in \mathbf{k}\} \ ,$$
$$\mathcal{P}_1 = \{(d,d') \mid d \mid d,d' \in \mathbf{k}\} \ , \quad \mathcal{P}_2 = \{(d,d') \mid d' \mid d,d' \in \mathbf{k}\} \ ,$$
$$\mathcal{S} = \{(d,d') \mid d \mid d,d' \in \mathbf{k}\,,\ d > d'\} \qquad \text{and} \qquad \mathcal{I} = \{(d,d') \mid d' \mid d,d' \in \mathbf{k}\,,\ d < d'\} \ .$$

The function max is realised by the following pure sequential transducer:

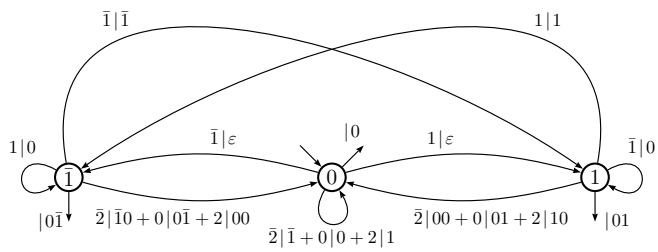


1.21 (a) *Show that the composition of a proper local function of degree d from A^* to B^* with a local function of degree d' from B^* to M is a local function of degree at most $d+d'$ from A^* to M. (Proposition 1.11).*

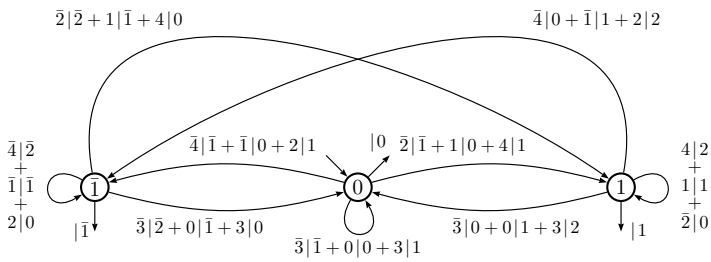
(b) *Show, with a simple example, that the assumption that the local function α is proper is necessary for the composition $\alpha\circ\beta$ with a local function β to be a local function.*

Solution: (a) If α is a proper local function of degree d, the image of a factor f of length $d+d'$ by α is at least of length d' and its image under β is well defined.

(b) Let $\alpha\colon \{a,b,c\}^* \to \{a,b\}^*$ be the projection which removes the letters c and $\beta\colon \{a,b\}^* \to \{a,b\}^*$ be the local function defined by $(x\,y)\overleftrightarrow{\beta} = a$ if $x = y$, and $(x\,y)\overleftrightarrow{\beta} = b$ otherwise. The function $\alpha\circ\beta$ is not local.



(a) in base 2



(b) in base 3

Figure SE.1.5: Avizienis adders

1.22 ***Soft addition.*** *We choose a base* $b \geqslant 3$ *and write the numbers using the alphabet of digits* $\mathfrak{b}'' = \{0, 1, \ldots, b-1, b, b+1\}$.

(a) *Show that addition is a local function of degree* 2 *from* $(\mathfrak{b}'' \times \mathfrak{b}'')^*$ *to* $\mathfrak{b}''^*$.

(b) *Give a local transducer for the case of base* 3.

(c) *Can we generalise the result to base* 2*?*

Solution: (a) *Cf.* [174] for example. (b) *Cf.* Figure SE.1.6.

(c) As for Avizienis's adder, there is no local transducer of degree 1 in base 2 (and over the alphabet $\{0, 1, 2, 3\}$), but there is one of degree 2.

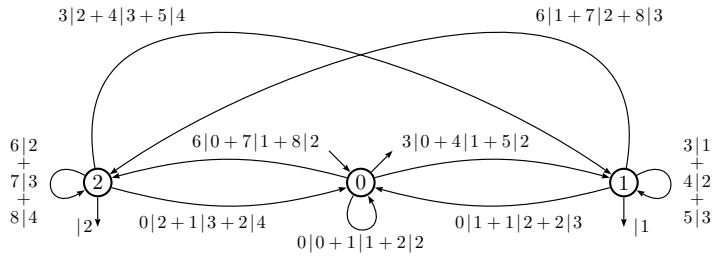


Figure SE.1.6: Soft addition in base 3, right transducer

2. Uniformisation of rational relations

2.1 *Show that every recognisable relation from* A^* *to* B^* *is uniformised by a recognisable function, and that every recognisable equivalence relation has a rational cross-section. Can we generalise these results to recognisable relations between arbitrary monoids?*

cf. Cor. II.2.20, p. 254 and also Sec. IV.1.4, p. 539

Solution: A recognisable relation θ (from M to N, arbitrary monoids) is of the form

$$\widehat{\theta} = \bigcup_{i \in I} R_i \times S_i \ ,$$

with I *finite*, $R_i \in \operatorname{Rec} M$ and $S_i \in \operatorname{Rec} N$. We choose a w_i in S_i for each i in I, and the relation τ, defined by

$$\widehat{\tau} = \bigcup_{i \in I} R_i \times w_i \ ,$$

is a function, the uniformisation of θ. It is a recognisable function *if the singletons are recognisable* in N, which is the case if $N = B^*$; it is a *rational* function if $\operatorname{Rec} M \subseteq \operatorname{Rat} M$, which is the case if $M = A^*$.

If θ is an equivalence relation (and then $B^* = A^*$), the uniformisation τ constructed above is constant on each of the classes: its image is finite, hence recognisable in A^*.

2.2 (a) *Describe the construction of the matrix representation of the Schützenberger co-covering of a transducer, then of that of an S-co-immersion.*

cf. p. 666 (b) *Apply the above construction to the transducer* $\mathcal{E}_2$ *of Example 2.2.*

cf. p. 671 *Solution*: (a) The construction is exactly that of Proposition 2.3 except that rows play the role of columns, and the co-determinisation plays the role of the determinisation.

Let $\mathcal{C}$ be a transducer whose representation is (λ, μ, ν) (of dimension Q), $\mathcal{A}$ its underlying input automaton and (η', κ', ξ') the (boolean) representation, of dimension $\mathfrak{P}(Q)$, of $\mathcal{A}_{\mathsf{cod}}$:

$$\begin{array}{rrcl} \forall P, S \in \mathfrak{P}(Q)\,,\ \forall a \in A & a\kappa'_{P,S} = 1 & \iff & P = \{p \mid \exists q \in S \quad a\mu_{p,q} \neq 0\}\ , \\ & \eta'_S = 1 & \iff & \exists q \in S \quad \lambda_q \neq 0 \\ \text{and} & \xi'_P = 1 & \iff & P = \{p \mid \nu_p \neq 0\}\ . \end{array}$$

The representation $(\zeta', \sigma', \omega')$ of the Schützenberger co-covering of $\mathcal{C}$ is defined by

$$\forall P, S \in \mathfrak{P}(Q)\,,\ \forall a \in A \qquad a\sigma'_{(P,Q),(S,Q)} = \begin{cases} a\mu^{|S|} & \text{if} \quad a\kappa'_{P,S} = 1 \\ 0 & \text{otherwise} \end{cases}$$

and by

$$\forall P, S \in \mathfrak{P}(Q) \qquad \zeta'_{(S,Q)} = \lambda^{|S|} \qquad \text{and} \qquad \omega'_{(P,Q)} = \begin{cases} \nu & \text{if} \quad \xi'_P = 1 \\ 0 & \text{otherwise} \end{cases}$$

The construction of a representation (χ', τ', ψ') of an S-co-immersion in $\mathcal{C}$ is analogous to that described for the representation of an S-immersion. For each letter a of A, each non-null block $a\sigma'_{(P,Q),(S,Q)}$ in $a\sigma'$ is replaced by a *row-monomial* block which has the same non-null rows as the original block. In other words, every non-null non-monomial row of each $Q{\times}Q$-block of $a\sigma'$ is made monomial, *but non-null*, by replacing each non-null entry (a subset of B^*) *either by* 0 *(the empty set), or by an arbitrarily chosen word in the entry.* The same operation is applied to the $1{\times}Q$-blocks of ζ'. *cf. p. 671*

(b) Recall the representation of $\mathcal{E}_2$:

$$\lambda_1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}\ , \quad a\mu_1 = \begin{pmatrix} a & b^+a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{pmatrix}\ , \quad b\mu_1 = \begin{pmatrix} b & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & b \end{pmatrix}\ , \quad \nu_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}\ .$$

The representation of $\mathcal{A}_{1\mathsf{cod}}$ is

$$\eta'_1 = \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix}\ , \quad a\kappa'_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\ , \quad b\kappa'_1 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}\ , \quad \xi'_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}\ .$$

The representation $(\zeta'_1, \sigma'_1, \omega'_1)$ of the S-co-covering of $\mathcal{E}_2$ is then (*cf.* Figure SE.2.1)

$$\zeta'_1 = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}\ , \quad \omega'_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}\ ,$$

$$a\sigma'_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \begin{matrix} a & b^+a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{matrix} & \begin{matrix} a & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{matrix} & \begin{matrix} 0 & b^+a & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{matrix} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & a \end{matrix} \end{pmatrix}\ , \quad b\sigma'_1 = \begin{pmatrix} \begin{matrix} b & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & b \end{matrix} & \begin{matrix} b & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & b \end{matrix} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & b \end{matrix} & \begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & b \end{matrix} \\ 0 & 0 & 0 & 0 \end{pmatrix}\ .$$

The representation $(\chi_1', \tau_1', \psi_1')$ of the S-uniformisation shown in Figure SE.2.1 is then

$$a\tau_1' = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \begin{matrix} \boxed{0}\,\boxed{ba}\,0 \\ 0\;0\;0 \\ 0\;0\;a \end{matrix} & \begin{matrix} a\;0\;0 \\ 0\;0\;0 \\ 0\;0\;a \end{matrix} & \begin{matrix} 0\,\boxed{b^2a}\,0 \\ 0\;0\;0 \\ 0\;0\;a \end{matrix} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \begin{matrix} 0\;0\;0 \\ 0\;0\;0 \\ 0\;0\;a \end{matrix} \end{pmatrix}$$

and $b\tau_1' = b\sigma_1'$, $\chi_1' = \zeta_1'$, $\psi_1' = \omega_1'$.

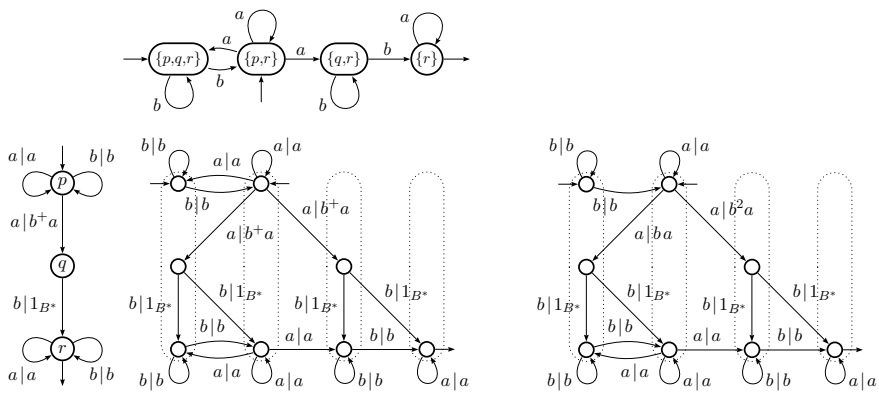
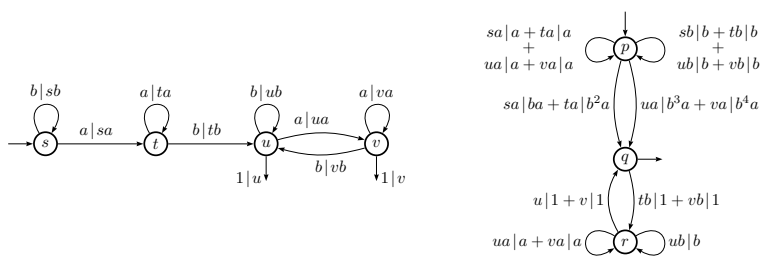


Figure SE.2.1: The transducer $\mathcal{E}_2$, its S-co-covering, and an S-co-uniformisation of θ_2

2.3 *Decompose the S-uniformisation of the function θ_2 defined by the transducer in Figure 2.2 (p. 667) into the product of a sequential transducer by a co-sequential transducer.*

Solution:



3. Cross-section of rational functions

3.1 ***The return of the Rational Cross-Section Theorem.*** *The aim of this exercise is to prove*[57] *the Rational Cross-Section Theorem directly:* every morphism from A^* to B^* has the rational cross-section property. *Let* $A = \{a_1, a_2, \ldots, a_n\}$ *be a totally ordered alphabet:* $a_1 < a_2 < \cdots < a_n$.

Th. 3.1, p. 681

[57] This is a guided reading of [80, § IX.7].

(a) *An elementary projection is the projection of a free monoid A^* on to one of the free monoids $(A \setminus a)^*$ (where a is one of the letters of A). Show that every morphism $\theta\colon A^* \to B^*$ can be decomposed into the product of an injective morphism by elementary projections.*

(b) *Let π be the (elementary) projection which removes the greatest letter of A. Let $\tau\colon A^* \to A^*$ be the relation which makes a word f of A^* correspond to all the words of A^* strictly greater than f in the lexicographic order and which have the same image as f under π. Show that τ is a rational relation.*

(c) *Let R be a subset of A^*. Show that $R \setminus R\tau$ is a cross-section of R for π.*

(d) *Show that every elementary projection has the rational cross-section property. Show that every morphism $\theta\colon A^* \to B^*$ has the rational cross-section property.*

Solution: (a) We can assume without loss of generality that A and B are disjoint. We set $Z = A \cup B$; then for every morphism $\theta\colon A^* \to B^*$, the morphism $\sigma\colon A^* \to Z^*$ defined by $a\sigma = a(a\theta)$ for all a in A is injective. We have $\theta = \sigma\,\mu$ where μ is the projection of Z^* onto B^* and every projection is a composition of elementary projections. *cf. Exer. 0.4.4, p. 26*

(b) Let z be the greatest letter of A. Two words f and g in A^* have the same image under π if they can be written as

$$f = z^{k_0}\, a_1\, z^{k_1}\, a_2 \cdots a_n\, z^{k_n} \quad , \qquad g = z^{l_0}\, a_1\, z^{l_1}\, a_2 \cdots a_n\, z^{l_n} \qquad \text{with } a_i \in A \text{ and } a_i \neq z\,.$$

Since z is the largest letter of A, $f \prec g$ if for the smallest index i such that $k_i \neq l_i$ we have $k_i < l_i$. We deduce immediately a rational expression for the graph of τ:

$$\widehat{\tau} = \mathcal{D}^*\,(1, z^+)\,[\mathcal{D}\,(z^*, z^*)]^* \quad , \qquad \text{with } \mathcal{D} = \sum_{a \in A} (a, a)\,.$$

Remark: The hypothesis that π removes the greatest letter is not necessary for τ to be rational (it is for the next question) but makes the solution simpler. Here is the solution in the general case.

We write z for the letter removed by π. Let B and C be the subsets of A formed respectively from letters less than and greater than z. We set

$$\mathcal{B} = \sum_{a \in B} (a, a) \quad , \quad \mathcal{C} = \sum_{a \in C} (a, a) \quad .$$

We now have, with the preceding notation, that $f \prec g$ if for the smallest index i such that $k_i \neq l_i$ either $k_i < l_i$ and $a_{i+1} \in B$, or $k_i > l_i$ and $a_{i+1} \in C$. The graph of τ can then be written

$$\widehat{\tau} = \mathcal{D}^*\,[(1, z^+)\,\mathcal{B} + (z^+, 1)\,\mathcal{C}]\,[\mathcal{D}\,(z^*, z^*)]^* \quad .$$

(c) Let f and g be in R such that $f\pi = g\pi$; if $f \prec g$ then $g \in f\tau$ and $g \notin R \setminus R\tau$. Thus $T = R \setminus R\tau$ is the set of smallest elements in the lexicographic order, *when they exist*, in the intersection of R with each class of the map equivalence of π.

It is a property of π that this smallest element always exists; let $s = \{f_i\}$ be a decreasing sequence of elements of R which all have the same image under π:

$$f_i = z^{k_{i,0}}\, a_1\, z^{k_{i,1}}\, a_2 \cdots a_n\, z^{k_{i,n}} \quad .$$

From what we saw in the previous question, we deduce that the sequence of integers $k_{i,0}$ is decreasing, and is hence stationary beyond some point. Beyond this point it is the sequence of the $k_{i,1}$ which is decreasing and hence stationary beyond some point. We thus show successively that all the sequences $k_{i,j}$ are stationary and hence that the sequence s is finite: it contains a smallest element.

Remark: If the letter removed by π is not the greatest of the alphabet A, this smallest element might not always exist. Take the example of a set which does not contain a smallest

element for the lexicographic order $a \prec b$: $R = a^*b$. If we take $a\pi = 1$ and $b\pi = b$, R is completely contained in a single class modulo π and $R \setminus R\tau = \emptyset$.

(d) Let $\pi\colon A^* \to A^*$ be an elementary projection. We can order A in such a way that the letter removed by π is the greatest letter. From the previous question, and with its notation, we deduce that if $R \in \operatorname{Rat} A^*$ then $T = R \setminus R\tau \in \operatorname{Rat} A^*$ since τ is a rational relation, π is injective on T, since T contains a single element in each class modulo π, and $T\pi = R\pi$, since T contains a single element of each class.

As every injective morphism obviously has the rational cross-section property, it follows from (a), and from Lemma 3.3 that every morphism $\theta\colon A^* \to B^*$ has the rational cross-section property.

cf. p. 681

3.2 *Show that if M has the rational cross-section property:*

cf. p. 682

(a) *every rational function from a free monoid to M has this property (Property 3.1);*

(b) $\operatorname{Rat} M = \operatorname{URat} M$ *(Property 3.2).*

Solution: (a) Let $\theta\colon A^* \to M$ be a rational function. It is enough to find a rational cross-section of $\mathsf{Dom}\,\theta$ (for a rational cross-section of R, we replace θ by $\iota_R\theta$).

Let $\theta = \varphi^{-1} \circ \iota_K \circ \psi$ be a realisation by morphisms of θ, U a rational cross-section of $A^*\varphi^{-1} \cap K$ for ψ and $T = U\varphi$. It holds that $\mathsf{Im}\,\theta = (A^*\varphi^{-1} \cap K)\psi = U\psi$, and since $T\varphi^{-1} \cap K$ contains $T\varphi^{-1} \cap U = U$, we have $T\theta = \mathsf{Im}\,\theta$.

$$\begin{tikzcd} Z^* \arrow[r, "\iota_K"] \arrow[d, "\varphi"'] & Z^* \arrow[d, "\psi"] \\ A^* \arrow[r, "\theta"'] & M \end{tikzcd}$$

Furthermore, since θ is a *function*, $f\theta \cap g\theta \neq \emptyset$ implies that $f\theta = g\theta = m$, which implies that $f\varphi^{-1} \cap K \subseteq m\psi^{-1}$ and that $g\varphi^{-1} \cap K \subseteq m\psi^{-1}$ and hence there exists u in U such that $f\varphi^{-1} \cap U = g\varphi^{-1} \cap U = u$, from which $f = g = u\varphi$ if f and g are in $U\varphi$ and θ is injective on T.

(b) Let $\alpha\colon A^* \to M$ be a surjective morphism. If $S \in \operatorname{Rat} M$, there exists $R \in \operatorname{Rat} A^*$ such that $R\alpha = S$. Let T be a rational cross-section of R for α. Like all the rationals of A^*, T is unambiguous, and since α is injective on T, $S = T\alpha$ is an unambiguous rational of M.

cf. Prop. II.1.12 (i), p. 231

3.3 *Show that $\mathbb{N}^2$ does not have the rational cross-section property.*

Solution: Let $\gamma\colon \{a,b\}^* \to \{a,b\}^{\oplus} = \mathbb{N}^2$ be the canonical morphism, and let us choose the rational set $R = a^*(ab^2)^* \cup (a^2b)^*a^*$. It follows that

$$R\gamma = \left(a^{\oplus} + (a+2b)^{\oplus}\right) \cup \left((2a+b)^{\oplus} + b^{\oplus}\right) \quad .$$

Suppose that T is the rational cross-section of R for γ and let N be the integer associated with T by the Star Lemma.

The word $u = a^{5\,N!}(ab^2)^{N!}$ belongs to T; indeed, there is no other word in R which has the same commutative image: all the words of the component $a^*(ab^2)^*$ have distinct images and the words of the component $(a^2b)^*a^*$ have a commutative image $na + mb$ such that $2m \geqslant n$ and $u\gamma = 5\,N!\,a + 2\,N!\,b$.

By the (Block) Star Lemma, $u = fgh$ with $f = a^{5\,N!}(ab^2)^r$, $g = (ab^2)^s$, $h = (ab^2)^t$, $s \leqslant N$ and fg^nh in T for all n. We can choose n such that

cf. Lem. I.1.14, p. 71

$$w = fg^nh = a^{5\,N!}(ab^2)^{5\,N!}$$

is in T. Symmetrically, the word $v = (ab^2)^{N!}a^{5\,N!}$ belongs to T and we deduce again by the Block Star Lemma that

$$z = (ab^2)^{5\,N!}a^{5\,N!}$$

is in T as well. However, $w\gamma = z\gamma$ and thus T is not a cross-section.

3.4 (a) *Show that every morphism $\alpha\colon M \to N$ between commutative monoids has the rational cross-section property.*

(b) *Deduce:*

Proposition 3.5 *Every rational relation between commutative monoids is uniformised by an unambiguous rational function.*

Solution: (a) We start by supposing $M = \mathbb{N}^k$ and write $\eta = \alpha\alpha^{-1}$ for the map equivalence α (which is a congruence). A rational X of $\mathbb{N}^k$ is a finite union of simple sets and we shall prove the proposition by induction on the number of elements of this union.

If $X = x + B^{\oplus}$ is simple, two elements $u = x = u'$ and $v = x + v'$ of X are congruent modulo η if and only if u' and v' are congruent modulo the congruence η_x: *cf. Exer. II.7.5, p. 329*

$$\forall u, v \in \mathbb{N}^k \qquad u \equiv v \bmod \eta_x \quad \Longleftrightarrow \quad x + u \equiv x + v \bmod \eta \ .$$

The trace on $B^{\oplus}$ of η_x is a congruence of $B^{\oplus}$ which has a rational cross-section S; the set $T = x + S$ is a cross-section of X for η. *cf. Cor. II.7.22, p. 340*

Let $X = \bigcup_{i=1}^{n} x_i + B_i{}^{\oplus}$ be a rational of $\mathbb{N}^k$; we set

$$X = Y \cup Z \ , \qquad \text{with} \quad Y = \bigcup_{i=1}^{n-1} x_i + B_i{}^{\oplus} \qquad \text{and} \qquad Z = x_n + B_n{}^{\oplus} \ .$$

Let P be a cross-section of Y for η (obtained by the induction hypothesis) and R be the cross-section of Z obtained as above. The set of all the elements congruent to at least one element of Y modulo η is the rational $(Y)\alpha\alpha^{-1}$. It follows that $T = P \cup \{R \setminus (Y)\alpha\alpha^{-1}\}$ is a cross-section of X for η.

Let X be a rational of M, an arbitrary commutative monoid. There exists a finitely generated submonoid M' of M which contains X; let $\beta\colon \mathbb{N}^k \to M'$ be a surjective morphism: $Y = (X)\beta^{-1}$ is a rational of $\mathbb{N}^k$. Let, according to the above, the set S be a cross-section of Y for $\beta\alpha$; $T = (S)\beta$ is a cross-section of X for α.

(b) First let $\theta\colon M \to N$ be a rational relation and $\theta = \varphi^{-1}\,\iota_K\,\psi$ be its realisation by morphisms (*cf.* Exer. IV.1.18, p. 538), with $\varphi\colon \mathbb{N}^k \to M$ and $\psi\colon \mathbb{N}^k \to N$ the morphisms and $K \in \operatorname{Rat}\mathbb{N}^k$ (adjacent diagram).

$$\begin{tikzcd} \mathbb{N}^k \arrow[r, "\iota_K"] \arrow[d, "\varphi"'] & \mathbb{N}^k \arrow[d, "\psi"] \\ M \arrow[r, "\theta"'] & N \end{tikzcd}$$

We now reuse the reasoning of Scholium 3.1 verbatim. Let T be a rational cross-section of K for φ. Then $\tau = \varphi^{-1}\,\iota_T\,\psi$ is a (rational) uniformisation of θ.

$\boxed{\mathbf{3.5}}$ (a) *Prove:*[58] *Let $\alpha\colon A^* \to \mathbb{N}^k$ be a morphism. For every rational of R of A^*, there exists an integer n and a rational S contained in R such that $S\alpha = R\alpha$ and each class modulo α contains at most n elements of S.* *Prop. 3.6, p. 684*

(b) *Show that if $k \geqslant 3$, the integer n in the last proposition cannot be chosen independently of R.*

(c) *Show that if $k = 2$, we can take $n = 2$ in the above proposition, independently of R.*

Solution [partial]: (a) We can assume without loss of generality that R is an (elementary) *bounded language*: *cf. Prop. II.7.4, p. 326*

$$R = f_1\, g_1^*\, f_2\, g_2^* \cdots f_n\, g_n^*\, f_{n+1} \ .$$

Set $x = (f_1\, f_2 \cdots f_n\, f_{n+1})\alpha$ and, for each i in $[n]$,

$$b_i = (g_i)\alpha \qquad \text{and} \qquad B = \{b_i \mid i \in [n]\} \ .$$

We have $R\alpha = x + B^{\oplus}$. We can furthermore suppose that the map α restricted to the g_i is *injective*. In fact, if that is not the case, we restrict ourselves to a subset of R by eliminating some of the g_i.

[58] Unpublished result due to M. Latteux and the author.

Lem. II.7.16, p. 335 By the Freedom Lemma, there exists an integer m, m elements y_j of $\mathbb{N}^k$ and m subsets D_j of B such that

$$R\alpha = x + B^{\oplus} = \bigcup_{j=1}^{m} y_j + {D_j}^{\oplus} \ ,$$

and that each of the ${D_j}^{\oplus}$ is a free commutative monoid. For each j of $[m]$, and as y_j belongs to $R\alpha$, there exist integers[59] $l_1, l_2, \ldots, l_n$ such that

$$(f_1\, g_1^{l_1}\, f_2\, g_2^{l_2}\, \cdots f_n\, g_n^{l_n}\, f_{n+1})\alpha = y_j \ .$$

We then define the sets S_j by

$$S_j = u_1\, v_1^{l_1}\, u_2\, v_2^{l_2}\, \cdots u_n\, v_n^{l_n}\, u_{n+1}$$

with, for each i in $[n]$, $u_i = f_i\, g_i^{l_i}$ and

$$v_i = \begin{cases} g_i & \text{if} \quad b_i \in D_j \ , \\ 1_{A^*} & \text{otherwise.} \end{cases}$$

Each S_j is contained in R and we have $(S_j)\alpha = y_j + {D_j}^{\oplus}$. As $y_j + {D_j}^{\oplus}$ is simple and as α is injective on B, we have α is injective on S_j.

(b) Let $A = \{a, b, c\}$ and $\gamma\colon A^* \to A^{\oplus}$ be the canonic morphism. We reuse the technique of Exercise 3.3: for each positive integer j, let

$$P_j = (a\, b^j)^*(a^{j-1}c^j)^* \ ,$$

and, for each integer n, set

$$R_n = \bigcup_{j=1}^{n} P_j \ .$$

Let S be a rational contained in R_n which has the same commutative image as R_n and let N be the integer which is associated with it by the (Block) Star Lemma. Each of the words

$$u_j = (a\, b^j)^{N!}(a^{j-1}c^j)^{N!\,\frac{n!}{j}}$$

belongs to S since there is no commutative equivalent in R_n. By the Block Star Lemma, S also contains the n distinct words

$$v_j = (a\, b^j)^{N!\,\frac{n!}{j}}(a^{j-1}c^j)^{N!\,\frac{n!}{j}}$$

which all have the same commutative image.

(c) This part is left to the reader.

$\boxed{\textbf{3.6}}$ *Show*[60] *that the Rational Uniformisation Theorem holds true for relations from* $A_1^* \times A_2^* \times \cdots \times A_k^*$ *to* $B_1^* \times B_2^* \times \cdots \times B_l^*$ *if and only if either* $k = 1$ *or* $\|A_i\| = \|B_j\| = 1$ *for all* $i \in [k]$ *and all* $j \in [l]$.

see p. 665 *Solution*: We know that the condition $k = 1$ is sufficient (Scholium 2.1). To show that it is necessary if at least one of the alphabets A_i or B_j contains at least two letters, there are two cases to consider: (i) relations from $a^* \times b^*$ to $\{c, d\}^*$; and (ii) relations from $\{a, b\}^* \times c^*$ to d^*.

[59] For full rigour, these integers l_i, like the words u_i and v_i that follow, should be indexed by j; we have allowed ourselves to simplify the notation.

[60] Result taken from [58].

Case (i) We have $a^* \times b^* = \mathbb{N}^2$ and if the rational relations from $\mathbb{N}^2$ to $\{c,d\}^*$ can be rationally uniformised, then $\mathbb{N}^2$ has the rational cross-section property, which it does not (Exercise 3.3).

Case (ii) Consider (once again) $V_1 = (a,c)^*(b,1)^*$ and $W_1 = (a,1)^*(b,c)^*$. If the relation $(V_1,1) \cup (W_1,d)$ were uniformised by a rational function τ, the inverse of τ would be a rational function from d^* to $\{a,b\}^* \times c^*$, hence *unambiguous*. The inverse images $(1)\tau^{-1}$ and $(d)\tau^{-1}$ would be two disjoint unambiguous rationals; their union, $V_1 \cup W_1$, would be an unambiguous rational, which it is not. *cf. Exer. II.1.12, p. 231 and Fact II.1.1, p. 231*

To show that the condition $\|A_i\| = \|B_j\| = 1$ for all i and j is sufficient, is the same as showing that the Rational Uniformisation Theorem holds for relations from $\mathbb{N}^k$ to $\mathbb{N}^l$, and that is what we proved with Proposition 3.5 (Exercise 3.4 (b)).

3.7 *Let $\alpha\colon \{a,b,c,d,e,f\}^* \to \{x,y,z\}^*$ be the morphism defined by*

$$a\alpha = x \ , \quad b\alpha = y\,x\,y\,x \ , \quad c\alpha = x\,y \ , \quad d\alpha = y\,z \ , \quad e\alpha = z\,y\,z\,y \ , \qquad \text{and} \qquad f\alpha = z \ .$$

Show that the radix cross-section of $\{a,b,c,d,e,f\}^*$ for α is not rational.[61]

Solution: Let T be the radix cross-section of $\{a,b,c,d,e,f\}^* = A^*$ and let us evaluate the language $T \cap a\,b^*(d^2)^*d$. We have $(a\,b^n d^{2m+1})\alpha = (x\,y)^{2n} x\,y\,z\,(y\,z)^{2m}$.

Since no image of a letter in A contains $x\,y\,z$ as a factor, a word w belonging to $((x\,y)^{2n} x\,y\,z\,(y\,z)^{2m})\alpha^{-1}$ necessarily factorises into $w = u\,v$, and one of the following assertions is true:

$$u\alpha = (x\,y)^{2n} x \qquad \text{and} \qquad v\alpha = (y\,z)^{2m+1} \tag{a}$$

$$\text{or} \qquad u\alpha = (x\,y)^{2n+1} \qquad \text{and} \qquad v\alpha = z\,(y\,z)^{2m} \ . \tag{b}$$

If (a) is true, then $v = d^{2m+1}$ is the only possibility and $u = a\,b^n$ is the shortest deciphering. If (b) is true, then $u = c^{2n+1}$ is the only possibility and $v = e^m f$ is the shortest deciphering.

Thus, $a\,b^n d^{2m+1}$ belongs to T if $1 + n + 2m + 1 \leqslant 2n + 1 + m + 1$ and hence

$$T \cap a\,b^*(d^2)^*d = \{a\,b^n d^{2m+1} \mid m \leqslant n\} \ ,$$

which is not rational.

3.8 *Show that if M and N are two matrices over $\mathfrak{P}(B^*)$, respectively of dimensions $P \times Q$ and $Q \times R$, and if M is* prefix, *then $M_{\mathsf{lex}} N_{\mathsf{lex}}$ is included in $(M\,N)_{\mathsf{lex}}$.* *Lem. 3.11, p. 689*

Solution: Let $p \in P$, $r \in R$ and $w \in (M_{\mathsf{lex}} N_{\mathsf{lex}})_{p,r} \neq 0$; there then exists q in Q such that

$$u = (M_{\mathsf{lex}})_{p,q} \ , \quad v = (N_{\mathsf{lex}})_{q,r} \qquad \text{and} \qquad w = u\,v \ .$$

By definition, u is the smallest element of $M_{p\bullet}$ and v is the smallest element of $N_{q\bullet}$. Suppose that there exists an element w' of $(M\,N)_{p\bullet}$, $w' \in (MN)_{p,r'}$, smaller than w. There then exists q' in Q such that

$$u' = (M_{\mathsf{lex}})_{p,q'} \ , \quad v' = (N_{\mathsf{lex}})_{q',r'} \qquad \text{and} \qquad w' = u'\,v' \ .$$

By Property 3.3, $w' \prec w$ implies that u is a prefix of u'. Since $\{M_{p,s}\}_{s \in Q}$ is a prefix family, $u = u'$ and $q = q'$. Then, $w' \prec w$ and $u = u'$ implies that $v' \prec v$, and since $q = q'$, we have $v = v'$. Hence, $[(MN)_{\mathsf{lex}}]_{p,r} = w$. *see p. 689*

[61] Taken from [207].

4. Sequential functions

4.1 *Show, with an example, that a deterministic rational relation can be functional without being sequential. Verify that a sequential function is deterministic.*

cf. Exer. IV.5.5, p. 588

Solution: The relation $\sigma_4 \colon \{a,b\}^* \to \{a,b\}^*$ defined by $a^{2p}\sigma_4 = a^{4p}$, $a^{2p+1}\sigma_4 = b^{2p+1}$ is not uniformly bounded.

Each transition $p \xrightarrow{a|u} q$ of a sequential transducer is transformed into a sequence of transitions $p \xrightarrow{a|1_{B^*}} x_1 \xrightarrow{1_{A^*}|b_1} x_2 \xrightarrow{1_{A^*}|b_2} \cdots x_n \xrightarrow{1_{A^*}|b_n} q$, where $b_1\, b_2 \cdots b_n = u$ with the b_i in B and where the x_i are new states; we thereby obtain a deterministic transducer.

cf. p. 693

4.3 *Verify that for all f in A^*, $\mathsf{Dom}\,\alpha/f = f^{-1}\mathsf{Dom}\,\alpha$ (Property 4.1).*

Solution: For all f and g in A^*, we have

$$g \in \mathsf{Dom}\,\alpha/f \iff f\,g \in \mathsf{Dom}\,\alpha \quad \text{and} \quad f \in \mathsf{Dom}\,\overset{\circ}{\alpha}$$
$$\iff f\,g \in \mathsf{Dom}\,\alpha \iff g \in f^{-1}\mathsf{Dom}\,\alpha\,.$$

cf. p. 693

4.4 *Verify that for all f and g in A^*, $(f\,g)\overset{\circ}{\alpha} = \Big[(f)\overset{\circ}{\alpha}\Big]\Big[(g)\overset{\circ}{\alpha/f}\Big]$ (Property 4.2).*

Solution: For all f and g in A^*, we have

$$\begin{aligned}(f\,g)\overset{\circ}{\alpha} &= \bigwedge_{\substack{h\in A^*,\\ fgh\in\mathsf{Dom}\,\alpha}} (f\,g\,h)\alpha = \bigwedge_{\substack{h\in A^*,\\ fgh\in\mathsf{Dom}\,\alpha}} \Big[(f)\overset{\circ}{\alpha}\Big]\Big[(g\,h)\alpha/f\Big]\\ &= \Big[(f)\overset{\circ}{\alpha}\Big]\Big[\bigwedge_{\substack{h\in A^*,\\ gh\in\mathsf{Dom}\,(\alpha/f)}} (g\,h)\alpha/f\Big] = \Big[(f)\overset{\circ}{\alpha}\Big]\Big[(g)\overset{\circ}{\alpha/f}\Big]\,.\end{aligned}$$

4.5 *Prove that for every function $\alpha \colon A^* \to B^*$, and for all f and g in A^*, we have:*
cf. p. 693
$\alpha/(f\,g) = (\alpha/f)/g$ (Lemma 4.1).

Solution: First, we have the sequence of equalities:

$$\mathsf{Dom}\,(\alpha/(f\,g)) = (f\,g)^{-1}\mathsf{Dom}\,\alpha = g^{-1}(f^{-1}\mathsf{Dom}\,\alpha) = g^{-1}(\mathsf{Dom}\,(\alpha/f)) = \mathsf{Dom}\,((\alpha/f)/g)\,.$$

Then we have, for all h in A^*,

$$\begin{aligned}(h)[\alpha/f\,g] &= \Big[(f\,g)\overset{\circ}{\alpha}\Big]^{-1}\Big[(f\,g\,h)\alpha\Big] = \Big[(f)\overset{\circ}{\alpha}\,(g)\overset{\circ}{\alpha/f}\Big]^{-1}\Big[(f\,g\,h)\alpha\Big]\\ &= \Big[(g)\overset{\circ}{\alpha/f}\Big]^{-1}\Big[\Big[(f)\overset{\circ}{\alpha}\Big]^{-1}\Big[(f\,g\,h)\alpha\Big]\Big] = \Big[(g)\overset{\circ}{\alpha/f}\Big]^{-1}\Big[(g\,h)\alpha/f\Big]\\ &= (h)\Big[(\alpha/f)/g\Big]\,.\end{aligned}$$

4.8 ***Uniform bound and Lipschitz condition.*** (a) *Let α be a function whose domain is prefix-closed. Show that α is uniformly bounded if and only if it is Lipschitz.*

(b) *Give an example of a uniformly bounded function which is not Lipschitz.*

Solution: (a) The condition is obviously sufficient. Conversely, suppose that α is uniformly bounded and let K be the integer such that

$$\forall f,g \in \mathsf{Dom}\,\alpha \qquad \mathbf{d}_{\mathrm{p}}\,(f,g) \leqslant 1 \implies \mathbf{d}_{\mathrm{p}}\,(f\alpha, g\alpha) \leqslant K\,.$$

Let f and g be in $\mathsf{Dom}\,\alpha$. The word $h = f \wedge g$ belongs to $\mathsf{Dom}\,\alpha$ as well and $\mathbf{d}_{\mathrm{p}}(f,g) = \mathbf{d}_{\mathrm{p}}(f,h) + \mathbf{d}_{\mathrm{p}}(h,g) = |h^{-1}f| + |h^{-1}g|$. Moreover, if f is written $f = h\,a_1\,a_2 \cdots a_k$ then for each i in $[k]$, $h\,a_1\,a_2 \cdots a_i$ belongs to $\mathsf{Dom}\,\alpha$. By the triangle inequality, we then obtain

$$\begin{aligned}\mathbf{d}_{\mathrm{p}}(h\alpha, f\alpha) &\leqslant \mathbf{d}_{\mathrm{p}}(h\alpha, (h\,a_1)\alpha) + \mathbf{d}_{\mathrm{p}}((h\,a_1)\alpha, (h\,a_1\,a_2)\alpha) + \cdots + \mathbf{d}_{\mathrm{p}}((h\,a_1\,a_2 \cdots a_{k-1})\alpha, f\alpha) \\ &\leqslant k\,K \;=\; K\,\mathbf{d}_{\mathrm{p}}(h,f) \quad .\end{aligned}$$

Likewise, $\mathbf{d}_{\mathrm{p}}(h\alpha, g\alpha) \leqslant K\,\mathbf{d}_{\mathrm{p}}(h,g)$ and α is Lipschitz.

(b) We take $\alpha\colon \{a\}^* \to \{a^*\}$ defined by $(a^{n^2})\alpha = a^{n^3}$; α is undefined for the words whose length is not a square. For every $k \in \mathbb{N}$, there is only a finite number of pairs f, g in $\mathsf{Dom}\,\alpha$ such that $\mathbf{d}_{\mathrm{p}}(f,g) \leqslant k$ and α is trivially uniformly bounded. Furthermore, α is not Lipschitz since:

$$\lim_{n\to+\infty} \mathbf{d}_{\mathrm{p}}\left((a^{(n+1)^2})\alpha, (a^{n^2})\alpha\right) / \mathbf{d}_{\mathrm{p}}\left(a^{(n+1)^2}, a^{n^2}\right) = +\infty \quad .$$

$\boxed{\mathbf{4.9}}$ *Correct the slight imprecision in the equation:*

$$\begin{aligned}\mathbf{d}_{\mathrm{p}}((f\,h)\alpha, (f\,k)\alpha) &= \mathbf{d}_{\mathrm{p}}\left(\left[\left[(f)\overset{\circ}{\eta}\right]\left[((f)\eta^\sharp)_p\right]\,u\right], \left[\left[(f)\overset{\circ}{\eta}\right]\left[((f)\eta^\sharp)_q\right]\,v\right]\right) \\ &= |((f)\eta^\sharp)_p| + |((f)\eta^\sharp)_q| + |u| + |v| \quad .\end{aligned} \tag{4.13}$$

which allows us to complete the proof of Theorem 4.2. *cf. p. 697*

Solution: If $((f)\eta^\sharp)_p \wedge ((f)\eta^\sharp)_q = 1_{B^*}$ and $((f)\eta^\sharp)_q \neq 1_{B^*}$, Equation (4.13) is correct; but if we have

$$((f)\eta^\sharp)_p \wedge ((f)\eta^\sharp)_q = 1_{B^*} \qquad \text{and} \qquad ((f)\eta^\sharp)_q = 1_{B^*} \quad ,$$

then v can share a non-trivial common prefix with $((f)\eta^\sharp)_p$, in which case we have

$$\mathbf{d}_{\mathrm{p}}((f\,h)\alpha, (f\,k)\alpha) \geqslant |((f)\eta^\sharp)_p| + |u| - |v| \quad ,$$

which can also be made arbitrarily large, and this completes the proof of the theorem.

$\boxed{\mathbf{4.12}}$ ***Decidability of sequentiality.*** *Let $\mathcal{T}$ be a real time transducer with n states, m transitions, and outputs of length bounded by k.*

(a) *Let $\mathcal{V}$ be the accessible part of $\mathcal{T} \times \mathcal{T}$ and $\mathcal{V}'$ the sub-automaton of $\mathcal{V}$ comprised of the states which are co-accessible to (at least) one loop whose label is distinct from $(1_{B^*}, 1_{B^*})$. Evaluate the complexity of the computation of $\mathcal{V}'$.*

(b) *Let $\mathcal{W}$ be the product of $\mathcal{V}$ with ω_B and let $\mathcal{W}'$ be the product of $\mathcal{V}'$ with ω_B. Verify that the conditions of Proposition 4.3 hold for $\mathcal{W}$ if and only if they hold for $\mathcal{W}'$.* *cf. p. 699*

(c) *Verify that if $\big((p,q),(1_{B^*},w)\big)$ and $\big((p,q),(1_{B^*},w')\big)$ are two states of $\mathcal{W}'$ then w and w' are comparable (for the prefix order) or condition (ii) of Proposition 4.3 is not satisfied.*

(d) *Let $\big((p,q),(1_{B^*},w)\big)$ be a state of $\mathcal{W}'$; show that if the conditions of Proposition 4.3 are satisfied then $|w| \leqslant k\,n^2$.*

(e) *Prove that we can decide with complexity less than $O(k^2n^4m^2)$ whether $\mathcal{T}$ realises a sequential function (Proposition 4.5).* *cf. p. 703*

Solution [partial]:[62] (a) The transducer $\mathcal{V}$ can be computed in $O(m^2)$, and so can $\mathcal{V}'$, which is obtained by a series of transitive closure operations: that which finds the circuits, then that which determines all the co-accessible states from some of them.

(b) Left to the reader.

[62] Proof taken from [19].

(c) Since (p,q) is in $\mathcal{V}'$, it is co-accessible to a state (r,s) which belongs to a loop labelled $(u,v) \neq (1_{B^*}, 1_{B^*})$; that is, there exists a computation $(p,q) \xrightarrow{f_3|(x,y)} (r,s)$. If w and w' are incomparable, then at least one of the two sets

$$X = \{(1_{B^*}, w) \cdot (x,y) \cdot (u,v)^n \mid n \in \mathbb{N}\} \qquad \text{and} \qquad X' = \{(1_{B^*}, w') \cdot (x,y) \cdot (u,v)^n \mid n \in \mathbb{N}\}$$

contains $\mathbf{0}$ and the state $\big((r,s), \mathbf{0}\big)$ is in $\mathcal{W}'$.

(d) In $\mathcal{W}'$, the *shortest* computation c which joins an initial state $\big((i,j), (1_{B^*}, 1_{B^*})\big)$ to a state $\big((p,q), (1_{B^*}, w)\big)$ has a length less than n^2 ; otherwise, we could decompose c into

$$\big((i,i'), (1_{B^*}, 1_{B^*})\big) \xrightarrow{f_1|(u_1,v_1)} \big((r,s), w_1\big) \xrightarrow{f_2|(u_2,v_2)} \big((r,s), w_2\big) \xrightarrow{f_3|(u_3,v_3)} \big((p,q), (1_{B^*}, w)\big) .$$

cf. p. 700 By Lemma 4.4, the set $X = \{w_1 \cdot (u_2, v_2)^n \mid n \in \mathbb{N}\}$ must be a singleton and $w_1 = w_2$, which would give a shorter computation.

Thus, the length of w is bounded by $k\,n^2$.

(e) To construct $\mathcal{W}'$, we keep two arrays T_1 and T_2 indexed by $Q \times Q$ which are initialised to 1_{B^*}. The states of $\mathcal{W}'$ are computed one after another. For each state $\big((p,q), h\big)$ of $\mathcal{W}'$ and each transition $(p,q) \xrightarrow{a|(u,v)} (p',q')$ of $\mathcal{V}'$, we compute $h' = h \cdot (u,v)$.

(i) If $h' = \mathbf{0}$, condition (ii) of Proposition 4.3 is not satisfied.

(ii) If h' is an element $(w, 1_{B^*})$ (resp. $(1_{B^*}, w)$), we test whether w and $T_1[p',q']$ (resp. w and $T_2[p',q']$) are comparable in the prefix order. There are two possible cases:

• if they are not comparable, then by (c), condition (ii) of Proposition 4.3 is not satisfied;

• if they are comparable then we put the longer of the two in $T_1[p',q']$ (resp. in $T_2[p',q']$).

Thus the algorithm always terminates, either because condition (ii) of Proposition 4.3 is not satisfied, or because the length of the words in T_1 and T_2 exceeds $k\,n^2$, and, by (d), Proposition 4.3 is not satisfied, or because no new state of $\mathcal{W}'$ needs to be computed. In the last case, $\mathcal{W}'$ is finite and $\mathcal{T}$ realises a sequential function.

We construct at most $2k\,n^4$ states for $\mathcal{W}'$ and examine at most $m^2 \times (2k\,n^2)$ transitions since for each transition going from (p,q) to $\mathcal{V}'$, we examine at most one transition from each state $\big((p,q), h\big)$ of $\mathcal{W}'$. The complexity of the construction is at most $2k^2 n^4 m^2$ since that of the comparisons performed to know whether two words are comparable is at most $k\,n^2$.

NOTES & REFERENCES

Functional transducers are rarely treated as such in textbooks, with the notable exception of two works which we have already frequently cited:

S. Eilenberg, *Automata, Languages and Machines*, volume A. Academic Press, 1974. [80]

J. Berstel, *Transductions and Context-Free Languages*. Teubner, 1979. [27]

We also mention, for sequential transducers:

N. Pippenger, *Theories of Computability*. Cambridge University Press, 1997. [189]

E. Roche and Y. Schabes, *Finite-state Language Processing*. MIT Press, 1997. [201]

In the first, one of the most original works on automata, the point of view is inverted: all the theory, or nearly all, is worked out for sequential transducers, and classic finite automata are defined as a special case. The second gives an overview of the use of sequential transducers for the treatment of natural languages.

1. Functionary

The example in Figure 1.1 is due to Choffrut and has already been reproduced in [27]. Theorem 1.1 is due to Schützenberger [224, 225] and has been rediscovered several times. A common reference in the English literature is [35]. *cf. Exer. 1.3, p. 719*

The definition of sequential functions (which he calls sub-sequential) is due to Schützenberger [226]. I have explained the change of terminology that I am trying to introduce; it has already been asked for by Bruyère and Reutenauer in [40]: 'the word *sub-sequential* is unfortunate since these functions should be called simply *sequential*; [...] we hope that someone will find some day a definite terminology.' *cf. Rem. 1.3, p. 652*

The different ways of writing numbers in more or less standard bases, the ways of transforming one into another and the realisation by transducers of operations on the written form of these numbers is an almost inexhaustible source of problems and interesting applications of automata, of which the exercises in this section offer a glimpse. On this rich subject,[63] see [80, Ch. V and XV] and [159, Ch. VII], this last reference containing a broad bibliography. For more details on Avizienis's system, see [93]; for soft addition, see [247].

2. Uniformisation of rational relations

The Rational Uniformisation Theorem (Theorem 2.1) probably appeared for the first time in [134, Th. 2]. A proof which is in substance the one we give appeared in [11]. In [80], this theorem is Proposition IX.8.2, obtained as a corollary of the Cross-Section Theorem. *cf. Exer. 3.1, p. 728*

The Strong Decomposition Theorem is due to Elgot and Mezei [84]. The proof which we give is based on the semi-monomial representation of rational functions due to Schützenberger [225]. We can note here that the transducers corresponding to semi-monomial representations are exactly the *bimachines* of [80].

Relations of finite norm were introduced in [225] in a more general context: Schützenberger studied the possibilities of the growth of the cardinal of the image of a word under a given rational relation as a function of the length of the word. An error in the proof of a key lemma leaves this article incomplete. The decidability of the equivalence of relations of finite norm (Theorem 2.7) was proved by Culik and Karhumäki [69] in a slightly more general context.

[63] Which will certainly be a chapter in another volume, if such a volume ever sees the light of day.

The works of Weber on this subject, which culminate in the beautiful Theorem 2.8 [248], are highly technical, and certainly deserve to be reworked with more algebraic methods.[64]

3. Cross-section of rational functions

Theorem 3.2 appears in [205]; the proof given here is more modern. This result is the starting point for an algebraic theory of deterministic pushdown automata which remains to be developed (*cf.* [205]).

Theorem 3.3 is due to Johnson [131], as is Problem 3.1. Corollary 3.9 appears in [207] where it is proved directly. The proof of Theorem 3.3 by the Schützenberger co-covering construction is taken from [181].

4. Sequential functions

Theorem 4.1 is implicit in [56, *cf.* Proposition 4] and explicit in [197], whose proof we follow. It is an extension of the Nerode criterion given by Raney [192] and by Eilenberg [80, Theorem XII.4.2] to define *their* sequential functions.

The 'Little Sequentiality Theorem' (Theorem 4.2) is due to Choffrut [55, Proposition 3.4]. It can also be seen in [226, Property 2] but without the notion of a uniformly bounded function being made explicit. The proof which we give is in substance that of [226].

cf. Rem. 4.3, p. 703

As already mentioned, the twinning property, equivalent to the conditions attached to Proposition 4.3, was defined by Choffrut in [55]. The fact that it can be tested in polynomial complexity was proved in [249, 19].

cf. Exer. 4.12, p. 704

The existence of a minimal transducer is part of Theorem 4.1 and is hence due to Choffrut; we follow the proof of [197]. The search for an algorithm of complexity proportional to the number of transitions of a transducer was undertaken by Mohri [172, 173]. The proof that we give here follows [18] (*cf.* also [57] for a survey of this subject).[65]

As already mentioned, the Great Sequentiality Theorem is due to Choffrut, and is the generalisation of a theorem of Ginsburg and Rose. The proof which we give is that of Bruyère and Reutenauer [40]; the notion of 'differential of a function', which is instrumental in this proof, has been introduced by Eilenberg [80]. Proposition 4.11 comes from [55], and Theorem 4.6 is due to O. Vaysse [247].

*

The quotations on pages 51, 155 and 476 are obviously the result of machination; however the rest are authentic, or at least, I take them to be so. The second epigraph to the preface was taken from Jorge Amado.[66]

— I do not know, my daughter, whether you will not find that for the most part
I have taught you things that you had no wish to know.

Mme de La Fayette

[64] *Note added in translation:* This program, as well as the definition of more powerful types of covering for computing uniformisation, has been achieved in recent work by R. de Souza in collaboration with the author ([213, 212, 238]).

[65] *Note added in translation:* In recent work with M.-P. Béal and S. Lombardy ([20, 21]), I have used the term 'conjugacy' (which appears in this proof) with a broader meaning (*cf.* Note 88, p. 519).

[66] *Translator's note:* For the derivation of 'tetrapyloctomy' (p. 126), see *Foucault's Pendulum* by Umberto Eco (Harcourt, 1989).

Bibliography

[1] I. J. Aalbersberg and H. J. Hoogeboom, Characterizations of the decidability of some problems for regular trace languages, *Math. Systems Theory*, vol. 20 (1989), 1–19.

[2] I. J. Aalbersberg and E. Welz, Trace languages defined by regular string languages, *RAIRO Theoret. Informatics and Appl.*, vol. 22 (1986), 103–119.

[3] A. V. Aho, Algorithms for finding patterns in strings, in: *Handbook of Theoretical Computer Science* [246], vol. A, 1990, 255–300.

[4] A. V. Aho, J. E. Hopcroft and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, 1974.

[5] A. V. Aho, J. E. Hopcroft and J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, 1983.

[6] J. Almeida, *Finite Semigroups and Universal Algebra*, World Scientific, 1994.

[7] A. W. Anisimov and F. D. Seifert, Zur algebraischen Charakteristik der durch kontext-freie Sprachen definierten Gruppen, *Elekt. Inf. und Kyber.*, vol. 11 (1975), 695–702.

[8] V. Antimirov, Partial derivatives of regular expressions and finite automaton constructions, *Theoret. Computer Sci.*, vol. 155 (1996), 291–319.

[9] A. Arnold, *Systèmes de transitions finis et sémantique des processus communiquants*, Masson, 1992. Translation: *Finite Transitions Systems*, Prentice-Hall, 1994.

[10] A. Arnold, A. Dicky and M. Nivat, A note about minimal non-deterministic automata, *Bull. EATCS*, vol. 47 (1992), 166–169.

[11] A. Arnold and M. Latteux, A new proof of two theorems about rational transductions, *Theoret. Computer Sci.*, vol. 8 (1979), 261–263.

[12] J.-M. Autebert and J. Beauquier, Une caractérisation des générateurs standard, *RAIRO Theoret. Informatics and Appl.*, vol. 8 (1974), 63–83.

[13] J.-M. Autebert and L. Boasson, *Transductions rationnelles*, Masson, 1988.

[14] J.-M. Autebert, L. Boasson and G. Sénizergues, Groups and NTS languages, *J. Computer System Sci.*, vol. 35 (1987), 243–267.

[15] Y. Bar-Hillel, M. Perles and E. Shamir, On formal properties of simple phrase structure grammars, *Z. Phonetik Sprachwiss. Kommunikat.*, vol. 14 (1961), 143–172.

[16] J. Barwise (Ed.), *Handbook of Mathematical Logic*, North Holland, 1977.

[17] M.-P. Béal, *Codage symbolique*, Masson, 1993.

[18] M.-P. Béal and O. Carton, Computing the prefix of an automaton, *RAIRO Theoret. Informatics and Appl.*, vol. 34 (2000), 503–514.

[19] M.-P. Béal, O. Carton, C. Prieur and J. Sakarovitch, Squaring transducers, *Theoret. Computer Sci.*, vol. 292 (2003), 45–63.

[20] M.-P. Béal, S. Lombardy and J. Sakarovitch, On the equivalence of $\mathbb{Z}$-automata, in: *Proc. ICALP 2005*, L. Caires (ed.), *Lecture Notes in Comput. Sci.* n° 3580, 2005, 397–409.

[21] M.-P. Béal, S. Lombardy and J. Sakarovitch, Conjugacy and equivalence of weighted automata and functional transducers, in: *Proc. CSR 2006*, D. Grigoriev, J. Harrison and E. Hirsch (ed.), *Lecture Notes in Comput. Sci.* n° 3967, 2006, 58–69.

[22] M.-P. Béal and D. Perrin, Symbolic dynamics and finite automata, in: *Handbook of Formal Languages* [202], vol. 2, 1997, 463–524.

[23] D. Beauquier, J. Berstel and P. Chrétienne, *Éléments d'algorithmique*, Masson, 1992.

[24] M. Benois, Parties rationnelles du groupe libre, *C. R. Acad. Sci. Paris Sér. A*, vol. 269 (1969), 1188–1190.

[25] M. Benois and J. Sakarovitch, On the complexity of some extended word problems defined by cancellation rules, *Inform. Process. Lett.*, vol. 23 (1986), 281–287.

[26] G. Berry and R. Sethi, From regular expressions to deterministic automata, *Theoret. Computer Sci.*, vol. 48 (1986), 117–126.

[27] J. Berstel, *Transductions and Context-Free Languages*, Teubner, 1979.

[28] J. Berstel, Fonctions rationnelles et addition, in: *Proc. École de printemps de théorie des langages*, M. Blab (ed.), LITP, 1982, 177–183.

[29] J. Berstel and D. Perrin, *Theory of Codes*, Academic Press, 1985.

[30] J. Berstel and J.-E. Pin, Local languages and the Berry-Sethi algorithm, *Theoret. Computer Sci.*, vol. 155 (1996), 439–446.

[31] J. Berstel and C. Reutenauer, *Les séries rationnelles et leurs langages*, Masson, 1984. Translation: *Rational Series and Their Languages.* Springer, 1988.

[32] A. Bertoni, G. Mauri and N. Sabadini, Unambiguous regular trace languages, in: *Algebra, Combinatorics, and Logic in Computer Science*, J. Demetrovics, G. Katona and A. Salomaa (ed.), *Math. Soc. Janos Bolyay* n° 42, North Holland, 1985.

[33] M. Bird, The equivalence problem for deterministic two-tape automata, *J. Computer System Sci.*, vol. 7 (1973), 218–236.

[34] G. Birkhoff, *Lattice Theory*, Amer. Math. Soc. Publications, 1948.

[35] M. Blattner and T. Head, Single valued a-transducers, *J. Computer System Sci.*, vol. 7 (1977), 310–327.

[36] S. L. Bloom and Z. Ésik, Matrix and matricial iteration theories, I, *J. Computer System Sci.*, vol. 46 (1993), 381–408.

[37] N. Bourbaki, *Algèbre*, Hermann, 1956.

[38] A. Brügemann-Klein, Regular expressions into finite automata, *Theoret. Computer Sci.*, vol. 120 (1993), 197–213.

[39] V. Bruyère, G. Hansel, C. Michaux and R. Villemaire, Logic and p-recognizable sets of integers, *Bull. Belg. Soc. Math.*, vol. 1 (1994), 191–238.

[40] V. Bruyère and C. Reutenauer, A proof of Choffrut's theorem on subsequential functions, *Theoret. Computer Sci.*, vol. 215 (1999), 329–335.

[41] J. A. Brzozowski, Canonical regular expressions and minimal state graphs for definite events, in: *Symposium on Mathematical Theory of Automata*, 1963, 529–561.

[42] J. A. Brzozowski, Derivatives of regular expressions, *J. Assoc. Comput. Mach.*, vol. 11 (1964), 481–494.

[43] J. A. Brzozowski and E. J. McCluskey, Signal flow graph techniques for sequential circuit state diagrams, *IEEE Trans. Electronic Computers*, vol. 12 (1963), 67–76.

[44] J. R. Büchi, On a decision method in restricted second order arithmetic, in: *Logic, Methodology and Philosophy of Science*, E. Nagel and al. (ed.), 1962, 1–11. Reprinted in [46], pp. 425–435.

[45] J. R. Büchi, Regular canonical systems, *Arch. für Math. Logik und Grundlagenforschung*, vol. 6 (1964), 91–111. Reprinted in [46], pp. 317–337.

[46] J. R. Büchi, *The Collected Works*, Springer, 1990. Collected by S. McLane and D. Siefkes.

[47] H.-D. Burkhard, Über Experimente an nicht-deterministischen Automaten, *Elekt. Inf. und Kyber.*, vol. 5 (1969), 347–376. Continued in vol. 6 (1970), 3–14.

[48] A. Cardon and M. Crochemore, Détermination de la représentation standard d'une série reconnaissable, *RAIRO Theoret. Informatics and Appl.*, vol. 14 (1980), 371–379.

[49] P. Caron and M. Flouret, Glushkov construction for multiplicities, in: *CIAA 2000*, A. Paun and S. Yu (ed.), *Lecture Notes in Comput. Sci.* n° 2088, 2001, 67–79.

[50] C. Carrez, On the minimalization of non-deterministic automaton. Report of the Computing Laboratory of the Science Faculty of Lille University, 1970.

[51] O. Carton, The growth ratio of synchronous rational relations is unique, *Theoret. Computer Sci.*, vol. 376 (2007), 52–59.

[52] O. Carton, C. Choffrut and S. Grigorieff, Decision problems among the main subfamilies of rational relations, *RAIRO Theoret. Informatics and Appl.*, vol. 40 (2006), 255–275.

[53] A. Cauchy, *Œuvres complètes, 1ère série, Tome V*, Gauthier-Villars, 1892.

[54] K. T. Chen, R. H. Fox and R. Lyndon, Free differential calculus IV – The quotient groups of the lower central series, *Ann. Math.*, vol. 68 (1958), 81–95.

[55] C. Choffrut, Une caractérisation des fonctions séquentielles et des fonctions sous-séquentielles en tant que relations rationnelles, *Theoret. Computer Sci.*, vol. 5 (1977), 325–337.

[56] C. Choffrut, Une caractérisation des codes à délai borné par leur fonction de décodage, in: *Proc. École de printemps de théorie des codes*, D. Perrin (ed.), LITP–ENSTA, 1979, 47–56.

[57] C. Choffrut, Minimizing subsequential transducers: a survey, *Theoret. Computer Sci.*, vol. 292 (2003), 131–143.

[58] C. Choffrut and S. Grigorieff, Uniformization of rational functions, in: *Jewels are forever*, J. Karhumäki, H. Maurer, G. Paŭn and G. Rozenberg (ed.), Springer, 1999, 59–71.

[59] C. Choffrut and J. Karhumäki, Combinatoric on words, in: *Handbook of Formal Languages* [202], vol. 1, 1997, 329–438.

[60] N. Chomsky and G. A. Miller, Finite state languages, *Inform. and Control*, vol. 1 (1958), 91–112.

[61] N. Chomsky and M. P. Schützenberger, The algebraic theory of context-free languages, in: *Computer Programming and Formal Systems*, P. Braffort and D. Hirschberg (ed.), North Holland, 1963.

[62] A. Cobham, On the base-dependance of the sets of numbers recognizable by finite automata, *Math. Systems Theory*, vol. 3 (1969), 186–192.

[63] R. Cohen, Star height of certain families of regular events, *J. Computer System Sci.*, vol. 4 (1970), 281–297.

[64] R. Cohen and J. A. Brzozowski, General properties of star height of regular events, *J. Computer System Sci.*, vol. 4 (1970), 260–280.

[65] P. M. Cohn, *Algebra*, John Wiley, 1974. 2nd edition: vol. I, 1982; vol. II, 1989; vol. III, 1991.

[66] J. H. Conway, *Regular Algebra and Finite Machines*, Chapman and Hall, 1971.

[67] B. Courcelle, D. Niwinski and A. Podelski, A geometrical view of the determinization and minimization of finite-state automata, *Math. Systems Theory*, vol. 24 (1991), 117–146.

[68] M. Crochemore, C. Hancart and T. Lecroq, *Algorithmique du texte*, Vuibert, 2001.

[69] K. Culik and J. Karhumäki, The equivalence of finite valued transducers on HDTOL languages is decidable, *Inform. Process. Lett.*, vol. 23 (1986), 29–31.

[70] K. Culik and J. Karhumäki, Finite automata computing real functions, *SIAM J. Comput.*, vol. 23 (1994), 789–814.

[71] M. Davis, *The Undecidable*, Raven Press, 1965.

[72] P. Dehornoy, *Complexité et décidabilité*, Springer, 1993.

[73] F. Dejean and M. P. Schützenberger, On a question of Eggan, *Inform. and Control*, vol. 9 (1966), 23–25.

[74] D. Derencourt, J. Karhumäki, M. Latteux and A. Terlutte, On continuous functions computed by finite automata, *RAIRO Theoret. Informatics and Appl.*, vol. 28 (1994), 387–408.

[75] L. E. Dickson, Finiteness of the odd perfect and primitive abundant numbers with n distinct prime factors, *Amer. J. Math.*, vol. 35 (1913), 413–422.

[76] R. Diestel, *Graph Theory*, Springer, 1997.

[77] L. C. Eggan, Transition graphs and the star-height of regular events, *Michigan Math. J.*, vol. 10 (1963), 385–397.

[78] A. Ehrenfeucht, D. Haussler and G. Rozenberg, On regularity of context-free languages, *Theoret. Computer Sci.*, vol. 27 (1983), 311–332.

[79] A. Ehrenfeucht, R. Parikh and G. Rozenberg, Pumping lemmas for regular sets, *SIAM J. Comput.*, vol. 10 (1981), 536–541.

[80] S. Eilenberg, *Automata, Languages and Machines*, vol. A, Academic Press, 1974.

[81] S. Eilenberg, *Automata, Languages and Machines*, vol. B, Academic Press, 1976.

[82] S. Eilenberg, C. C. Elgot and J. C. Shepherdson, Sets recognized by n-tape automata, *J. Algebra*, vol. 13 (1969), 447–464.

[83] S. Eilenberg and M. P. Schützenberger, Rational sets in commutative monoids, *J. Algebra*, vol. 13 (1969), 173–191.

[84] C. C. Elgot and J. E. Mezei, On relations defined by generalized finite automata, *IBM J. Res. and Develop.*, vol. 9 (1965), 47–68.

[85] Z. Ésik and W. Kuich, Inductive $*$-semirings, *Theoret. Computer Sci.*, vol. 324 (2004), 3–33.

[86] P. C. Fischer and A. L. Rosenberg, Multitape one-way nonwriting automata, *J. Computer System Sci.*, vol. 2 (1968), 88–101.

[87] M. Fliess, Séries reconnaissables, rationnelles et algébriques, *Bull. Sci. Math.*, vol. 94 (1970), 231–239.

[88] M. Fliess, Transductions algébriques, *RAIRO Theoret. Informatics and Appl.*, vol. 4 (1970), 109–125.

[89] M. Fliess, Deux applications de la représentation matricielle d'une série non commutative, *J. Algebra*, vol. 19 (1971), 344–353.

[90] M. Fliess, Matrices de Hankel, *J. Math. Pures et Appl.*, vol. 53 (1974), 197–222. Erratum in: vol. 54 (1975).

[91] M. Fliess, Sur divers produit de séries formelles, *Bull. Soc. Math. France*, vol. 102 (1974), 181–191.

[92] M. Flouret and E. Laugerotte, Noncommutative minimization algorithms, *Inform. Process. Lett.*, vol. 64 (1997), 123–126.

[93] C. Frougny, On-line finite automata for addition in some numeration systems, *RAIRO Theoret. Informatics and Appl.*, vol. 33 (1999), 79–101.

[94] C. Frougny and J. Sakarovitch, Rational relations with bounded delay, in: *Proc. STACS '91*, M. Jantzen (ed.), *Lecture Notes in Comput. Sci.* n° 480, 1991, 50–63.

[95] C. Frougny and J. Sakarovitch, Synchronized relations of finite and infinite words, *Theoret. Computer Sci.*, vol. 108 (1993), 45–82.

[96] C. Frougny and J. Sakarovitch, Automatic conversion from Fibonacci numeration system to golden mean base, and generalization, *Int. J. of Algebra and Computation*, vol. 9 (1999), 351–384.

[97] C. Frougny, J. Sakarovitch and P. E. Schupp, Finiteness conditions on subgroups and formal language theory, *Proc. London Math. Soc.*, vol. 58 (1989), 74–88.

[98] A. Gibbons and W. Rytter, On the decidability of some problems about rational subsets of free partially commutative monoids, *Theoret. Computer Sci.*, vol. 48 (1986), 329–337.

[99] S. Ginsburg, *The Mathematical Theory of Context-Free Languages*, McGraw-Hill, 1966.

[100] S. Ginsburg and E. H. Spanier, Quotients of context-free languages, *J. Assoc. Comput. Mach.*, vol. 10 (1963), 487–492.

[101] S. Ginsburg and E. H. Spanier, Bounded ALGOL-like Languages, *Trans. Amer. Math. Soc.*, vol. 113 (1964), 333–368.

[102] S. Ginsburg and E. H. Spanier, Semigroups, Presburger formulas and languages, *Pacif. J. Math.*, vol. 16 (1966), 285–296.

[103] Y. Give'on, On some properties of the free monoids with applications to automata theory, *J. Computer System Sci.*, vol. 1 (1967), 137–154.

[104] V. M. Glushkov, The abstract theory of automata, *Russian Math. Surveys*, vol. 16 (1961), 1–53.

[105] R. Godement, *Algèbre*, Hermann, 1970.

[106] P. Gohon, An algorithm to decide whether a rational subset of $\mathbb{N}^k$ is recognizable, *Theoret. Computer Sci.*, vol. 41 (1985), 51–59.

[107] L. Greenberg, Discrete groups of motions, *Canad. J. Math.*, vol. 12 (1960), 415–426.

[108] D. Gries, Describing an algorithm by Hopcroft, *Acta Informatica*, vol. 2 (1973), 97–109.

[109] M. Gross and A. Lentin, *Introduction aux grammaires formelles*, Gauthier-Villars, 1967.

[110] L. H. Haines, On free monoids partially ordered by divisibility, *J. Comb. Theory*, vol. 6 (1969), 94–98.

[111] T. Harju and J. Karhumäki, The equivalence problem of multitape finite automata, *Theoret. Computer Sci.*, vol. 78 (1991), 347–355.

[112] M. Harrison, *Introduction to Formal Language Theory*, Addison-Wesley, 1974.

[113] K. Hashiguchi, A decision procedure for the order of regular events, *Theoret. Computer Sci.*, vol. 8 (1979), 69–72.

[114] K. Hashiguchi, The star height of reset-free events and strictly locally testable events, *Inform. and Control*, vol. 78 (1979), 167–184.

[115] K. Hashiguchi, Limitedness theorem on finite automata with distance functions, *J. Computer System Sci.*, vol. 24 (1982), 233–244.

[116] K. Hashiguchi, Regular languages of star height one, *Inform. and Control*, vol. 53 (1982), 199–210.

[117] K. Hashiguchi, Algorithms for determining relative star height and star height, *Inform. and Comput.*, vol. 78 (1988), 124–169.

[118] G. Higman, Ordering by divisibility in abstract algebra, *Proc. London Math. Soc.*, vol. 2 (1952), 326–336.

[119] J. E. Hopcroft, An $n \log n$ algorithm for minimizing states in a finite automaton, in: *Theory of Machines and Computations*, Z. Kohavi and A. Paz (ed.), Academic Press, 1971.

[120] J. E. Hopcroft, R. Motwani and J. D. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, 2000.

[121] J. E. Hopcroft and J. D. Ullman, *Formal Languages and their Relation to Automata*, Addison-Wesley, 1969.

[122] J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages and Computation*, Addison-Wesley, 1979.

[123] J. M. Howie, *Automata and Languages*, Oxford University Press, 1991.

[124] A. G. Howson, On the intersection of finitely generated free groups, *J. London Math. Soc.*, vol. 29 (1954), 428–434.

[125] O. Ibarra, The unsolvability of the equivalence problem for ϵ-free NGSM's with unary input (output) alphabet and applications, *SIAM J. Comput.*, vol. 7 (1978), 524–532.

[126] G. Jacob, Représentations et substitutions matricielles dans la théorie algébrique des transductions. Thèse Sci. Math. Univ. Paris VII, 1975.

[127] G. Jacob, Sur un théorème de Shamir, *Inform. and Control*, vol. 27 (1975), 218–261.

[128] N. Jacobson, *Lectures in Abstract Algebra*, Springer, 1952.

[129] J. Jaffe, A necessary and sufficient pumping lemma for regular languages, *SIGACT News*, 1978, 48–49.

[130] G. Jirásková and A. Okhotin, State complexity of cyclic shift, *RAIRO Theoret. Informatics and Appl.*, vol. 42 (2008), 335–360.

[131] H. J. Johnson, Rational equivalence relations, *Theoret. Computer Sci.*, vol. 47 (1986), 39–60.

[132] D. Kirsten, Distance desert automata and the star height problem, *RAIRO Theoret. Informatics and Appl.*, vol. 39 (2005), 455–509.

[133] S. C. Kleene, Representation of events in nerve nets and finite automata, in: *Automata Studies*, C. Shannon and J. McCarthy (ed.), Princeton Univ. Press, 1956, 3–41.

[134] K. Kobayashi, Classification of formal languages by functional binary transductions, *Inform. and Control*, vol. 15 (1969), 95–109.

[135] D. König, Egy véssegi tétel és alkalmázasai, *Mathematikai és Fisikai Lapok*, vol. 39 (1932), 27–29. In Hungarian, with a summary in German.

[136] D. König, *Theorie der endlichen und unendlichen Graphen*, Akademie Verlagsgesellschaft, Leipzig, 1936. Reprinted by Chelsea, 1950; English translation, Birkhäuser, 1990.

[137] D. Kozen, A completeness theorem for Kleene algebras and the algebra of regular events, in: *Proc. LICS '91*, 1991, 214–225.

[138] D. Kozen, *Automata and Computability*, Springer, 1997.

[139] D. Krob, Complete systems of B-rational identities, *Theoret. Computer Sci.*, vol. 89 (1991), 207–343.

[140] J. B. Kruskal, The theory of well quasi-ordering: a frequently discovered concept, *J. Comb. Theory, Ser. A*, vol. 13 (1972), 297–305.

[141] W. Kuich, The Kleene and Parikh theorem in complete semirings, in: *Proc. ICALP '87*, *Lecture Notes in Comput. Sci.* n° 267, 1987, 212–225.

[142] W. Kuich and A. Salomaa, *Semirings, Automata, Languages*, Springer, 1986.

[143] M. Kunc, The power of commuting with finite sets of words, *Theory Comput. Syst.*, vol. 40,4 (2007), 521–551.

[144] G. Lallement, *Semigroups and Combinatorial Applications*, John Wiley, 1979.

[145] S. Lang, *Algebra*, Addison-Wesley, 1965. Revised third edition, Springer, 2002.

[146] M. Latteux, Mots infinis et langages commutatifs, *RAIRO Theoret. Informatics and Appl.*, vol. 12 (1978), 185–192.

[147] J. Leguy, Transductions rationnelles décroissantes, *RAIRO Theoret. Informatics and Appl.*, vol. 15 (1981), 141–148.

[148] F. W. Levi, Über die Untergruppen der freie Gruppen, I, *Math. Z.*, vol. 32 (1930), 315–318.

[149] F. W. Levi, On semigroups, *Bull. Calcutta Math. Soc.*, vol. 36 (1944), 141–146. and vol. 38 (1946) 123–124.

[150] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press, 1995.

[151] L. P. Lisovik, The identity problem for regular events over the direct product of free and cyclic semigroups, *Dokl. Akad. Nauk Ukrainskoj RSR ser. A*, vol. 6 (1979), 410–413. In Ukrainian; an English translation by Andreas Weber circulates in manuscript.

[152] S. Lombardy, On the construction of reversible automata for reversible languages, in: *Proc. ICALP '02*, *Lecture Notes in Comput. Sci.* n° 2380, 2002, 170–182.

[153] S. Lombardy and J. Sakarovitch, Derivation of rational expressions with multiplicity, in: *Proc. MFCS '02*, *Lecture Notes in Comput. Sci.* n° 2420, 2002, 471–482.

[154] S. Lombardy and J. Sakarovitch, Star height of reversible languages and universal automata, in: *Proc. LATIN '02*, S. Rajsbaum (ed.), *Lecture Notes in Comput. Sci.* n° 2286, 2002, 76–90.

[155] S. Lombardy and J. Sakarovitch, On the star height of rational languages, in: *Words, Languages and Combinatorics III, Kyoto 2000*, M. Ito (ed.), World Scientific, 2003.

[156] S. Lombardy and J. Sakarovitch, Derivation of rational expressions with multiplicity, *Theoret. Computer Sci.*, vol. 332 (2005), 141–177.

[157] S. Lombardy and J. Sakarovitch, The universal automaton, in: *Logic and Automata: History and Perspectives*, E. Gädel, J. Flum and T. Wilke (ed.), Amsterdam University Press, 2007.

[158] M. Lothaire, *Combinatorics on Words*, Addison-Wesley, 1983. Reprint: Cambridge University Press, 1997.

[159] M. Lothaire, *Algebraic Combinatorics on Words*, Cambridge University Press, 2002.

[160] M. Lothaire, *Applied Combinatorics on Words*, Cambridge University Press, 2005.

[161] R. C. Lyndon and P. E. Schupp, *Combinatorial Group Theory*, Springer, 1977.

[162] W. Magnus, A. Karass and D. Solitar, *Combinatorial Group Theory*, John Wiley, 1966.

[163] A. I. Malcev, On the embedding of group algebras in division algebras, *Dokl. Akad. Nauk SSSR (N. S.)*, vol. 60 (1948), 1409–1501. In Russian.

[164] A. N. Maslov, Estimates of the number of states of finite automata, *Soviet Math. Dokl.*, vol. 11 (1970), 1373–1375.

[165] W. McCulloch and W. Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophysics*, vol. 5 (1943), 115–133.

[166] J. McKnight, Kleene's quotient theorems, *Pacific J. Math.*, vol. 14 (1964), 43–52.

[167] S. McLane, *Categories for the Working Mathematician*, Springer, 1971.

[168] R. McNaughton, The loop complexity of pure-group events, *Inform. and Control*, vol. 11 (1967), 167–176.

[169] R. McNaughton and H. Yamada, Regular expressions and state graphs for automata, *IRE Trans. Electronic Computers*, vol. 9 (1960), 39–47.

[170] G. H. Mealy, A method for synthesizing sequential circuits, *Bell Syst. Tech. J.*, vol. 34 (1955), 1045–1079.

[171] B. Mitchell, *Theory of Categories*, Academic Press, 1965.

[172] M. Mohri, Minimization of sequential transducers, in: *Proc. ICALP '94*, *Lecture Notes in Comput. Sci.* n° 807, 1994, 151–163.

[173] M. Mohri, Minimization algorithms for sequential transducers, *Theoret. Computer Sci.*, vol. 234 (2000), 177–201.

[174] J.-M. Muller, *Arithmétique des ordinateurs*, Masson, 1989.

[175] G. Navarro and M. Raffinot, *Flexible Pattern Matching in Strings*, Cambridge University Press, 2002.

[176] B. H. Neumann, On ordered division ring, *Trans. Amer. Math. Soc.*, vol. 66 (1949), 202–252.

[177] B. H. Neumann, On ordered groups, *Amer. J. Math.*, vol. 71 (1949), 1–18.

[178] M. Nivat, Transductions des langages de Chomsky, *Ann. Inst. Fourier (Grenoble)*, vol. 18 (1968), 339–455.

[179] M. Nivat, Congruences de Thue et t-langages, *Studia Sci. Math. Hungar.*, vol. 6 (1971), 243–249.

[180] B. Pascal, *Œuvres complètes*, Seuil, 1963. The treatise *De numeris multiplicibus*, written with the other arithmetical treatises before 1654, was published by Guillaume Desprez in 1665.

[181] M. Pelletier and J. Sakarovitch, On the representation of finite deterministic 2-tape automata, *Theoret. Computer Sci.*, vol. 225 (1999), 1–63.

[182] D. Perrin, Ehrenfeucht conjecture, *Bull. EATCS*, vol. 27 (1985), 71–82.

[183] D. Perrin, Finite automata, in: *Handbook of Theoretical Computer Science* [246], vol. B, 1990, 1–53.

[184] D. Perrin, Les débuts de la théorie des automates, *Technique et Science Informatique*, vol. 14 (1995), 409–443.

[185] J.-E. Pin, *Variétés de langages formels*, Masson, 1984. Translation: *Varieties of Formal Languages*, Oxford University Press, 1986.

[186] J.-E. Pin, Syntactic monoids, in: *Handbook of Formal Languages* [202], vol. 1, 1997, 679–746.

[187] J.-E. Pin and J. Sakarovitch, Une application de la représentation matricielle des transductions, *Theoret. Computer Sci.*, vol. 35 (1985), 271–293. In English as: Some operations and transductions that preserve rationality, *in: Proc. 6th G.I. Conf. on Theoret. Comput. Sci.*, *Lecture Notes in Comput. Sci.* n° 145(1982), 277–288.

[188] J.-E. Pin, H. Straubing and D. Thérien, Some results on the generalized star-height problem, *Inform. and Comput.*, vol. 101 (1992), 219–250.

[189] N. Pippenger, *Theories of Computability*, Cambridge University Press, 1997.

[190] M. O. Rabin and D. Scott, Finite automata and their decision problems, *I.B.M. J. Res. Develop.*, vol. 3 (1959), 125–144. Reprinted in *Sequential Machines: Selected Papers* (E. Moore, ed.), Addison-Wesley, 1965.

[191] F. P. Ramsey, On a problem of formal logic, *Proc. London Math. Soc.*, vol. 30 (1928), 338–384.

[192] G. Raney, Sequential functions, *J. Assoc. Comput. Mach.*, vol. 5 (1958), 177–180.

[193] L. Redei, *The Theory of Finitely Generated Commutative Semigroups*, Pergamon Press, 1965.

[194] V. N. Redko, On defining relations for the algebra of regular events, *Ukrain. Math. Ž.*, vol. 16 (1964), 120–126. In Russian.

[195] A. Restivo and C. Reutenauer, On cancellation properties of languages which are support of rational series, *J. Computer System Sci.*, vol. 29 (1984), 153–159.

[196] C. Reutenauer, Sur les variétés de langages et de monoïdes, in: *Proc. Theoretical Computer Science 4th GI Conference*, K. Weihrauch (ed.), *Lecture Notes in Comput. Sci.* n° 67, 1979, 260–265.

[197] C. Reutenauer, Subsequential functions: characterizations, minimization, examples, in: *Proc. IMYCS '90*, *Lecture Notes in Comput. Sci.* n° 464, 1990, 62–79.

[198] C. Reutenauer, *Free Lie Algebras*, Oxford University Press, 1993.

[199] C. Reutenauer and M. P. Schützenberger, Minimization of rational word functions, *SIAM J. Comput.*, vol. 20 (1991), 669–685.

[200] A. R. Richardson, Simultaneous linear equations over a division algebra, *Proc. London Math. Soc.*, vol. 28 (1928), 395–420.

[201] E. Roche and Y. Schabes, *Finite-state Language Processing*, MIT Press, 1997.

[202] G. Rozenberg and A. Salomaa (Ed.), *Handbook of Formal Languages*, Elsevier, 1997. Vol. 1, 2 and 3.

[203] J. M. Rutten, Automata, power series, and coinduction: taking input derivatives seriously, in: *Proc. ICALP '99*, J. Wiedermann, P. van Emde Boas and M. Nielsen (ed.), *Lecture Notes in Comput. Sci.* n° 1644, 1999, 645–654.

[204] J. M. Rutten, Behavioural differential equations: a coinductive calculus of streams, automata, and power series, *Theoret. Computer Sci.*, vol. 308 (2003), 1–53.

[205] J. Sakarovitch, Un théorème de transversale rationnelle pour les automates à pile déterministes, in: *Proc. 4th G.I. Conference on Theoret. Comput. Sci.*, K. Weihrauch (ed.), *Lecture Notes in Comput. Sci.* n° 67, 1979, 276–285.

[206] J. Sakarovitch, Description des monoïdes de type fini, *Elekt. Inf. und Kyber.*, vol. 17 (1981), 417–434.

[207] J. Sakarovitch, Deux remarques sur un théorème de S. Eilenberg, *RAIRO Theoret. Informatics and Appl.*, vol. 17 (1983), 23–48.

[208] J. Sakarovitch, A problem on rational subsets of the free group, *Amer. Math. Monthly*, vol. 91 (1984), 499–501.

[209] J. Sakarovitch, On regular trace languages, *Theoret. Computer Sci.*, vol. 52 (1987), 59–75.

[210] J. Sakarovitch, The "last" decision problem for rational trace languages, in: *Proc. LATIN '92*, I. Simon (ed.), *Lecture Notes in Comput. Sci.* n° 583, 1992, 460–473.

[211] J. Sakarovitch, A construction on automata that has remained hidden, *Theoret. Computer Sci.*, vol. 204 (1998), 205–231.

[212] J. Sakarovitch and R. de Souza, On the decidability of bounded valuedness for transducers, in: *Proc. MFCS 2008*, E. Ochmanski and J. Tyszkiewicz (ed.), *Lecture Notes in Comput. Sci.* n° 5162, 2008, 588–600.

[213] J. Sakarovitch and R. de Souza, On the decomposition of k-valued rational relations, in: *Proc. STACS 2008*, S. Albers and P. Weil (ed.), 2008, 621–632. available from `http://stacs-conf.org`.

[214] J. Sakarovitch, T. Nishida and Y. Kobuchi, Recurrent words for substitutions, *Math. Systems Theory*, vol. 17 (1984), 199–215.

[215] M. Sakarovitch, *Optimisation combinatoire*, vol. Graphes et programmation linéaire, Hermann, 1984.

[216] A. Salomaa, Two complete axiom systems for the algebra of regular events, *J. Assoc. Comput. Mach.*, vol. 13 (1964), 158–169.

[217] A. Salomaa, *Jewels of Formal Language Theory*, Computer Science Press, 1981.

[218] A. Salomaa and M. Soittola, *Automata-Theoretic Aspects of Formal Power Series*, Springer, 1977.

[219] M. P. Schützenberger, Quelques propriétés combinatoires de la théorie des automates. Lecture notes taken by Jean-François Perrot, 1968.

[220] M. P. Schützenberger, A remark on finite transducers, *Inform. and Control*, vol. 4 (1961), 381–388.

[221] M. P. Schützenberger, Certain elementary families of automata, in: *Proc. Symposium on Mathematical Theory of Automata*, 1962, 139–153.

[222] M. P. Schützenberger, On a theorem of R. Jungen, *Proc. Amer. Math. Soc.*, vol. 13 (1962), 885–889.

[223] M. P. Schützenberger, Parties rationnelles d'un monoïde libre, in: *Proc. International Mathematics Conference*, 1970, 281–282.

[224] M. P. Schützenberger, Sur les relations rationnelles, in: *Proc. Automata Theory and Formal Languages 2nd GI Conference*, H. Brackhage (ed.), *Lecture Notes in Comput. Sci.* n° 33, 1975, 209–213.

[225] M. P. Schützenberger, Sur les relations rationnelles entre monoïdes libres, *Theoret. Computer Sci.*, vol. 3 (1976), 243–259.

[226] M. P. Schützenberger, Sur une variante des fonctions séquentielles, *Theoret. Computer Sci.*, vol. 4 (1977), 47–57.

[227] M. P. Schützenberger, Une propriété de Hankel des relations rationnelles entre monoïdes libres, *Advances in Math.*, vol. 24 (1977), 274–280.

[228] L. Schwartz, *Méthodes mathématiques pour la physique*, Hermann, 1965.

[229] J. I. Seiferas and R. McNaughton, Regularity preserving relations, *Theoret. Computer Sci.*, vol. 2 (1976), 147–154.

[230] G. Sénizergues, On the rational subsets of the free group, *Acta Informatica*, vol. 33 (1996), 281–296.

[231] E. Shamir, A representation theorem for algebraic and context-free power series in non commuting variables, *Inform. and Control*, vol. 11 (1967), 239–254.

[232] C. Shannon and J. McCarthy (Ed.), *Automata Studies*, Princeton University Press, 1956.

[233] J. C. Shepherdson, The reduction of two-way automata to one-way automata, *I.B.M. J. Res. Develop.*, vol. 3 (1959), 198–200.

[234] P. V. Silva, Free group languages: rational versus recognizable, *RAIRO Theoret. Informatics and Appl.*, vol. 38 (2004), 49–67.

[235] I. Simon, Limited subsets of a free monoid, in: *Proc. FOCS '78*, 1978, 143–150.

[236] I. Simon, Recognizable sets with multiplicities in the tropical semiring, in: *Proc. MFCS '88*, Chytil, Janiga and Koubek (ed.), *Lecture Notes in Comput. Sci.* n° 324, 1988, 107–120.

[237] E. D. Sontag, On some questions of rationality and decidability, *J. Computer System Sci.*, vol. 11 (1975), 375–385.

[238] R. de Souza, On the equivalence of bounded valuedness transducers, in: *Proc. DLT 2008*, M. Ito and M. Toyama (ed.), *Lecture Notes in Comput. Sci.* n° 5257, 2008, 252–263.

[239] J. Stallings, Topology of finite graphs, *Inventiones Math.*, vol. 71 (1983), 551–565.

[240] P. Starke, *Abstract Automata*, North Holland, 1972. Translated from *Abstrakte Automaten*, VEB, 1969.

[241] R. E. Stearns and J. Hartmanis, Regularity preserving modifications of regular expressions, *Inform. and Control*, vol. 6 (1963), 55–66.

[242] W. Thomas, Automata and quantifier hierarchies, in: *Formal Properties of Finite Automata and Applications*, J.-E. Pin (ed.), *Lecture Notes in Comput. Sci.* n° 386, 1988, 104–119.

[243] K. Thompson, Regular expression search algorithm, *Comm. Assoc. Comput. Mach.*, vol. 11 (1968), 419–422.

[244] L. G. Valiant, The equivalence problem for deterministic finite-turn pushdown automata, *Inform. and Control*, vol. 25 (1974), 123–133.

[245] J. van Leeuwen, Graph algorithms, in: *Handbook of Theoretical Computer Science* [246], vol. A, 1990, 525–631.

[246] J. van Leeuwen (Ed.), *Handbook of Theoretical Computer Science*, Elsevier, 1990. Vol. A and B.

[247] O. Vaysse, Addition molle et fonctions p-locales, *Semigroup Forum*, vol. 34 (1986), 157–175.

[248] A. Weber, Decomposing a k-valued transducer into k unambiguous ones, in: *Proc. LATIN '92*, I. Simon (ed.), *Lecture Notes in Comput. Sci.* n° 583, 1992, 503–515.

[249] A. Weber and R. Klemm, Economy of description for single-valued transducers, *Inform. and Comput.*, vol. 118 (1995), 327–340.

[250] D. Wood, *Theory of Computation*, John Wiley, 1987.

[251] S. Yu, Regular languages, in: *Handbook of Formal Languages* [202], vol. 1, 1997, 41–111.

Index

Bold page numbers indicate definitions, and italics indicate exercises and their solutions.

action, 10, **232**, *439*, 583, 693
  Lead or Delay, *238*, *323*, *351*, *367*, **648**, 699
alphabet, **18**
  input, **51**, 52, 53
  symmetric, **305**
antichain, *see* chain (in ordered sets)
Arden's Lemma, **100**, 131, 132, 134, *201*, *215*, 395, 472, 474
automaton, **51**
  accessible, **68**, 77, 107, *176*, 220, 261, 270, 280
  accessible part of, 78, 266, 267, 605, 666, 700, *723*, *735*
  ambiguous, **74**, 647
  behaviour of, **55**, 220, 227, 256
  characteristic, **408**
  co-accessible, **68**, 107, 220, 272, 280
  co-deterministic, **78**, 89, 108, 654
  complete, **76**, 107, 113, 156, *188*, 345, *352*, 660
  deterministic, **77**, 89, 101, 104, 171, *260*, 345, 580, 643, 653, 654
    complete, 270
    expansion of, **286**
  dimension of, **403**
  equivalent automata, **55**, 68, 75, 110, 220, 266
  finite, **53**, 220
  generalised, **97**, 163, **227**, 708
  left, 59, 654
  left regular, *243*, *352*
  local, 662, 718
  minimal, 111, **114**, *260*, 285, 345, 422, 643, 695, 704
  normalised, **88**, *89*, 90, 146, 161, *189–191*, 257
  of an action, **242**, 267
  pushdown, *see* pushdown automaton
  of quotients, **112**, 144
  quotient of, 115, **262**, *264*, 274, 418
    minimum, **262**, 287
  right, 59, 654
  standard, **92**, 146, *191*
    of an expression, *264*, *451*
  sub-automaton, **53**, 56, 75
  transposed, 54, **58**, 108, 256, 259
  trim, **68**, 74, 75, 107, *178*, 220, 280, 709
  trim part of, **68**, 70, 75, 285, 647, 648
  two-way, **172**, 546, *633*
  unambiguous, **74**, 78, 89
  underlying graph of, **51**, 67, 160
  universal, **279**, *304*, 342, 347, *374*
    relative to G, **283**
  valuation of, *see* valuation
  with spontaneous transitions, 97
$\mathbb{K}$-automaton, **403**
  accessible, **408**
  behaviour of, **404**
  characteristic, **408**, *423*, *507*
  co-accessible, **408**
  final function, **403**
  initial function, **403**
  proper, **410**
  $\mathbb{K}$-quotient of, **417**, *451*
    minimal, **421**
  co-$\mathbb{K}$-quotient of, **420**
  standard, **414**
  support, **408**
  trim, **408**
Avizienis adder, 657, *658*, 663, 724, *724*
Avizienis system, *657*, 663, *724*
axiom of choice, *491*

backward closure, *see* transition, spontaneous
ball, *see* graph
behaviour, *see* $\mathbb{K}$-automaton, *see* automaton
  pure, *see* transducer, with...
Benois's Theorem, **314**, *373*

bijection, **10**
bimachine, 737
binomial coefficient, *432*
bisimulation, *see* morphism (of automata)
Block Star Lemma, *see* Star Lemma
Boolean algebra, **9**, 62, *101*, 317, 323
  atom of, *14*, *39*
  effective, 110
  generated, *14*, *39*, **251**, **355**
border, *see* language, prefix-closed
border (of a word), *see* word
bounded growth, *see* language
bouquet, **259**
Brzozowski–McCluskey (BMC) algorithm, 95, **97**, *see* state elimination method
Büchi grammar, **321**
  pure –, *322*, *365*
Büchi system, **319**, *538*

category, *24*, 247, 476
  final object in, 249
  initial object in, 248
Cauchy, *see* product (of series)
Cayley graph, *208*, *304*, 343, *362*, *699*
Cayley–Hamilton Theorem, *213*
chain (in ordered sets), **293**
  antichain, **293**, 300, 324
    wide, *326*, *367*
  ascending chain condition, 294, 326, 328, 332, *369*
  strictly decreasing, 324
circuit, *see* graph
closure, *see* graph, transitive, *see* transition, spontaneous
Cobham's Theorem, *186*, *374*
code with deciphering delay d, *656*, *721*
Coloured Transition Lemma, 269, *632*
commutative monoid, **14**, 30, *65*, *66*, *181*, *182*, *204*, 323, 379, *538*, *630*, *684*, *731*
  affine subset in, **331**, 332, *370*
    slice, *331*
    stable element in affine subset, **331**
  finitely generated, 326, 327, 331
  free, *xviii*, 218, 222, **324**, *373*, 484, *498*, 571, *732*
  free partially –, *520*, *642*
  lexicographic decomposition (in free –), 325, **328**, 332
  linear subset, **334**
  semi-linear subset, **334**
  semi-simple subset, **334**
  simple subset, **334**
  simplifiable, *330*, *333*, *369*
  stable submonoid, *333*, *370*, 600, *602*, *637*
  subtractive submonoid, **330**
commutative operation, **14**
computation, **54**, 219, 404
  almost simple, **303**
  delay of, **602**
  destination of, **54**, 219
  in a $\mathbb{K}$-automaton, **403**
  label of, **54**, 58, 97, 219, 403
  lag in, **603**
  length of, **54**
  prefix of, 603
  result of, **403**
  simple, **303**
  source of, **54**, 219
  successful, **54**, 58, 219, 265, 404
concatenation, **19**, 54, 59
confluence, 308
congruence
  generated, **17**, *26*, 237, 307
  of monoids, **16**
constant term
  of a language, **89**, 139
  of a series, **395**, 396, *398*, 415, 443, *501*, 559
  of an expression, **139**, 399
  of a $\mathbb{K}$-expression, 444
context, **249**, *355*
convergence
  simple, 386, 552
  uniform, *392*, *500*, *553*, *633*
Correspondence, *see* Post Correspondence Problem
Corrugated Cardboard Lemma, **581**, 688
coset, *see* subgroup
covering, *see* morphism (of automata)
cross-section, *231*, 665
  rational – property, **680**

decidable (problem, property), **36**, 69, 341, 571, 574, 589, 643, 648, 650, 703
decomposition (lexicographic –), *see* commutative monoid
Decomposition Theorem
  Strong –, *524*, **673**
  Weak –, *524*, **673**, 689
dense subset, 390, *392*

denumerable, **9**
depth, *see* rational expression
derived term
– automaton, **151**, *264*
– of an expression, 127, **151**
– $\mathbb{K}$-automaton, **449**, *451*, *520*
– of a $\mathbb{K}$-expression, **447**
phantom –, **450**, *520*
destination, *see* computation, transition
detachable identity, *see* free product
determinisation (by the subset construction), **106**, 153, *176*, *243*, 270, 272, 482, *484*
of letter-to-letter automata, 600, 613
diagonal (of a Cartesian square), **12**
Dickson's Lemma, **325**, 328, *373*
dimension, *see* automaton, *see* vector space
discriminating, *see* length
distance, **386**, 388, 393, *499*
gap, **387**, 556
Hamming, *528*
prefix, **696**, *699*, 715
ultrametric, **386**, *392*
division relation, *see* order
division ring, *see* skew field
divisor of the identity, *27*, *225*, 283, *348*, 384, **476**, 477, 480, 485, 487, 488, *499*, *516*
domain
of a function, **10**
of a relation, **11**
Dyck congruence, **306**
Dyck reduction, **310**, 321, *323*, *367*, *373*, 533
Dyck word, **306**, *313*, *362*

echelon system, *see* Gaussian elimination
edge (of a graph), **33**
Eggan's Theorem, 157, **162**, *215*
Ehrenfeucht–Haussler–Rozenberg Theorem, 217, **298**, *373*
Ehrenfeucht–Parikh–Rozenberg Theorem, 49, **119**, *215*, 465
Elgot–Mezei Composition Theorem, **534**
Elgot–Mezei Structure Theorem, *xix*, **673**, *737*
elimination, *see* state elimination method
elimination, *see* Gaussian elimination
endmarker, **172**, 661, *see also* transducer
equidivisible, *see* monoid
Levi's Theorem, *see* Levi
equivalence (relation), **12**, 34, 665
index of, 713
map equivalence, **12**, 16, 236, 248, 262, 328, 341, *354*, 419, 422, 490, 592, 679, 713, *729*
quotient by, **12**, 293
right regular, 248, 249, *354*
expression, *see* rational expression

factor (of a word), **20**, 68, 83, *181*, *228*
factorisation, **21**, 71, 119, *122*
proper, **21**
iterating, **71**, *177*, *208*, *254*, *356*
left, **20**
proper, **20**
right, **20**
factorisation (of a subset), **275**, 345
factor in, **276**
sub-factorisation, **276**
term of a sub-factorisation, **276**
Fatou property, 301, **303**, 317, 340, *342*, *348*, *371*, 429, 520, 596, 714
field, **29**, 379, *463*, 491
skew, **29**, 379, 451, 463, 491, 495, 520, 575, 616, 625
final monoid, *see* free product (of monoids)
Fine–Wilf's Theorem, *23*
finite power property, *504*
finite type, *see* monoid
finitely decomposable, *see* monoid element
Floyd's algorithm, 95
free product (of monoids), **476**, *516*, *520*
bipartite automaton, **478**
deterministic, **479**
proper, **478**
detachable identity property, **480**
final monoid, 477
immiscible product, 477
initial monoid, 477
rational automaton on, **479**
Freedom Lemma, **335**, *732*
function (general case), **10**
bijective, **10**
characteristic, **29**, 175, 381, 404, 408, 550, 660, 662, 693, 703
injective, **10**
partial, **10**
surjective, **10**
total, **10**
function (on free monoids)
adjacent functions, **715**
differential of a function, **711**

frank, **693**, 706
Lipschitz, **697**
prefix-closed, **660**, 717
rational, *see* rational function
translation of function, 643, **693**, 704
uniformly bounded, 644, **696**
functional (on a vector space), **493**
Fundamental Theorem of Finite Automata, 225, **409**, 433, 442, *519*, 529, 530, 562, 568, 600
future of a state, 99, *260*, **274**

gap, *see* distance
Gaussian elimination, 99, 133, **459**
echelon system, **459**
gradation, **384**
graph, **33**
ball in, **160**, 344, 345
circuit, 33
depth-first search algorithm, *183*
incidence matrix, 33
labelled, 51, *217*
– morphism, 256
multigraph, 33
partial graph, 53
partial subgraph, 53
path, 33, 67, 404
– of a relation, **10**, 30
simple, 33
strongly connected component, **34**, 160, 709
subgraph, 53
transitive closure of, **33**, 35, *73*, 79
– traversal, 67, 69, *73*, *183*, *184*
group, **15**, 27, *246*, 342, *353*, 517, 704
commutative, *330*, 490
finitely generated –, **302**, 317, *364*
free, *xviii*, 65, 218, 301, 305, 314, *314*, 317, *363*, *373*, 377, 470, 477, 482, 575, 616, 618–620, *642*, 679, 682, 699, 711
of units, *18*, 619
ordered, *see* ordered set
residually finite, *314*, *363*
simple, *215*
subgroup, 339, *370*
coset of, **301**, *318*
finite index –, **302**, 303, *318*, *361*, *364*
normal, *18*, *40*, *361*, 620
group language, **342**, *374*
growth (bounded –), *see* language

Hadamard, *see* product (of series)
Hamilton, *see* Cayley
Hamming distance, *528*
Hankel matrix, **452**, *514*
Haussler, *see* Ehrenfeucht
Higman's Theorem, **296**, *373*, 622
Howson's Theorem, 317

ideal, *see* monoid, *see* ordered set
ideal, *see* ring, *see* syntactic
identity (rational), **129**, 394
aperiodic, **130**, 395
classical, 132
cyclic, **131**, 395
natural, **130**
trivial, **130**
image
of a function, **10**
of a relation, **11**
of an element, **10**
immersion, *see* morphism (of automata)
immiscible product, *see* free product (of monoids)
incidence matrix, *see* graph
index, of an automaton, of a transition, *see* star height
infiltration
of two automata, *182*
of two languages, *66*, *182*
of two series, *432*
of two words, *66*, *182*, *432*, *529*, *628*
initial monoid, *see* free product (of monoids)
iteration, 84

Johnson's Problem, 688, *726*, *738*
Johnson's Theorem, 686, *738*

$\mathbb{K}$-automaton, *see* under letter a
$\mathbb{K}$-function, *see* $\mathbb{K}$-relation
$\mathbb{K}$-relation, *see* under letter r
$\mathbb{K}$-representation, *see* representation
Kleene's Theorem, 82, **87**, 94, 110, 123, 127, 145, 157, 225, 239, 242, 250, 433, 533, 540
Kleene–Schützenberger Theorem, **433**, 438, 442, 462, *519*, 543, 562, 564–566, 574
Knuth–Morris–Pratt algorithm, 156
Kronecker delta, *432*, 493

label, *see* computation, *see* transition
language, **22**

accepted, **55**
bounded, 731
denoted by an expression, **125**
factor-closed, **23**
first half of, first third of, *177*, *209*
local, *228*, 662
of bounded growth, *177*, *210*
prefix, **22**, *65*, *85*, *181*, *189*, 595, *635*
prefix-closed, **23**, *111*, 455, 659, 711, 717
border of a prefix-closed language, **457**
prefix-free, *22*
rational, *see* rational, language
recognisable, *see* recognisable language
recognised, **55**, 97
reversible, **347**
root of, *175*, *204*, **252**
strictly locally testable, 664
suffix, **22**
suffix-closed, **23**
thin, *177*, *210*
Laurent series, *498*
length, *see* computation, *see* word
– function, **384**
discriminating, **111**, *195*, *462*, *515*
literal, *see* rational expression
length discrepancy, *see* relation
letter, **18**
Levi's Theorem, 26, *43*
lexicographic
decomposition, *see* commutative monoid
order, *see* order
Lindenmayer system, *538*
linear combination, 489
literal length, *see* rational expression
loop complexity, *see* star height

Magnus representation, 617, **618**, 620
Malcev–Neumann series, xix, *524*, **616**
Malcev–Neumann Theorem, 575, 616, *642*
map, **10**, 16
matrix, **30**
block decomposition, **31**, 401
boolean, **30**, 33, 233
column-monomial, **31**, 233
determinant of, *515*
diagonal, **31**, 704
dimension of, 30
incidence, 53
nilpotent, 406
prefix, *see* prefix, matrix
proper, 396, 410
rank of, 452, *496*
row-monomial, **31**, 233
semi-monomial, **670**
square, 30
trace of, *32*, *426*, *509*
transition, *see* transition matrix
transpose, **32**, 58
upper triangular, *31*, 393
stict, *31*, 406, 411, 412
Vandermonde –, *515*
McCluskey, *see* Brzozowski
McKnight's Theorem, **239**, 242, 303, *372*, 436, 561, *633*
McNaughton–Yamada (MNY) algorithm, **95**, 123, 158, 163, *192*, 230, 401, 402, 411
Mealy machine, 175, *641*
Mezei's Theorem, **254**, *372*
mirror image, **21**, *26*, 58, 537, 693
module, 380, **489**
basis of, **490**
free, **490**
of finite type, **490**
monoid, **14**, 82
commutative, *see* commutative monoid
equidivisible, **26**, *32*, *43*, *45*, *246*, *254*, *353*, *356*, *362*, *423*, *508*
finitely generated, **15**, *225*, *349*, 384, 558
free, **24**, 477
generating set, 20
generating set of, **15**
graded, 26, **384**, 392, 395, 548, 556, 558, 559
strictly, **575**
ideal in, *17*, 324
left, *17*, 324, 343
right, *17*, 324, 343
two-sided, *17*, 324
involutive, 305, **306**, 682
locally finite, 384
of finite type, **15**, 384
of maps, **16**, 176
of relations, **16**
presentation of, *385*
rationally enumerable, *231*, *350*
simplification, *see* simplification
submonoid, **15**
syntactic, 342
transition monoid, **232**, *351*, *352*

monoid element
  finitely decomposable element, **383**
  inverse element, **15**, *18*
  neutral element, **14**
  zero, **15**
Moore machine, 174
Moore's algorithm, **116**, 117, *117*, *118*, *196*, *197*, 263, 287, 423, 708
morphism (of action), **247**
morphism (of automata), 255, **256**, 257, *260*, 284, 289, 344, *357*, *372*, 418
  conformal, **258**, 261, 343
  covering, *113*, 255, **265**, 266, 267, 269, *357*, 486, *513*, 579, 588, 605, 707
    S-covering, **270**
    Schützenberger –, *xviii*, *xix*, **270**, *372*, 466, 643, 666, 689
  co-covering, **265**, 605
    S-co-covering, **272**, 690
    Schützenberger –, **272**
  immersion, **267**, 268, 579
    S-immersion, **272**, 666, 669
  co-immersion, **267**
    S-co-immersion, **272**
  In-,Out-bijective, **260**, 265
  In-,Out-injective, **260**, 267
  In-,Out-surjective, **259, 260**, 270, 272
  simulation, *262*, *372*
  bisimulation, *264*
  totally surjective, **262**
morphism (of monoids), **16**, 24, 37, 40, 43, 44, 219, 221, 223, 225, 231, 526, 532, 564, 585, 592, 661
  alphabetic, **25**
    strictly, **25**, *42*, *228*, *349*
  continuous, **25**, *26*, 557, 559, 592, *634*, 663
  inverse, 585
  prefix, **595**
  projection, **25**, *26*, *42*, 559
  split, **337**
morphism (of ordered sets), **298**
  strict, **622**, 624
morphism (of semirings), **28**, 31, 221, 425
  continuous, **391**, 394
Morris, *see* Knuth
Morris–Pratt algorithm, 156

Nerode's equivalence, **114**, 614
Nivat's Theorem, **532**, 563, *641*, *731*
notation, *see* rational expression
numbers (k-recognisable set of –), *see* recognisable

operation, *see* rational
operation (commutative –), **14**
order, **12**, 293, 665
  cartesian product of, 324
  division relation, 295, **295**, 297, 299, *361*, 622
  finite basis property, **294**
  genealogical, **528**
  ideal in ordered set, **294**, 296, 324
  lexicographic, **24**, *41*, *329*, *528*, 610, 618, 625, *627*, 686
    on $\mathbb{N}^k$, **326**
  military, **528**, **664**
  natural, **324**
  partial, **12**, 324
  prefix, **21**, 581
  product, 274
  quasi-order, **293**
    well quasi-order, **294**
  radix, 452, **528**, 610, 619, *627*, 664
  regular, 295
  total, **12**, 73, 95, 293, 575
  well order, **12**, *24*, *41*, 294, 326, 575, 622, 665
  well partial order, **294**, 325, 622
ordered set
  group, 381, **575**, 616, 617, 623
  ring, 619
  semigroup, **623**
ordering, *see* order

palindrome, **22**, *73*, *185*
parenthetic word, **306**
Parikh, *see* Ehrenfeucht
partial order, *see* order
partially commutative monoid, *see* commutative monoid
partition, **9**
past of a state, *260*, **274**
path, *see* graph
  in a $\mathbb{K}$-automaton, **403**
PCP, *see* Post Correspondence Problem
period, *see* word
Pigeonhole Principle, 70
polynomial, 381, **381**, 382, 411
Post Correspondence Problem, **36**, 469, 524, 569–571, 589, *635*
Post's Theorem, **37**, 569, 650
power series, *see* series

Pratt, *see* Knuth, *see* Morris
prefix
– family of languages, *85*, **590**, *636*
– matrix, *85*, *189*, **590**, 689, *691*, *733*
– representation, **590**, 592
in $A^* \times B^*$, **581**
longest common –, 581
of a computation, **580**
prefix (of a word), **20**, 57, 63, 68, 83, *85*, 112, 153, 154, *181*, *189*, 245, *528*, 580, *627*, 647, 703
longest common –, *111*, 692, 705, 709
proper, **20**, *181*
prefix, *see* computation, *see* distance
prefix, *see* language, *see* notation
prefix, *see* order
prefix-closed, *see* function, *see* language
prefix-free, *see* language
prefix-preserving, *see* function, prefix-closed
Presburger arithmetic, *373*
problem
equivalence –, **38**, 569
inclusion –, **38**, 569
intersection –, **38**, 569
universality –, **38**, 569
product
infiltration, *see* infiltration
of languages, **83**, 86, 90
shuffle, *see* shuffle
synchronised – (of automata), *639*
tensor, **426**
unambiguous, *92*, **228**
product (cartesian –)
of an automaton and an action, **267**, 605, 647
of automata, **61**, 75, 78, *182*, *220*, 257, 267, 647
of sets, **9**
of transducers, 573, **647**
product (of series)
Cauchy –, **380**
Hadamard –, 380, **382**, 428, 435, 464
scalar –, 380, **382**
projection, *see* morphism (of monoids)
proper, *see* $\mathbb{K}$-automaton, *see* matrix
pushdown automaton, *xix*, 218, 301, 305, 377, *642*, 679, *738*

quasi-ordering, *see* order, quasi-order
quotient (by a word)
bi-, *251*, *355*
of a language, **63**, 70, 112, 120, 140, 173, *199*, 439, 693
of a series, **438**, 451
of a subset, **245**, 248, 278
of a word, **63**, 112, *237*, 439

Rabin–Scott automaton, **171**, *216*, 524, 541, 545, 576, 577, 595, 602, *641*
Rabin–Scott Intersection Theorem, 569, *641*
Ramsey's Theorem, 118, **120**, *123*, *199*, *215*
rank, *see* matrix, *see* series
rational
– closure, **86**, 222
– language, **86**
bounded rational language, *326*, *368*
– operation, **85**, 123, 222, 398, 412, 474
unambiguous, **228**, 382, 397
– subset, **222**, 525
ambiguous, **230**, *415*
inherently ambiguous, **230**
unambiguous, **229**
Rational Cross-Section Theorem, *524*, 681, *684*
rational expression, 86, 91, 123, **124**, 222, 527, 543
depth, **125**, 151, 223, 444
derivative, 127, **139**
derived term, *see* derived term
equivalent expressions, **126**
interpretation, **127**, 223
free interpretation, **128**
literal length, **147**, 151
notation
infix,postfix,prefix, 125, *128*, *200*
star height, *see* star height
weighted, 138, **399**
rational $\mathbb{K}$-expression, 394, **399**, 402, 444
denoted series, 400, 444
depth, 399, 448
derived term, *see* derived term
equivalent $\mathbb{K}$-expressions, 444
interpretation, **400**
literal length, 444
valid, **400**, 444
rational function, 645, 650
local, **661**, 662, 718
degree of, **661**
sequential, 643, **651**, 673, 692, 694
pure, **658**, 673, 717

co-sequential, **654**, 673
pure, **658**, 673
subsequential, 652, 654
unambiguous, 665
rational relation, **525**
bld-relation, 596, 598, 604, 610
composition of, 534, 543
deterministic, **584**
functional, *see* rational function
inverse, 526
iteration of, **536**, *538*, *631*
proper, **536**
pure deterministic, **585**
synchronous, 596, **609**, 612
left, **609**
right, **610**
transitive closure, 536
unambiguous, 558, 650
universal, 584, 588
rational $\mathbb{K}$-relation, **557**, 567
composition of, 565
inverse, **558**
Rational Skimming Theorem, **465**, 677
Rational Uniformisation Theorem, *see* uniformisation
ray, *177*, *210*, *326*
recognisable
– subset, **234**
k-recognisable set of numbers, *74*, *186*
language, **56**
recognisable relation, **539**, 588, 611, *666*, *726*
iteration of, *540*, *631*
proper, *540*, *631*
reduction, *see* Dyck, Shamir
regular expression, *see* rational expression
regulator, 533
relabelling, **225**, 257
relation, **10**
additive extension of, **11**, *14*, 63
bld-relation, **597**
bounded-length discrepancy –, **597**
composition of, **11**
equivalence, *see* equivalence relation
of finite co-image, *385*
functional, **11**
identity, **12**
injective, *14*
inverse, **11**
length discrepancy of a pair of words, **596**, *602*, *636*
length-preserving, **597**
norm of, **679**
rational, *see* rational relation
recognisable, *see* recognisable relation
restriction of, **13**
transitive closure of, **13**, 33
universal, **12**, 17
$\mathbb{K}$-relation, **550**
characteristic, 552
composition of, **553**
$\mathbb{K}$-function, **552**, 646
of finite image, 551, 553
of finite co-image, 551, 553, 557
injective, **552**
multiplicative, **555**
rational, *see* rational $\mathbb{K}$-relation
regulated, 551
unambiguous, 552, 558
representation (matrix), 542
column-monomial, 654
composition product of, 543, **567**, 656, 659
$\mathbb{K}$-representation, **424**
reduced, **453**
row-monomial, 651
semi-monomial, **670**
trim, 646
rewriting, 297
ring, **29**, 379, 388, 394, 490
commutative, *462*
ideal in, *294*, *462*
ordered, *see* ordered set
Roy–Warshall algorithm, *73*, 95, *184*
Rozenberg, *see* Ehrenfeucht

Schützenberger product, *45*, *356*, *360*
Scott, *see* Rabin
semigroup, **15**, 233
ordered, *see* ordered set
simplifiable, 623
semiring, **27**, 30, 129, 221
Boolean, 28, 221
centre of, *29*, *44*, 427, 548, 557
commutable sub-semirings, **427**, *514*, 565
commutative, **28**, *32*, 379, 427, 428, 430, 489
complete, 221, 381
continuous, 548, 558, 559
equisubtractive, *423*, *508*
positive, **28**, 408, 415, 426, 463, 646

reasonable, **559**
strong, **396**, 548, 558
summable family in –, **389**, 393
topological, **388**, 393, 548
tropical, **28**
sequential transducer, **653**, 694
blocked, **705**
blocked state in, **705**
left, 654, 655
minimal, 643, 695, **707**, 718
reduced transducer, **706**
right, 654, 655, *657*
separate states, **706**
translation transducer, 705, 707, 708
Sequentiality Theorem
Great –, 644, **711**, 716, *738*
Little –, 644, **697**, *738*
series, **379**
characteristic, **381**, 408, 415, 439
commutable, 561
$\mathbb{K}$-rational, **399**
commutable, **435**, 436
$\mathbb{K}$-recognisable, **424**
commutable, **429**
Laurent, *498*
proper, **395**
proper part of, **396**, 411, 443
rank of, **452**, *463*
rational, **399**
support of, **381**, 408, 415, *634*
valuation of, **387**, 390, 556, 620
series (family of –)
locally finite family, **390**, 395, 557, 625
proper family, **411**
summable family, **390**, 395, 404
Shamir congruence, **306**, 311
Shamir reduction, **310**, 319, 321, *373*
shuffle
of two automata, *182*
of two languages, *65*, *181*
of two series, 430, 436
of two words, *65*, *181*, *529*, *628*
Simon's automaton, 157
simplification, 305, **307**
– monoid, **314**, 679, 682
reduction associated with, **310**
simulation, *see* morphism (of automata)
slice, *see* commutative monoid, affine subset
source, *see* computation, *see* transition
of a transition, 219
spontaneous, *see* transition
stable
– module, 440, 441, 451, *514*
– subset, **440**
element, *see* commutative monoid, affine subset
submonoid, *see* commutative monoid
star
in a semiring, **393**
of a language, **83**, 86, 91
of a subset, **221**
of a word, 71, **83**
proper, **83**
unambiguous, **229**
star height, *xviii*, 49, *161*, *204*, *215*, 333, 342, *374*, *504*
generalised, 157, **170**, *204*
index of a transition, **163**
index of an automaton, **163**
loop complexity, 157, **160**, *161*, 162, *203*, *204*, *215*, *304*, 342, 343, 347, *362*, *371*, *374*
of a language, **159**
of a rational expression, *125*, **158**, 304, *416*, 599, *637*
Star Lemma, *xviii*, **71**, 73, 118, *187*, *349*, 537, *538*, 586, *630*
Block –, **71**, 118, 120, 169, *208*, 346, *730*, *732*
state, **51**, 219, 403
accessible, **67**, 69, 220, **408**
co-accessible, **67**, 69, 220, 272, **408**
final, **51**, 69, 88, 99, 133, 219, 259, 404
future of, *see* future
initial, **51**, 63, 69, 88, 92, 219, 259, 404, 414
past of, *see* past
rank of, *see* transducer
sink, **77**, 585
terminal, 51
useful, **67**, 97, 220, **408**
state elimination method, **97**, 123, 163, *191*, 401, 411
strongly connected component, *see* graph
subgroup, *see* group
subset (of a monoid)
disjunctive, *251*, *355*
recognisable, **234**
recognised by a morphism, **236**, 277
recognised by an action, **234**
saturated by an equivalence, 236

substitution, **252**, 526, 533, 665
subword, *see* word
suffix (of a word), **20**, 68, 83, 153, 154, *181*, 245, *528*, *627*
  proper, **20**
syntactic
  – algebra, *462*
  – congruence, **249**, *355*
  – ideal, *463*
  – monoid, **249**
  – property, **250**, 277
system of equations, **100**

Thue system, **319**
topology, 376
  discrete, 376, **387**, *392*, 393, *500*, 696
  product, 386
trace, *see* matrix
transducer, 59, 268, **529**
  conjugate, **704**
  deterministic, **578**
  final function, 653
  in normal form, **529**, 578
  initial function, 653
  k-tape automaton, **545**, 595
    deterministic, **545**, 576, 577, 589
  letter-to-letter, 175, 565, **596**
    composition of, 608
  of bounded lag, 268, **603**
  output function, **653**
  rank of a state, **604**
  real-time, 646, 666
  resynchronisation algorithm, *609*, 705
  sequential, *see* sequential transducer
  translation –, **694**, 707
  underlying input automaton, **529**, 647, 653, 654, 705, 706, 718
  with an endmarker, **577**, 592
    pure behaviour of, **577**, 592
transition, **51**, 219, 255
  – function, **102**
  – matrix, **54**, 403
  active, **156**
  destination of, **54**, 219, 255
  label of, **54**, 219, 255
  source of, **54**, 255
  spontaneous, **79**, *89*, *190*, *220*, 286, *409*, 411
    backward closure, **80**, 93, 148, 286, 411
    forward closure, **80**, 93, 411, *504*
  table of, **102**
transpose, *see* matrix, *see* word
trim, *see* automaton, $\mathbb{K}$-automaton
Turing machine, 36, 171, 172, 545
twinning property, **703**, 738

undecidable (problem, property), **36**, 468, 569, 570, 573, 589, 615
underlying graph, *see* automaton
uniformisation, **664**
  lexicographic selection, **665**, 685, 686
  radix, 664, 685
  Rational Uniformisation Theorem, *524*, *643*, 665, 677, 679, 685, *685*, 689, *726*, *732*, *737*
  S-uniformisation, **666**, *677*, *728*
universal, *see* automaton, *see* relation

valuation, *see* series
valuation of an automaton (by an action), **268**, *358*, 648
vector
  column, **30**, 54, 403, 424, *433*
  final, **403**
  initial, **403**
  row, **30**, 54, 403, 424
vector space, 451, 491, 495
  dimension of, **491**, *515*
  dual, **493**
  functional on, **493**
  orthogonal of a subspace, **494**
  supplementary subspace, **492**, 494
vertex (of a graph), **33**

wide antichain, *see* chain (in ordered sets)
Wilf, *see* Fine
Williams's Theorem, *519*
word, **18**
  adult, *538*, *631*
  border of, **154**, *157*
  conjugate, *27*, *42*, *92*, *190*, *279*, *360*, *538*, *630*
  empty, **19**
  first half of, first third of, *209*
  length of, **20**, **25**
  period of, *23*, *157*
  recurrent, *538*, *631*
  reduced, **308**
  subword, **21**, *65*, *66*, 295, *432*, *528*, *627*
  transpose, **21**, 58